PhD Course and Research Review

Spring 2025
Version 3

Miles Woollacott

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

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

Contents

1 ST 701: Statistical Theory I

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

1.1 Probability

Return to Table of Contents

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

1.2 Distributions

Return to Table of Contents

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

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

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

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

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

1.3 Moments and Expectations

Return to Table of Contents

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

1.4 Bivariate Random Variables

Return to Table of Contents

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

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

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

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

  Case 1: $z < 0$.

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

  Case 2: $z > 0$.

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

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

1.5 Statistics and Order Statistics

Return to Table of Contents

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

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

1.6 Convergence

Return to Table of Contents

- An estimator a is **consistent** if $a \xrightarrow{\text{p}} \mathbb{E}(X)$.
- A sequence of RVs $X_1, X_2, \ldots$ **converges in probability** to RV X if for every $\epsilon > 0$, $\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$, or $\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$, denoted as $X_n \xrightarrow{\text{p}} X$.
- **Weak law of large numbers**: Let $X_1, X_2, \ldots, X_n$ be iid RVs with $\mathbb{E}(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$. Then, $\bar{X}_n \xrightarrow{\text{p}} \mu$.
- X_n **converges in distribution to** X if $F_{X_n}(x) \to F_X(x)$ for all x where $F_X(x)$ is continuous.
  - Equivalent to showing $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ for some bounded and continuous f.
  - If $M_{X_n}(t) \to M_X(t)$ for all t in an open neighborhood with 0, then $X_n \xrightarrow{\text{d}} X$.
  - If $X_n \xrightarrow{\text{p}} X$, then $X_n \xrightarrow{\text{d}} X$.
  - $X_n \xrightarrow{\text{p}} \mu$ iff $X_n \xrightarrow{\text{d}} \mu$.
- **Central limit theorem**, or **CLT**: Suppose $X_1, X_2, \ldots \overset{\text{iid}}{\sim} D(\mu, \sigma^2)$, where $\sigma^2 < \infty$. Then, $\sqrt{n}(X_n - \mu) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2)$.
- **First-order delta method**: Suppose $\sqrt{n}(X_n - \mu) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2)$. If $g'(\mu) \neq 0$, then $\sqrt{n}\left(g(X_n) - g(\mu)\right) \xrightarrow{\text{d}} \mathcal{N}\left(0, \sigma^2[g'(\mu)]^2\right)$.
  - If $g'(\mu) = 0$ but $g''(\mu) \neq 0$, then $n\left[g(X_n) - g(\mu)\right] \xrightarrow{\text{d}} \frac{\sigma^2}{2} g''(\mu)\chi^2_1$.
- **Example**: Suppose visits to the NCSU statistics department website follows a Poisson process with rate equal to 0.5 visits/minute. Denote by X the time (in minutes) from the last visit to the nth ($n \geq 1$) visit.
  a. Show that X has a χ^2_m distribution, where $m = 2n$.
  b. Use the CLT to show that $\frac{X-m}{\sqrt{2m}} \xrightarrow{\text{d}} \mathcal{N}(0,1)$.
  c. Use (b) to show that $\sqrt{2X} - \sqrt{2m} \xrightarrow{\text{d}} \mathcal{N}(0,1)$.

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

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

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

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

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

c.

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

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

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

- **Slutsky's theorem**: Suppose $X_n \xrightarrow{\text{d}} X$, and $Y_n \xrightarrow{\text{p}} a$. Then,
  - $Y_n X_n \xrightarrow{\text{d}} aX$.
  - $X_n + Y_n \xrightarrow{\text{d}} X + a$.
  - $\frac{X_n}{Y_n} \xrightarrow{\text{d}} \frac{X}{a}$ if $a \neq 0$.
  - $\frac{Y_n}{X_n} \xrightarrow{\text{d}} \frac{a}{X}$ if $P(X = 0) = 0$.
- **Continuous mapping theorem**: Suppose g is a continuous function, and $X_n \xrightarrow{\text{d}} X$. Then, $g(X_n) \xrightarrow{\text{d}} g(X)$.
- **Example**: Suppose that $U_1, \ldots, U_n \overset{\text{iid}}{\sim} U(0,1)$. Define $S_n = \sum_{i=1}^n U_i$ and $V_n \prod_{i=1}^n U_i$.
  a. Find the PDF of V_n.
  b. Determine $\mathbb{E}\left(\frac{U_1}{S_n}\right)$.
  c. Show that $(V_n)^{-1/S_n} \xrightarrow{\text{p}} c$, and find c.
  d. Compute $\lim_{n\to\infty} P\left(\frac{-\log(V_n)}{S_n} \geq 2\right)$.

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

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

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

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

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

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

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

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

2 ST 702: Statistical Theory II

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

2.1 Consistency and Sufficiency

Return to Table of Contents

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

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

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

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

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

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

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

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

2.2 Estimation

Return to Table of Contents

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

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

2.3 Hypothesis Tests and CIs

Return to Table of Contents

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

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

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

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

a.

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

b.

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

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

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

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

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

d.

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

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

e.

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

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

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

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

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

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

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

2.4 Introduction to Bayesian Inference

Return to Table of Contents

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

3 ST 703: Statistical Methods I

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

3.1 Hypothesis Tests and CIs

Return to Table of Contents

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

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

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

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

3.2 ANOVA Model

Return to Table of Contents

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

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

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

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

3.3 Multiple Comparisons

Return to Table of Contents

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

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

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

3.4 Two-Way Classificiation Models

Return to Table of Contents

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

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

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

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

3.5 Mixed Effects Models

Return to Table of Contents

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

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

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

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

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

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

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

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

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

3.6 Repeated Measures Designs

Return to Table of Contents

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

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

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

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

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

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

a. Complete the ANOVA table.

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

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

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

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

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

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

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

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

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

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

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

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

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

d.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4 ST 704: Statistical Methods II

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

4.1 Linear Regression

Return to Table of Contents

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

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

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

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

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

4.2 Model Assessment

Return to Table of Contents

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

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

4.3 Biased Regression and Dimension Reduction

Return to Table of Contents

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

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

4.4 GLMs

Return to Table of Contents

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

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

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

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

4.5 Mixed Models

Return to Table of Contents

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

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

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

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

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

5 ST 705: Linear Models and Variance Components

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

5.1 Linear Algebra Review

Return to Table of Contents

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

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

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

5.2 The Normal Equations

Return to Table of Contents

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

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

5.3 Estimability

Return to Table of Contents

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

5.4 Gauss-Markov/Aitken Theorem and Model Misspecification

Return to Table of Contents

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

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

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

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

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

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

5.5 Distributions/General Linear Hypotheses

Return to Table of Contents

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

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

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

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

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

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

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

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

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

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

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

where $x_1 = x_2 = 2$.

a.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.6 Cochran's Theorem

Return to Table of Contents

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

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

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

5.7 Variance Component Estimation

Return to Table of Contents

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

6 ST 740: Bayesian Statistical Inference

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

6.1 Basics of Bayesian Inference

Return to Table of Contents

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

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

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

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

6.2 Bayesian Inference

Return to Table of Contents

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

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

- A common approach to estimate estimator performance is

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

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

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

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

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

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

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

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

  In other words,

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

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

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

6.3 Prior Distributions

Return to Table of Contents

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Is this unbiased for any value of θ and λ?

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.4 MCMC and Computational Methods

Return to Table of Contents

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

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

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

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

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

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

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

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

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

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

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

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

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

  - We can often relax the strong drift condition.

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

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

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

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

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

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

6.5 Bayesian Linear Models

Return to Table of Contents

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

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

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

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

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

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

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

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

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

7 ST 793: Advanced Statistical Inference

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

7.1 Likelihood Functions

Return to Table of Contents

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The score function is

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We first note that

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

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

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

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

Verify that this is a maximizer with the second derivative.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Using the CLT for the MLE, we know that

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

Using the consistency result,

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

7.2 Asymptotics

Return to Table of Contents

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We can rewrite s_p to form

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

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

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

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

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

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

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

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

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

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

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

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

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

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

    Therefore,

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

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

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

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

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

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

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

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

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

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

- **Example**: ABAR for σ^2.

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

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

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

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

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

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

  then Q_n admits an ABAR of

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

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

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

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

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

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

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

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

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

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

(C) $\log f(y; \boldsymbol{\theta})$ has a continuous derivative wrt $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_0$ for each y in the support of $F(y; \boldsymbol{\theta}_0)$.
- Guarantees that $\hat{\boldsymbol{\theta}}_n$ is a solution to the likelihood equations.

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

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

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

(D) Uniform SLLN: $\exists h(y) : |\log f(y; \theta)| < h(y)$ for all $y \in Supp(Y)$, and $\boldsymbol{\theta}$ in a compact neighborhood of $\boldsymbol{\theta}_0$ where $\mathbb{E}_{\boldsymbol{\theta}_0}[h(Y)] < \infty$.
- In the univariate case, strong consistency relies on exactly two solutions to $|\boldsymbol{\theta} - \boldsymbol{\theta}_0| = \delta$, which is not true in higher dimensions.
- Ensures SLLN holds near $\boldsymbol{\theta}_0$.

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

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

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

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

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

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

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

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

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

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

A1:
$$F(y; \sigma) = \int_0^y \frac{2y}{\sigma^2} \exp\left\{ -\frac{y^2}{\sigma^2} \right\} dy = \cdots = \frac{2}{\sigma^2} \int_0^{\sqrt{u}} y \exp\left\{ -\frac{u}{\sigma^2} \right\} \frac{1}{2y} du = 1 - \exp\left\{ -\frac{y^2}{\sigma^2} \right\}.$$
If $\sigma_1 \neq \sigma_2$ and $y = 1$, then $F(1; \sigma_1) \neq F(2; \sigma_2)$.

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

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

A4: Define $M(Y)$ first.
$$\begin{aligned} \left| -\frac{4}{\sigma_0^3} + 24\frac{y^2}{\sigma_0^5} \right| = \left| -4\sigma_0^{-3} + 24y^2\sigma_0^{-5} \right| \leq & | -4\sigma_0^{-3}| + |24y^2\sigma_0^{-5}| \\ = & 4\sigma_0^{-3} + 24y^2\sigma_0^{-5} =: M(Y); \end{aligned}$$
$$\int_0^\infty M(Y) dF(y; \sigma_0) dy = \int_0^\infty M(Y) f(y; \sigma_0) dy = (4\sigma_0^{-3} + 24y^2\sigma_0^{-5}) \left(2\frac{y}{\sigma_0} \exp\left\{ -\frac{y^2}{\sigma_0^2} \right\} \right).$$

A5:
$$\mathbb{E}\left(-\frac{2}{\sigma} + 2\frac{y^2}{\sigma^3} \right) = -\frac{2}{\sigma} + \frac{2}{\sigma^3} \int_0^\infty 2\frac{y^3}{\sigma^2} \exp\left\{ -\frac{y^2}{\sigma^2} \right\} dy = -\frac{2}{\sigma} + \frac{2}{\sigma^3} \cdot \sigma^2 = 0.$$
$$\mathbb{E}\left[\left(\frac{\partial \ell}{\partial \sigma} \right)^2 \right] = \int_0^\infty \left[-\frac{2}{\sigma} + 2\frac{y^2}{\sigma^3} \right] \frac{2y}{\sigma^2} \exp\left\{ -\frac{y^2}{\sigma^2} \right\} dy = \cdots = \frac{4}{\sigma^2}, \text{ and}$$
$$\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \sigma^2} \right] = \int_0^\infty \left[\frac{2}{\sigma^2} + 6\frac{y^2}{\sigma^4} \right] \frac{2y}{\sigma^2} \exp\left\{ -\frac{y^2}{\sigma^2} \right\} dy = \cdots = \frac{4}{\sigma^2}. \blacksquare$$

7.3 Test Statistics and Confidence Intervals

Return to Table of Contents

Setup for this section:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

7.4 Misspecified Models and M-Estimation

Return to Table of Contents

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

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

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

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

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

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

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

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

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

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

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

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

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

Define μ_i as the ith central moment of Y.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Applying Slutsky's theorem,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

  Use MOM estimates for all remaining quantities. ■

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

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

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

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

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

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

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

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

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

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

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

7.5 Monte Carlo

Return to Table of Contents

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

    This lets us compute $SE(R_N)$.

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

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

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

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

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

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

8 ST 758: Computation for Statistical Research

Instructor: Dr. Srijan Sengupta
Semester: Spring 2025
Main Textbook: Lange, *Numerical Analysis for Statisticians*

8.1 Algorithms

Return to Table of Contents

- **Algorithm**: A list of instructions for the completion of a task.
  - Good algorithms possess finiteness (must terminate eventually), definiteness (precisely defined steps), inputs, outputs, and effectiveness (operations must be sufficiently basic).
  - **Statistical Algorithm**: An algorithm for a specific inferential task.
    - The input(s) is/are usually data and/or hyperparameters.
    - The output is often either a decision, fitted model, test result, or estimator.
- Computational problems may arise in intermediate steps of algorithms, even if the output is well-behaved.
  - **Overflow**: An error from assigning a number too large.
    - Is a problem when multiplying large numbers.
    - **Example**: $n!$ when n is large.
  - **Underflow**: An error from assigning a number indistinguishable from zero.
    - Is a problem when subtracting numbers very close in value, or dividing by a large number.
    - **Example**: $\frac{1}{n!}$ when n is large.
    - **Example**: Suppose $X \sim Bin(n, p)$. We want to calculate $P(X = 100)$ for $n = 200$ and $p = 0.99$.

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}.$$

When $n = 200$, $n!$ may be too large, so this might result in overflow. Or, if k is large, this might result in underflow. In addition, $(1-p)^{n-k}$ may be too small, which would result in underflow.

$$\begin{aligned} P(X = k) =& \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \frac{\prod_{i=1}^{n} i}{\prod_{i=1}^{k} i \prod_{i=1}^{n-k} i} p^k (1-p)^{n-k} \\ =& \frac{\prod_{i=k+1}^{n} i}{\prod_{i=1}^{n-k} i} p^k (1-p)^{n-k} = p^k \prod_{i=1}^{n-k} \frac{(k+i)(1-p)}{i}. \end{aligned}$$

This expression cancels a lot of the potentially problematic operations, and is less likely to result in underflow/overflow.

- **Computational Complexity**: The amount of resources required to execute an algorithm.
  - **Big-O**, or $\boldsymbol{O(f(n))}$: $f = O(g)$ if $\exists M > 0$ that is independent of n, and a number n_0 such that $|f(n)| \leq M|g(n)|$ for all $n \geq n_0$.
    - While M is often ignored in practice, it may still matter to us.
    - **Example**: Suppose $x, y \in \mathbb{R}^n$. $x + y \in O(n)$, and $x^T y \in O(n)$, although M may be larger for the latter case.
  - There is a trade off between computational cost and statistical error.
  - **Example**: Suppose $X_{n \times p}$. Show that $\hat{\beta}$ used in linear regression has complexity $O(np^2)$. Assume that inverting a matrix $A_{m \times m}$ is $O(m^3)$, and multiplying $A_{n \times m}$ and $B_{m \times p}$ is $O(mnp)$.

Multiplying a $p \times n$ matrix by an $n \times p$ matrix is $O(pnp) = O(np^2)$. Next, inverting the $p \times p$ matrix $X^T X$ is $O(p^3)$, but since $n > p$ is assumed, then $np^2 > p^3$, so $(X^T X)^{-1} \in O(np^2)$.
Next, we multiply the $p \times p$ matrix $(X^T X)^{-1}$ by a $p \times n$ matrix X^T, which is $O(ppn) = O(np^2)$. This means the most complex computation is still $O(np^2)$.
Lastly, multiplying this matrix by y is $O(np)$, since we can treat y as a one-column matrix, and multiplying a $p \times n$ matrix by an $n \times 1$ matrix is $O(np)$. $O(np)$ is trivially less complex than $O(np^2)$. Therefore, the entire calculation for $\hat{\beta} \in O(np^2)$. ■

- **Flops**: Basic arithmetic operations, such as adding or subtracting.
  - Computation complexity is thus the number of flops needed.
- Numbers are represented in a computer in three ways:
  - **Signed Integer**: The leading bit in the expression is 0 if the number is positive, 1 if negative.
    - There are two representations of zero: $000\ldots0$, and $100\ldots0$.
  - **One's Complement**: The negative version of a positive number negates every bit, so 0 becomes 1, and 1 becomes zero.
    - There are still two representations of zero: $000\ldots0$, and $11\ldots1$.
  - **Two's Complement**: One's complement, but then add 1.
    - There is only one representation of zero: $000\ldots0$.
- **Positional Number System**: Suppose $z > 0$. For some base B, we can decompose z into

$$z = a_k B^k + \cdots + a_1 B^1 + a_0 +_{-1} B^{-1} + \ldots$$

for some k, where $a_j \in \{0, 1, \ldots, B-1\}$.

  - Consider the representation $z = (a_k \ldots a_1 a_0 . a_{-1} \ldots)_B$. The period after a_0 is the **radix point**.
    - The radix appears implicitly in integers after the binary expression, but for floats, it must be accounted for somewhere.

Cauchy-Schwarz Inequality: $\left(\sum_{i=1}^n u_i v_i\right)^2 \le \left(\sum_{i=1}^n u_i^2\right)\left(\sum_{i=1}^n v_i^2\right)$.

  - **Example**: Prove that $\forall i,\ |x_i - \bar{x}| \le \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2} \cdot \sqrt{n-1}$.

    Define $a \in \mathbb{R}^n$ to be a vector which is 1 at the ith index, 0 otherwise. Note that $\bar{a} := \frac{1}{n}\sum_{i=1}^n a_i = \frac{1}{n}$. We will show that the squared inequality holds.

$$\begin{aligned}
(x_i - \bar{x})^2 &= \left[\sum_{i=1}^n a_i(x_i - \bar{x})\right]^2 = \left[\sum_{i=1}^n (a_i - \bar{a})(x_i - \bar{x})\right]^2 \\
&\le \left[\sum_{i=1}^n (a_i - \bar{a})^2\right]\left[\sum_{i=1}^n (x_i - \bar{x})^2\right] \quad \text{(Cauchy-Schwarz)} \\
&= \left[\sum_{i=1}^n (a_i^2 - 2a_i\bar{a} + \bar{a}^2)\right]\left[\sum_{i=1}^n (x_i - \bar{x})^2\right] \\
&= \left[1 - \frac{2}{n} + n \cdot \frac{1}{n^2}\right]\left[\sum_{i=1}^n (x_i - \bar{x})^2\right] \\
&= \left[1 - \frac{1}{n}\right]\left[\sum_{i=1}^n (x_i - \bar{x})^2\right] = \frac{(n-1)}{n}\sum_{i=1}^n (x_i - \bar{x})^2. \ \blacksquare
\end{aligned}$$

- Computers are very good at repetitiveness. We can exploit this using recurrence.
- **Recurrence Relation**: An equation that recursively defines a sequence of values.
  - Recursion can reduce the risk of overflow/underflow, and improve complexity.
  - **Example**: Suppose we observe $x_1, \ldots, x_n$, and compute $\bar{x}_n$ and s_n^2 (using $\frac{1}{n}$ for s_n^2). Then, we observe x_{n+1}. We want to update $\bar{x}_n$ and s_n^2 without repeating previous calculations.

$$\bar{x}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} x_i = \frac{1}{n+1}(n\bar{x}_n + x_{n+1}).$$

$$s_{n+1}^2 = \frac{1}{n+1}\sum_{i=1}^{n+1}(x_i - \bar{x}_{n+1})^2 = \frac{1}{n+1}\left[n s_n^2 + n(x_{n+1} - \bar{x}_n)^2 + (x_{n+1} - \bar{x}_{n+1})^2\right] = \frac{n}{n+1}s_n^2 + \frac{1}{n}(x_{n+1} - \bar{x}_{n+1})^2.$$

- **Example**: Let f_n be the number of subsets of $\{1, 2, \ldots, n\}$ that do not contain two consecutive integers. Show that $f_n = f_{n-1} - f_{n-2}$ for $n \geq 2$.
  f_0 is the number of subsets of non-consecutive integers that can be formed from $\{\}$. Only $\varnothing$ is in this set, so $f_0 = 1$.
  f_1 is the number of subsets of non-consecutive integers that can be formed from $\{1\}$. There are only two sets here that work: $\{1\}$, and $\varnothing$, so $f_1 = 2$.
  Every element of f_{n-1} is in f_n. This is because we are only adding on one term at the end of $\{1, \ldots, n\}$ and not removing anything, so if a subset didn't contain any non-consecutive integers earlier, it will still do so here. So, $f_n = f_{n-1} + a$, where $a \in \mathbb{Z}^+$.
  We now need to show that $a = f_{n-2}$. When we transition from f_{n-1} to f_n, we can only modify existing sets in f_{n-1} by adding on n to it. This would create a consecutive sequence iff $n-1$ is already in the sequence. However, if a sequence in f_{n-1} doesn't contain $n-1$, then adding on n to the sequence will still be a sequence of non-consecutive integers. The sequences in f_{n-1} that don't contain $n-1$ are f_{n-2} precisely, based on the previous logic we used to get $f_n = f_{n-1} + a$. This means that $a = f_{n-2}$, so $f_n = f_{n-1} + f_{n-2}$. ■

- **Pascal's Triangle**: $\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}$.

- **Horner's Method**: Suppose we want to compute $p(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x^1 + a_n$. Defining $b_0(x) := a_0$, $b_i(x) = xb_{i-1}(x) + a_i$.
  - For derivatives, $b_1'(x) = b_0(x)$, and $b_i'(x) = xb_{i-1}'(x) + b_{i=1}(x)$.

- **Example**: Write a recursive relationship to replace the integration in the expression $y_n(a) = \int_0^1 \frac{x^n}{x+a}dx$.

  Start with a base case.
  $$y_0(a) = \int_0^1 \frac{1}{x+a}dx = \log\left(\frac{a+1}{a}\right).$$
  Now, consider $y_n(a)$.
  $$y_n(a) = \int_0^1 \frac{x^n}{x+1}dx = \int_0^1 \frac{x^{n-1}[(x+a)-a]}{x+a}dx = \int_0^1 x^{n-1}dx - a\int_0^1 \frac{x^{n-1}}{x+a}dx = \frac{1}{n} - ay_{n-1}(a).$$

- **Example**: Suppose $X \sim Bin(n, p)$, with PMF $P(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$ for $k = 0, \ldots, n$.

  1. Prove that $\binom{n}{k}$ is an integer.
  2. Transform the CDF of X into an operation that only uses addition, subtraction, multiplication, and division. Then, compute $P(X \geq 75)$, where $n = 200$ and $p = 0.4$.
  3. Compute the Normal approximation of $P(X \geq 75)$, where $n = 200$ and $p = 0.4$, and compare.

  1. Proceed with induction. It can be assumed that $n \geq k$, and both n and k are integers.
     Base case: $n = k$.
     $$\frac{n!}{k!(n-k)!} = \frac{k!}{k!0!} = 1\checkmark$$
     Inductive step: Suppose $\frac{n!}{k!(n-k)!}$ is an integer for some $n \geq k$. Pascal's identity states that
     $$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}.$$
     We assumed the first term is an integer.
     $$\binom{n}{k} = \binom{n}{k-1} \cdot \frac{1}{k \cdot (n-k)}.$$
     Rewriting this expression yields
     $$k \cdot (n-k) \cdot \binom{n}{k} = \binom{n}{k-1}.$$
     Since n and k are assumed to be integers, this means that $\binom{n}{k-1}$ is the product of three integers, which yields an integer. Since $\binom{n+1}{k}$ consists of the sum of two integers, it must also be an integer.

2.
$$\begin{aligned}P(X=k) =& \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} = \frac{\prod_{i=n-k+1}^{n} i}{k!}\prod_{i=1}^{k} p \prod_{i=1}^{n-k}(1-p)\\ =& \frac{\prod_{i=1}^{k}(i+n-k)}{\prod_{i=1}^{k} i}\prod_{i=1}^{k} p \prod_{i=1}^{n-k}(1-p)\\ =& \left[\prod_{i=1}^{k}\frac{(i+n-k)\cdot p}{i}\right]\left[\prod_{i=1}^{n-k}(1-p)\right];\end{aligned}$$
Writing a recursive algorithm in `R` yields $P(X \geq 75) = 0.7858$.

3. Using the Normal approximation,
$$X \sim \mathcal{N}(np, np(1-p)) = \mathcal{N}(80, 48).$$
In `R`, this yielded 0.7648. This is off by roughly 2% from the value calculated previously, but with that in mind the values are still somewhat close. ■

- **Quicksort**: A sorting algorithm that chooses a **pivot** point x_k, places numbers smaller than x_k to the left, larger to the right, then repeats on the two sublists.
  - The chosen pivot dictates complexity. Amortized, this algorithm is $O(n \log n)$, but worst-case (pivots chosen are either minimum or maximum value every time) is $O(n^2)$.
  - The expected number of operations in quicksort is
$$e_n = (n-1) + \frac{2}{n}\sum_{i=1}^{n} e_{i-1}.$$
  - **Example**: Proof of $O(n \log n)$ amortized case: Let M_n be the true *position* of the pivot at step n. Suppose $M_i \to M_{i+1}$. If $x_{i+1} > x_i$, then $M_{i+1} = M_i$. If $x_{i+1} < x_i$, then $M_{i+1} = M_i + 1$. Denote N_n as the number of operations needed to sort a sequence of n numbers. We now need to find, given the position of the pivot, the number of operations needed to sort the sequence. Recall that this is a divide-and-conquer algorithm, so we will have to repeat this on the two subsequences.
$$\mathbb{E}[N_n|M_n] = (n+1) + \mathbb{E}[N_{M_{n-1}}|M_n] + \mathbb{E}[N_{n-M_n}|M_n].$$
Assuming a random order of the list, $P(M_n = i) = \frac{1}{n}$ for all $i \in \{1, \ldots, n\}$. Under this assumption,
$$\begin{aligned}e_n = \mathbb{E}[N_n] =& \sum_{i=1}^{n}\mathbb{E}[N_n|M_n = i]P(M_n = i) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[N_n|M_n = i]\\ =& \frac{1}{n}\sum_{i=1}^{n}[(n-1) + e_{i+1} + e_{n-i}];\end{aligned}$$
$$ne_n = n(n-1) + 2\sum_{i=1}^{n} e_{i-1}, \text{ and } (n-1)e_{n-1} = (n-1)(n-2) + 2\sum_{i=1}^{n-1} e_{i-1};$$
$$\begin{aligned}ne_n - (n-1)e_{n-2} =& 2(n-1) + 2e_{n-1}\\ \Rightarrow ne_n =& 2(n-1) + (n+1)e_{n-1};\end{aligned}$$
$$\begin{aligned}\frac{e_n}{n+1} = \frac{2(n-1)}{n(n+1)} + \frac{e_{n-1}}{n} = \cdots =& 2\sum_{k=1}^{n}\frac{k-1}{k(k+1)} = 2\sum_{k=1}^{n}\frac{k+1-2}{k(k+1)}\\ =& 2\sum_{k=1}^{n}\frac{1}{k} - 4\sum_{k=1}^{n}\frac{1}{k(k+1)} \approx 2\log(n) - 4.\end{aligned}$$
Therefore, $e_n \in O(n \log n)$. ■
  - **Example**: Show that Quicksort is worst-case $O(n^2)$.

We begin the same way as done previously, where we denote e_n as the expected number of operations in quicksort, N_n as the number of operations needed to sort a sequence of length n, and M_n as the true location of the pivot in the nth step.
$$e_n = \mathbb{E}[N_n|M_n] = \sum_{i=1}^{n}\mathbb{E}[N_n|M_n = i]P(M_n = i).$$

Recall that $\mathbb{E}[N_n|M_n] = (n-1) + \mathbb{E}[N_{M_N-1}|M_n] + \mathbb{E}[N_{n-M_n}|M_n]$. WLOG, consider choosing the minimum value of the unsorted list as the pivot at every step. This means that $M_n = 1$, and there are $n-1$ terms to be put after of x_1, and zero terms to be placed before it in this iteration.

$$\sum_{i=1}^{n} \mathbb{E}[N_n|M_n = i]P(M_n = i) = \mathbb{E}[N_n|M_n = 1]P(M_n = 1) = (n-1) + 0 + e_{n-1} = \sum_{i=1}^{n-1}(n-i).$$

Using the sum of integers formula,

$$\sum_{i=1}^{n-1}(n-i) = n^2 - \sum_{i=1}^{n} i = n^2 - \frac{1}{2}n(n+1) = \frac{n^2}{2} - \frac{n}{2}.$$

Therefore, given $M_n = 1$ for all n, $e_n \in O(n^2)$. ■

– **Example**: Let p be the probability that a randomly chosen permutation of n distinct numbers contains at least one pre-existing splitter in quicksort. Prove that $p = \sum_{i=1}^{n} \frac{(-1)^{i-1}}{i!} \approx 1 - e^{-1}$.

Let A_i denote the event that the ith element is in its correct sorted position. We use the inclusion-exclusion principle to obtain the event of at least one fixed point:

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{k=1}^{n}(-1)^{k-1} \sum_{1\le i_1<\cdots<i_k\le n} \left|\bigcap_{j=1}^{k} A_{i_j}\right|$$

This means there are k fixed points, so there are $n-k$ free points, which means $(n-k)!$ permutations:

$$= \sum_{k=1}^{n}(-1)^{k-1} \sum_{1\le i_1<\cdots<i_k\le n} (n-k)! = \underbrace{\sum_{k=1}^{n}(-1)^{k-1}\binom{n}{k}(n-k)!}_{:=N};$$

Since there are $n!$ total permutations,

$$p = \frac{N}{n!} = \frac{1}{n!}\sum_{k=1}^{n}(-1)^{k-1}\binom{n}{k}(n-k)! = \sum_{k=1}^{n}\frac{(-1)^{k-1}}{k!} = \sum_{i=1}^{n}\frac{(-1)^{i-1}}{i!}.$$ ■

- **Power Series**: An infinite series of the form $\sum_{n=0}^{\infty} a_n(x-c)^n$, where a_n is the coefficient of the nth term, and c is the center.
- MGFs and CGFs. If $Q(s) = \sum_{k=0}^{\infty} \frac{m_k}{k!}s^k$ is the MGF of an RV, then $P(s) = \sum_{k=0}^{\infty} \frac{c_k}{k!}s^k$ is the corresponding CGF, with $m_0 = 1$ and $c_0 = 0$.
  - **Example**: Moments of iid sums. Let ω_k be the kth moment of S_n, and μ_k be the kth moment of X_1, where $S_n = \sum_{i=1}^{n} X_i$, and $X_i \overset{\text{iid}}{\sim} F$. Then, $\omega_k = \frac{1}{k}\sum_{j=0}^{k-1}\binom{k}{j}\left[n(k-j)-j\right]\mu_{k=j}\omega_j$.
  - We can convert from m_k to c_k using $m_k = \sum_{j=0}^{k-1}\binom{k-1}{j}c_{k-j}m_j$, and the other way around using $c_k = m_k\sum_{j=1}^{k-1}\binom{k-1}{j}c_{k-j}m_j$.
- **Example**: CDF of the standard Normal. $F(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_0^x e^{-y^2/2}dy$. Rewrite this expression to a term that does not use integration.

$$\begin{aligned}\int_0^x e^{-y^2/2}dy &= \int_0^{\infty}\sum_{n=0}^{\infty}\frac{(-1)^n y^{2n}}{n!2^n}dy = \sum_{n=0}^{\infty}\frac{(-1)^n}{n!2^n}\int_0^x y^{2n}dy \\ &= \sum_{n=0}^{\infty}\frac{(-1)^n}{n!2^n}\left[\frac{y^{2n+1}}{2n+1}\right]\Bigg|_{y=0}^{x} \\ &= \sum_{n=0}^{\infty}\frac{(-1)^n x^{2n+1}}{n!2^n(2n+1)}.\end{aligned}$$

So, $F(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\int_0^x e^{-y^2/2}dy = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\sum_{n=0}^{\infty}\frac{(-1)^n x^{2n+1}}{n!2^n(2n+1)}$.

– A recursive relationship that uses differential equations lets $g(x) = e^{x^2/2}\int_0^x e^{-y^2/2}dy = \sum_{n=0}^{\infty} c_n x^{2n+1}$. $g(x)$ satisfies $g'(x) = xg(x) + 1 \implies c_n = \frac{c_{n-1}}{2n+1}$. This leads to $F(x) - \frac{1}{2} = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}g(x) = \sum_{n=0}^{\infty} a_n$, where $a_0 = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}x$, and $a_n = a_{n-1}\frac{x^2}{2n+1}$.

* This version is more stable than the one in the example.

- **Example**: **Incomplete Gamma Function**: The Gamma distribution is $P(a,bx) = \frac{1}{\Gamma(a)}\int_0^x b^a y^{a-1}e^{-by}dy \overset{z=by}{=} \frac{1}{\Gamma(a)}\int_0^{bx} z^{a-1}e^{-z}dz$. Convert this into a power series.

  Begin with IBP. Let $u = z^{a-1}$ and $dv = e^{-z}$.

$$\begin{aligned}
\frac{1}{\Gamma(a)}\int_0^{bx} z^{a-1}e^{-z}dz =& \frac{1}{\Gamma(a)}\left[-z^{a-1}e^{-z}\right]|_{z=0}^{x} + \frac{1}{\Gamma(a)}\int_0^x (a-1)z^{a-2}e^{-z}dz \\
=& -\frac{1}{\Gamma(a)}x^{a-1}e^{-x} + \frac{(a-1)\cdot\Gamma(a-1)}{\Gamma(a)}\cdot\frac{1}{\Gamma(a-1)}\int_0^x z^{(a+1)-1}e^{-z}dz \\
=& -\frac{1}{\Gamma(a)}x^{a-1}e^{-x} + \frac{\Gamma(a)}{\Gamma(a)}\cdot P(a-1,x) \\
=& -\frac{1}{\Gamma(a)}x^{a-1}e^{-x} + P(a-1,x) \\
= \cdots \implies P(a,x) =& -\sum_{n=1}^{\infty}\frac{x^{a-n}e^{-x}}{\Gamma(a-n+1)} = \sum_{n=1}^{\infty}\frac{x^{n-a}e^{-x}}{\Gamma(a-n+1)}.
\end{aligned}$$

  – Due to connections with other distributions (ex. χ^2 to Gamma), we can apply this to other distributions.

- **Example**: **Incomplete Beta Function**: The incomplete beta function is $I_X(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^x y^{a-1}(1-y)^{b-1}dy$. Convert this into power series form.

  We want $I_X(a,b) = x^a(1-x)^b\sum_{n=0}^{\infty}c_n x^n = \sum_{n=0}^{\infty}c_n x^{n+a}(1-x)^b$.

$$\begin{aligned}
\frac{d}{dx}I_X(a,b) =& \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1} = \sum_{n=0}^{\infty}c_n\left[(n+a)(1-x)-bx\right]x^{n+a-1}(1-x)^{b-1} \\
\implies \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} =& \sum_{n=0}^{\infty}c_n\left[(n+a)(1-x)-bx\right]x^n;
\end{aligned}$$

  For $n=0$, this results in $c_0 = \frac{\Gamma(a+b)}{\Gamma(a+1)\Gamma(b)}$. For $n>0$, this yields

$$c_n(n+1) - c_{n-1}(n-1+a+b) = 0 \implies c_n = \frac{c_{n-1}(n-1+a+b)}{n+1}.$$

- **Example**: Asymptotic expansions. Suppose we're interested in $\lim_{n\to\infty} f(n) = \lim_{n\to\infty}\frac{n^2+1}{n+1}$.

$$f(n) = \frac{n^2+1}{n+1} = \frac{n^2+1}{n(1+\frac{1}{n})} = \left(n+\frac{1}{n}\right)\sum_{k=0}^{\infty}\left(-\frac{1}{k}\right)^k = n - 1 - 2\sum_{k=1}^{\infty}\left(-\frac{1}{k}\right)^k.$$

  Thus, for large n, $f(n) \approx n-1$.

  – This lets us control precision with calculations.

- **Example**: Suppose we are drawing randomly from an infinite urn of red and blue marbles. In each draw, a red marble is selected with probability p.

  1. We want to have a sample with at least r red marbles and b blue marbles. Let N_{rb} be the number of samples needed to ensure this. Derive a recurrence relation to compute $\mathbb{E}[N_{rb}]$ for $r,b>0$.
  2. Derive a recurrence relation to compute $\mathbb{E}[N_{rb}^2]$ for $r,b>0$.
  3. Now, suppose we have a sampling limit. In other words, we stop sampling once we hit r red marbles and b blue marbles, or until we have collected m marbles in total, whichever comes first. Denote N_{rbm} as the number of samples drawn. Derive a recurrence relation to compute $\mathbb{E}[N_{rbm}]$ for $r,b,m>0$.

  1. We first derive the base cases. Suppose $r=0$. This means we are now finding the expected number of samples needed to obtain b blue marbles. This number is actually a negative binomial random variable: that is, $N_{0b} \sim \text{NegBin}(b, 1-p)$. Note that this parameterization is the number of *failures* before the

bth success, as opposed to the number of *trials*. Applying the mean of a negative binomial, we get that $N_{0b} = \frac{bp}{1-p}$. Similarly, $N_{r0} = \frac{rp}{1-p}$.

$$N_{0(b+1)} = \frac{(b+1)(1-p)}{p} = \frac{b+1}{b} N_{0b}. \text{ Similarly, } N_{(r+1)0} = \frac{r+1}{r} N_{r0}.$$

We note that we either draw a red or a blue marble on any given term. Thus, $N_{rb} = 1 + N_{(r-1)b}$ with probability p or $1 + N_{r(b-1)}$ with probability $1-p$. Thus,

$$\mathbb{E}[N_{rb}] = 1 + p\mathbb{E}\left[N_{(r-1)b}\right] + (1-p)\mathbb{E}\left[N_{r(b-1)}\right].$$

2. Note that

$$\mathbb{E}[N_{r0}^2] = Var(N_{r0}) + \mathbb{E}[R_{n0}]^2 = \frac{r(1-p)}{p^2} - \left[\frac{r(1-p)}{p}\right]^2 = \frac{r(1-p)}{p^2}\left[1 - r(1-p)\right].$$

Similarly, $\mathbb{E}[N_{0b}^2] = \frac{bp}{(1-p)^2}\left[1 - bp\right]$. Now, for the recursive step, $N_{rb}^2 = [1 + N_{(r-1)b}]^2 = 1 + 2N_{(r-1)b} + N_{(r-1)b}^2$ with probability p or $[1 + N_{r(b-1)}]^2 = 1 + 2N_{r(b-1)} + N_{r(b-1)}^2$ with probability $1-p$. So,

$$\mathbb{E}[N_{rb}^2] = 1 + p \cdot \left\{2\mathbb{E}[N_{(r-1)b}] + \mathbb{E}[N_{(r-1)b}^2]\right\} + (1-p)\left\{2\mathbb{E}[N_{r(b-1)}] + \mathbb{E}[N_{r(b-1)}^2]\right\}.$$

3. We first need to define base cases. When $m = 0$, then we won't draw any marbles, so $\mathbb{E}[N_{rb0}] = 0$. Next, we note that $\mathbb{E}[N_{rbm}]$ depends on whether or not $r+b \geq m$. If this is the case, then $\mathbb{E}[N_{rbm}] = m$, since we would stop at the mth marble. However, if this is not the case, then we can revert back to N_{rb}. Lastly, $\mathbb{E}[N_{00m}] = 0$ as well. Note that n, b, and m are fixed quantities.

$$N_{rbm} = \begin{cases} 1 + \mathbb{I}(\text{red})N_{(r-1)b(m-1)} + \mathbb{I}(\text{blue})N_{r(b-1)(m-1)} & m < b + r \\ m, & m \geq b + r. \end{cases}$$

Taking the expectation yields

$$\mathbb{E}[N_{rbm}] = \begin{cases} 1 + p \cdot \mathbb{E}[N_{(r-1)b(m-1)}] + (1-p)\mathbb{E}[N_{r(b-1)(m-1)}] & m < b + r \\ m, & m \geq b + r \end{cases}. \blacksquare$$

- **Example**: A family of discrete density functions $p_n(\theta)$ defined on $\{0, 1, \dots\}$ and indexed by a parameter $\theta > 0$ is a power series family if $\forall n$, $p_n(\theta) = \frac{c_n \theta^n}{g(\theta)}$, where $c_n \geq 0$ and $g(\theta) \sum_{k=0}^{\infty} c_k \theta^k$ is the appropriate normalizing constant.

1. Show that $\mu(\theta) = \frac{\theta g'(\theta)}{g(\theta)}$ and $\sigma^2(\theta) = \theta \cdot \mu'(\theta)$.
2. Now, suppose $X_1, \dots, X_m$ is a random sample from $p_n(\theta)$. Show that $S_m = \sum_{i=1}^m X_i$ follows a power series distribution with $P(S_m = n) = \frac{a_{mn}\theta^n}{g(\theta)^m}$, where a_{mn} is the coefficient of θ^n in $g(\theta)^m$.
3. If $a_{mn} = 0$ for $n < 0$, then show that $\frac{a_{m,S_m - r}}{a_{m,S_m}}$ is an unbiased estimator of θ^r.

1. Note that

$$g'(\theta) = \frac{d}{d\theta}\sum_{n=0}^{\infty} c_n \theta^n = \sum_{n=1}^{\infty} n c_n \theta^{n-1}.$$

$$\begin{aligned} \mu(\theta) = \sum_{n=0}^{\infty} n p_n(\theta) = \sum_{n=0}^{\infty} n \cdot \frac{c_n \theta^n}{g(\theta)} &= \frac{\sum_{n=1}^{\infty} n c_n \theta^n}{g(\theta)} \\ = \frac{\sum_{n=1}^{\infty} \theta \cdot n c_n \theta^{n-1}}{g(\theta)} &= \frac{\theta \sum_{n=1}^{\infty} n c_n \theta^{n-1}}{g(\theta)} = \frac{\theta g'(\theta)}{g(\theta)}; \end{aligned}$$

Next, mote that

$$\mu'(\theta) = \frac{g(\theta)[g'(\theta) + \theta g''(\theta)] - \theta g'(\theta) g'(\theta)}{g(\theta)^2} = \frac{g'(\theta) + \theta g''(\theta)}{g(\theta)} - \frac{\theta g'(\theta)^2}{g(\theta)^2}.$$

Also, we need

$$g''(\theta) = \frac{d}{d\theta}\sum_{n=1}^{\infty} n c_n \theta^{n-1} = \sum_{n=2}^{\infty} n(n-1) c_n \theta^{n-2}$$

$$\begin{aligned}
\sigma^2(\theta) =& \mu(\theta^2) - \mu(\theta)^2 = \sum_{n=0}^{\infty} n^2 p_n(\theta) - \frac{\theta^2 g'(\theta)^2}{g(\theta)^2} = \sum_{n=0}^{\infty} n^2 \cdot \frac{c_n \theta^n}{g(\theta)} - \theta \cdot \frac{\theta g'(\theta)^2}{g(\theta)^2} \\
=& \theta \left[\frac{\sum_{n=0}^{\infty} n^2 c_n \theta^{n-1}}{g(\theta)} - \frac{\theta g'(\theta)^2}{g(\theta)^2} \right]; \\
\sum_{n=0}^{\infty} n^2 c_n \theta^{n-1} =& \sum_{n=0}^{\infty} (n + n^2 - n) c_n \theta^{n-1} = \sum_{n=0}^{\infty} n c_n \theta^{n-1} + \sum_{n=0}^{\infty} n(n-1) c_n \theta^{n-1} \\
=& \sum_{n=1}^{\infty} n c_n \theta^{n-1} + \theta \sum_{n=2}^{\infty} n(n-1) c_n \theta^{n-2} = g'(\theta) + \theta g''(\theta); \\
\sigma^2(\theta) =& \theta \left[\frac{\sum_{n=0}^{\infty} n^2 c_n \theta^{n-1}}{g(\theta)} - \frac{\theta g'(\theta)^2}{g(\theta)^2} \right] = \theta \left[\frac{g'(\theta) + \theta g''(\theta)}{g(\theta)} - \frac{\theta g'(\theta)^2}{g(\theta)^2} \right] \\
=& \theta \cdot \mu'(\theta).
\end{aligned}$$

2. We will convert an S_m problem into an X_i problem, so we can apply the power series distribution.

$$\begin{aligned}
P(S_m = n) =& P\left(\sum_{i=1}^{m} X_i = n \right) = \sum_{x_1=0}^{n} P\left(\sum_{i=2}^{m} X_i = n - x_1 \right) = \dots \\
=& \sum_{x_1=0}^{n} \sum_{x_2=0}^{n-x_1} \cdots \sum_{x_{m-1}=0}^{n-\sum_{i=1}^{m-2} x_i} \prod_{i=1}^{m} \left[\frac{c_{x_i} \theta^{x_i}}{g(\theta)} \right] \\
=& \sum_{x_1=0}^{n} \sum_{x_2=0}^{n-x_1} \cdots \sum_{x_{m-1}=0}^{n-\sum_{i=1}^{m-2} x_i} \frac{\theta^{x_1 + \cdots + x_m} \prod_{i=1}^{m} c_{x_i}}{g(\theta)^m} \\
=& \frac{\theta^n}{g(\theta)^m} \underbrace{\sum_{x_1=0}^{n} \sum_{x_2=0}^{n-x_1} \cdots \sum_{x_{m-1}=0}^{n-\sum_{i=1}^{m-2} x_i} \prod_{i=1}^{m} c_{x_i}}_{=a_{mn}} \\
=& \frac{a_{mn} \theta^n}{g(\theta)^m}.
\end{aligned}$$

3.

$$\begin{aligned}
\mathbb{E}\left[\frac{a_{m,S_m - r}}{a_{m,S_m}} \right] =& \sum_{k=r}^{\infty} \mathbb{E}\left[\frac{a_{m,S_m - r}}{a_{m,S_m}} \middle| S_m = k \right] P(S_m = k) \\
=& \sum_{k=r}^{\infty} \frac{a_{m,k-r}}{a_{m,k}} P(S_m = k) \\
=& \sum_{k=r}^{\infty} \frac{a_{m,k-r}}{a_{m,k}} \cdot \frac{a_{m,k} \theta^k}{g(\theta)^m} = \sum_{k=r}^{\infty} \frac{a_{m,k-r} \theta^k}{g(\theta)^m} \\
=& \theta^r \sum_{\ell=0}^{\infty} \frac{a_{m,\ell} \theta^\ell}{g(\theta)^m} = \theta^r \sum_{\ell=0}^{\infty} P(S_m = j) \\
=& \theta^r;
\end{aligned}$$

Since $\mathbb{E}\left[\frac{a_{m,S_m - r}}{a_{m,S_m}} \right] = \theta^r$, $\frac{a_{m,S_m - r}}{a_{m,S_m}}$ is unbiased for θ^r. ■

8.2 Matrix Operations

Return to Table of Contents

- **Vector Norm**: A function $\mathbb{R}^m \to \mathbb{R}$ for $x \in \mathbb{R}^m$ that satisfy:
  1. $\|x\| \geq 0$ for all x, with equality iff $x = \underline{0}$.
  2. $\|cx\| = |c| \cdot \|x\|$ for all $c \in \mathbb{R}$ and x.
  3. Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$.

  – $\|x\|_1 = \sum_{i=1}^{m} |x_i|$, and $\|x\|_2 = \sqrt{\sum_{i=1}^{m} x_i^2}$.
  – Other valid vector norms include $\max_i |x_i|$ and $\max_{i,j} |A(i,j)|$.

- **Example**: Let $\|x\|$ be any norm. Show that $\exists k_l, k_u > 0$ such that $k_l \|x\|_1 \leq \|x\| \leq k_u \|x\|_1$ for all x.

  Proof: We first show that WLOG, we consider values of x such that $\|x\|_1 = 1$. This is because we can define $u := \frac{x}{\|x\|}$ for $x \neq \underline{0}$, which trivially has norm 1. So, if $k_l \|u\|_1 \leq \|u\| \leq k_u \|u\|_1$ holds for all x, then

  $$\|x\|_1 \cdot k_l \|u\|_1 \leq \|x\|_1 \cdot \|u\| \leq \|x\|_1 \cdot k_u \|u\|_1 \equiv k_l \|x\|_1 \leq \|x\| \leq k_u \|x\|_1$$

  Now, we show that $\|x\| \leq k_u \|x\|_1$. Let $e_1, \ldots, e_m$ be the basis vectors for $\mathbb{R}^m$, and suppose $x \in \mathbb{R}^m$. Then,

  $$x = \sum_{i=1}^{m} x_i e_i \implies \|x\| \leq \sum_{i=1}^{m} |x_i| \cdot \|e_i\| \leq \left(\max_i \|e_i\|\right) \sum_{i=1}^{m} |x_i| = \left(\max_i \|e_i\|\right) \|x\|_1 .$$

  Next, we show that $k_l \|x\|_1 \leq \|x\|$. Consider $f : \mathbb{R}^m \to \mathbb{R}^+$, where $f(u) = \|u\|$. Take any $u_1, u_2 \in \mathbb{R}^m$. $u_1 = u_1 - u_2 + u_2$, so by the triangle inequality, $\|u_1\| \leq \|u_1 - u_2\| + \|u_2\|$, so $\|u_1\| - \|u_2\| \leq \|u_1 - u_2\|$. Similarly, $\|u\|_2 - \|u_1\| \leq \|u_2 - u_1\|$ Therefore, $|f(u_1) - f(u_2)| \leq \big|\, \|u_1\| - \|u_2\| \,\big|$.
  As a result, $\big|\, \|u_1\| - \|u_2\| \,\big| \leq \|u_1 - u_2\| \leq k_u \|u_1 - u_2\|$. This means that f is uniformly continuous on the set of vectors u with $\|u\|_1 = 1$. So, f must have a minimum, and we define $k_l = \min_{u \in \{u : \|u\|_1 = 1\}} f(u)$. $k_l > 0$ by the definition of a norm, so $k_l > 0$ exists. ■

- If a sequence converges to a value in one norm, it will converge under any other norm.
  - The rate of convergence may be different, however.

- Linear algebra terms and results applicable to this section:
  - **Spectral Radius**, or $\boldsymbol{\rho(A)}$: The largest eigenvalue of A.
  - Recall that $(AB)_{ij} \sum_{k=1}^{m} A_{ik} B_{kj}$.
  - **Orthogonal Matrix**: A matrix U such that $U^T U = U U^T = I$.

- **Matrix Norm**: For square matrix A, matrix norms satisfy:
  1. All vector norm properties.
  2. Sub-multiplicativity: $\|AB\| \leq \|A\| \cdot \|B\|$.

  - **$\boldsymbol{\ell_2}$-Norm**: $\|A\|_2 = \sqrt{\rho(A^T A)}$.
  - **Frobenius Norm**: $\|A\|_F = \sqrt{\mathrm{tr}(A^T A)} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{i,j}^2}$.
    - **Example**: Show that $\|A\|_F$ is sub-multiplicative.

      $$\|AB\|_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} (AB)_{ij}^2 = \sum_{i,j=1}^{m} \left(\sum_{k=1}^{m} A_{ik} B_{kj} \right)^2 \leq \sum_{i,j=1}^{m} \left(\sum_{k=1}^{m} A_{ik}^2 \right) \left(\sum_{k=1}^{m} B_{kj}^2 \right) = \cdots = \|A\|_F^2 \cdot \|B\|_F^2 . \blacksquare$$

  - **Example**: Prove that for any nonsingular A, $\|A\|^{-1} \leq \left\|A^{-1}\right\|$.

    We know that $\|I\| \geq 1$ for any norm. Using the submultiplicity property of matrix norms,

    $$1 \leq \|I\| = \left\|A A^{-1}\right\| \leq \|A\| \cdot \left\|A^{-1}\right\| .$$

    Therefore, $\|A\|_1^{-1} \leq \left\|A^{-1}\right\|_1$. ■
  - **Example**: Define $\|A\| = \max_{ij} |A(i,j)|$ for square A. Show that this is a valid vector norm, but not a matrix norm. First, we show the properties of a vector norm are upheld:
    - $\|A\| \geq 0$. This is trivial, since we are taking the absolute value of $A(i,j)$ ✓
    - $\|A\| = 0$ iff $A = 0_{m \times m}$. If $A = 0_{m \times m}$, this is trivial. If $\|A\| = 0$, then $\max_{ij} |A(i,j)| = 0$. Again, since we are taking the absolute value of the entire expression, the absolute value is zero iff the original term is zero. In addition, we are taking the maximum of non-negative values, which can only be zero if every term is zero ✓
    - $\|cA\| = |c| \cdot \|A\|$. $\|cA\| = \max_{ij} |c \cdot A(i,j)| = |c| \cdot \max_{ij} |A(i,j)| = |c| \cdot \|A\|$ ✓
    - Triangle inequality. $\|A + B\| = \max_{ij} |A(i,j) + B(i,j)| \leq \max_{ij} (|A(i,j)| + |B(i,j)|) = \max_{ij} |A(i,j)| + \max_{ij} |B(i,j)| = \|A\| + \|B\|$ ✓

Now, we show a violation of a matrix property. The only property that could be violated is the product inequality. Consider $A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$. $AB = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}$, so $\|AB\| = \max_{ij} |AB(i,j)| = 2$, but $\|A\| \cdot \|B\| = 1 \cdot 1 = 1 < 2$, which means that in this case, $\|AB\| > \|A\| \cdot \|B\|$. Thus, the product inequality is not satisfied, and so this norm does not satisfy the conditions needed for a matrix norm. ■

- **Induced Matrix Norm**: $\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|$.
  - **Example**: Show that the two forms of the induced matrix norm are equivalent.

$$\sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \geq \sup_{x=1} \frac{\|Ax\|}{\|x\|} = \|Ax\|.$$

    Now, for any $x \neq 0$, define $u = \frac{x}{\|x\|}$, which trivially has norm 1.

$$\frac{\|Ax\|}{\|x\|} = \frac{\|A \cdot \|x\| \cdot u\|}{\|\|x\| \cdot \|u\|\|} = \frac{\|x\| \cdot \|A \cdot u\|}{\|x\| \cdot \|u\|} = \frac{\|Au\|}{\|u\|} = \|Au\| \leq \sup_{\|u\|=1} \|Au\|.$$

    Since $\sup_{\|x\|=1} \|Ax\| \leq \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \leq \sup_{\|x\|=1} \|Ax\|$, $\sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \equiv \sup_{\|x\|=1} \|Ax\|$. ■
  - $\|A\|_\infty = \max_i \sum_{j=1}^m |A(i,j)|$.
  - $\|A\|_1 = \max_j \sum_{i=1}^m |A(i,j)|$.
  - $\|A\|_2 = \sqrt{\rho(A^T A)}$.
    - **Example**: Show that $\|A\|_2 = \sqrt{\rho(A^T A)}$.

$$\|A\|_2^2 = \sup_{\|x\|_2=1} \|Ax\|_2^2 = \sup_{\|x\|_2=1} x^T A^T A x,$$

      where u_i are orthonormal. Let $x = \sum_{i=1}^m c_i u_i$, where $\sum_{i=1}^m c_i^2 = 1$.

$$x^T A^T A x = \left(\sum_{i=1}^m c_i u_i \right) \left(\sum_{i=1}^m \lambda_i u_i u_i^T \right) \left(\sum_{i=1}^m c_i u_i \right) = \sum_{i=1}^m c_i^2 \lambda_i \leq \lambda_m \sum_{i=1}^m c_i^2 = \lambda_m.$$

      This shows $\leq$, but we need $=$. Now, we let $x = u_m$.

$$\|Ax\|_2^2 = u_m^T A^T A u_m = u_m^T \lambda_m u_m u_m^T u_m = \lambda_m(1)(1) = \lambda_m.$$ ■
  - $\rho(A) \leq \|A\|$ for any induced norm $\|A\|$.
  - **Example**: For square matrix A, prove that the induced norms satisfy the following inequalities:

$$\frac{1}{\sqrt{m}} \|A\|_1 \leq \|A\|_2 \leq \sqrt{m} \|A\|_1$$
$$\frac{1}{\sqrt{m}} \|A\|_\infty \leq \|A\|_2 \leq \sqrt{m} \|A\|_\infty$$
$$\frac{1}{\sqrt{m}} \|A\|_F \leq \|A\|_2 \leq \|A\|_F.$$

    First, we show that $\frac{1}{\sqrt{m}} \|A\|_1 \leq \|A\|_2 \leq \sqrt{m} \|A\|_1$. Using the property that $\|x\|_p \leq m^{\frac{1}{p} - \frac{1}{q}} \|x\|_q$ for $p < q \in \mathbb{N}$,

$$\|A\|_1 \leq m^{1-1/2} \|A\|_2 \implies \frac{1}{\sqrt{m}} \|A\|_1 \leq \|A\|_2.$$

    In addition, since $\|A\|_q \leq \|A\|_p$ for $p < q \in \mathbb{N}$, $\|A\|_2 \leq \|A\|_1 \leq \sqrt{m} \|A\|_1$ for $m \geq 1$ (which is the only valid way to define an $m \times m$ matrix). Thus, $\frac{1}{\sqrt{m}} \|A\|_1 \leq \|A\|_2 \leq \sqrt{m} \|A\|_1$. Next, we show that $\frac{1}{\sqrt{m}} \|A\|_\infty \leq \|A\|_2 \leq \sqrt{m} \|A\|_\infty$. Using the earlier property yields a similar result that

$$\|A\|_2 \leq m^{0-1/2} \|A\|_\infty \implies \|A\|_2 \leq \sqrt{m} \|A\|_\infty.$$

In addition, once again using an earlier property, $\frac{1}{\sqrt{m}}\|A\|_\infty \leq \|A\|_\infty \leq \|A\|_2$ for $m \geq 1$. Thus, $\frac{1}{\sqrt{m}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{m}\|A\|_\infty$. Lastly, we show that $\frac{1}{\sqrt{m}}\|A\|_F \leq \|A\|_2 \leq \|A\|_F$. Denote $\lambda_1 \leq \cdots \leq \lambda_m$ as the eigenvalues of A^TA. Since A^TA is symmetric, $\lambda_1 \geq 0$.

$$\|A\|_F = \sqrt{\text{trace}(A^TA)} = \sqrt{\sum_{i=1}^{m}\lambda_i} \leq \sqrt{\sum_{i=1}^{m}\lambda_m} = \sqrt{m \cdot \rho(A^TA)} = \sqrt{m}\|A\|_2 .$$

Thus, $\frac{1}{\sqrt{m}}\|A\|_F \leq \|A\|_2$.

$$\|A\|_2 = \sqrt{\rho(A^TA)} = \sqrt{\lambda_m} \leq \sqrt{\sum_{i=1}^{m}\lambda_i} = \|A\|_F .$$

Thus, $\frac{1}{\sqrt{m}}\|A\|_F \leq \|A\|_2 \leq \|A\|_F$. ■

- **Example**: Show that for any induced matrix norm for square matrix A, $\rho(A) = \lim_{n\to\infty}\|A^n\|^{1/n}$.

  First, we show that $\rho(A) \leq \lim_{n\to\infty}\|A^n\|^{1/n}$. Define $\tilde{A} := \frac{1}{\rho(A)-\epsilon}A$ for all $\epsilon > 0$. As a result, $\rho(\tilde{A}) > 1$, and $\lim_{n\to\infty}\tilde{A}^k = 0_{m\times m}$, where $A \in \mathbb{R}^{m\times m}$. Since $\tilde{A}$ is convergent, then $\exists n_m$ such that $\forall n \geq n_m$, $\left\|\tilde{A}^n\right\| > 1$. Thus, $\forall n \geq n_m$, $\|A^n\| > (\rho(A)-\epsilon)^n \equiv \|A^n\|^{1/n} > \rho(A) - \epsilon$.
  Next, we show that $\rho(A) \geq \lim_{n\to\infty}\|A^n\|^{1/n}$. Similar to the previous case, define $A' := \frac{1}{\rho(A)+\epsilon}A$ for all $\epsilon > 0$. As a result, $\rho(A') < 1$, and $\lim_{n\to\infty}(A')^k = 0_{m\times m}$, where $A \in \mathbb{R}^{m\times m}$. Since A' is convergent, then $\exists n_M$ such that $\forall n \geq n_M$, $\|(A')^n\| < 1$. Thus, $\forall n \geq n_M$, $\|A^n\| < (\rho(A)+\epsilon)^n \equiv \|A^n\|^{1/n} < \rho(A) + \epsilon$. Take $n_0 = \max\{n_m, n_M\}$. Then, since $\rho(A) - \epsilon < \|A^n\|^{1/n} < \rho(A) + \epsilon$, then by definition, $\lim_{n\to\infty}\|A^n\|^{1/n} = \rho(A)$. ■

- **Orthogonal Transformation**: Pre or post-multiply a vector/matrix by orthogonal matrix O.
  - **Example**: Show orthogonal transformations are inner-product preserving.

    $$(Ou)^T(Ov) = u^TO^TOv = uv. \blacksquare$$

  - **Example**: Show orthogonal transformations are norm-preserving for the Frobenius norm.

    $$\|OA\|_F^2 = \text{tr}\{A^TO^TOA\} = \text{tr}\{A^TA\} = \|A\|_F^2 . \blacksquare$$

  - **Example**: Show orthogonal transformations are 2-norm preserving.

    We show this for vectors x and matrices A. Suppose O is orthogonal.

    $$\|Ox\|_2^2 = (Ox)^T(Ox) = x^TO^TOx = x^Tx = \|x\|_2^2 .$$

    This means that, for vectors, orthogonal transformations are 2-norm preserving.

    $$\|WA\|_2 = \sup_{\|x\|_2=1}\|W(Ax)\|_2 = \sup_{\|x\|_2=1}\|Ax\|_2 = \|A\|_2 ,$$

    Since 2-norm of orthogonal transformations is preserved for vectors. ■

- **Condition Number**, or ***cond*(*A*)**: $cond(A) = \|A\|\left\|A^{-1}\right\|$.
  - Suppose $Ax = b$ for invertible A. $cond(A)$ determines the stability of solving for x. That is, given $x = A^{-1}b$, how stable is this calculation wrt b?
  - In regression, we can't control y, but we may be able to control x, so it is very important to choose a stable x.
  - $cond(A)$ is the worst-case stability of the solution. That is, $\frac{\|\Delta x\|}{\|x\|} \leq cond(A) \cdot \frac{\|\Delta b\|}{\|b\|}$.
  - **Example**: Assuming $A + \Delta A$, show that

    $$\frac{\left\|(A+\Delta A)^{-1} - A^{-1}\right\|}{\|(A+\Delta A)^{-1}\|} \leq \text{cond}(A) \cdot \frac{\|\Delta A\|}{\|A\|},$$

where $\|A\|$ is the induced matrix norm of square A.

$$cond(A)\frac{\|\Delta A\|}{\|A\|} = \|A\| \cdot \left\|A^{-1}\right\| \frac{\|\Delta A\|}{\|A\|} = \left\|A^{-1}\right\| \|\Delta A\|.$$

We start by leveraging the fact that $(A + \Delta A)$ and A are invertible.

$$\begin{aligned}
(A+\Delta A)(A+\Delta A)^{-1} =& I;\\
A(A+\Delta A)^{-1} =& I - \Delta A(A+\Delta A)^{-1};\\
A^{-1}\left[A(A+\Delta A)^{-1}\right] =& A^{-1}\left[I - \Delta A(A+\Delta A)^{-1}\right];\\
(A+\Delta A)^{-1} - A^{-1} =& -A^{-1}\Delta A(A+\Delta A)^{-1};\\
\frac{(A+\Delta A)^{-1} - A^{-1}}{\|(A+\Delta A)^{-1}\|} =& -\frac{A^{-1}\Delta A(A+\Delta A)^{-1}}{\|(A+\Delta A)^{-1}\|};\\
\frac{\left\|(A+\Delta A)^{-1} - A^{-1}\right\|}{\|(A+\Delta A)^{-1}\|} =& \frac{\left\|(-1)A^{-1}\Delta A(A+\Delta A)^{-1}\right\|}{\|(A+\Delta A)^{-1}\|} = |-1| \cdot \frac{\left\|A^{-1}\Delta A(A+\Delta A)^{-1}\right\|}{\|(A+\Delta A)^{-1}\|}\\
\leq& \left\|A^{-1}\Delta A\right\| \cdot \frac{\left\|(A+\Delta A)^{-1}\right\|}{\|(A+\Delta A)^{-1}\|} = \left\|A^{-1}\Delta A\right\| \leq \left\|A^{-1}\right\| \|\Delta A\|\\
=& cond(A)\frac{\|\Delta A\|}{\|A\|}. \blacksquare
\end{aligned}$$

- **Example**: Let A and W be $m \times m$ matrices, and W be an orthogonal matrix.
  1. Show that $\|A\|_F = \|WA\|_F$.
  2. Show that $\text{cond}(A) = \text{cond}(WA)$ in the 2-norm.

  1. $$\|WA\|_F^2 = \text{tr}\{(WA)^T WA\} = \text{tr}\{A^T W^T WA\} = \text{tr}\{A^T A\} = \|A\|_F^2.$$
  2. $cond(WA) = \|WA\|_2 \left\|(WA)^{-1}\right\|_2$. Using the fact that the 2-norm for orthogonal transformations is preserved, $\|WA\|_2 \left\|(WA)^{-1}\right\|_2 = \|A\|_2 \left\|(WA)^{-1}\right\|_2$. Once again using this fact,

     $$\left\|(WA)^{-1}\right\|_2 = \left\|A^{-1}W^T\right\|_2 = \left\|A^{-1}\right\|_2 \implies \|A\|_2 \left\|(WA)^{-1}\right\|_2 = \|A\|_2 \left\|A^{-1}\right\|_2;$$

     Therefore, $cond(WA) = cond(A)$. $\blacksquare$

- **Example**: Prove that $\|A\|_F = \left\|U^T AU\right\|_F$, and $\text{tr}(A) = \text{tr}(U^T AU)$.

  First, we show that $\|A\|_F = \left\|U^T AU\right\|_F$. This proof uses the cyclic property of the trace.

  $$\left\|U^T AU\right\|_F^2 = \text{tr}\left\{(U^T AU)^T * U^T AU\right\} = \text{tr}\left\{U^T A^T AU\right\} = \text{tr}\left\{UU^T A^T A\right\} = \text{tr}\left\{A^T A\right\} = \|A\|_F^2.$$

  Next, we show the traces are equal. This is a straightforward application of the cyclic property of the trace.

  $$\text{tr}\left\{U^T AU\right\} = \text{tr}\left\{UU^T A\right\} = \text{tr}\{A\}. \blacksquare$$

- **Woodbury's Formula**: Suppose A is invertible, and U, V are rank-deficient and non-square matrices.

  $$(A + UV^T) = A^{-1} - A^{-1}U(I_p + V^T A^{-1}U)^{-1}V^T A^{-1}.$$

  - Since $(I_p + V^T A^{-1}U)$ is $d \times d$, finding the inverse might be cheaper than inverting $(A + UV^T)$ directly.

- **Example**: Suppose positive-definite $A_{m\times m}$ has inverse B. Show that $a_{ii}^{-1} \leq b_{ii}$ for all $i \in \{1, \ldots, m\}$, with equality iff $a_{ij} = a_{ji} = 0$ for all $j \neq i$. Let e_i denote the standard basis vector. Also note that since A and B are both positive-definite, there exists unique $A^{1/2}$ and $B^{1/2}$'s such that $A^{1/2}A^{1/2} = A$, same for $B^{1/2}B^{1/2} = B$. Apply the Cauchy-Schwarz inequality.

  $$1 = (e_i^T A^{1/2}B^{1/2}e_i)^2 = (e_i^T e_i)^2 \leq (e_i^T Ae_i)(e_i^T Be_i) = a_{ii}b_{ii} \implies a_{ii}^{-1} \leq b_{ii}.$$

  Next, we show the equality statements. Suppose $a_{ij} = 0$ for all $i \neq j$. This means that A is a diagonal matrix, and since A is invertible, none of the diagonal elements are zero. Therefore, $AB = I \equiv a_{ii} \cdot b_{ii} = 1$ for all $i \in \{1, \ldots, m\}$, which means $a_{ii}^{-1} = b_{ii}$. Now, suppose $a_{ii}^{-1} = b_{ii}$ for all $i \in \{1, \ldots, m\}$, and $i \neq j$. The Cauchy-Schwarz statement is an equality iff u and v are linearly dependent, where $u = A^{1/2}e_i$ and $v = B^{1/2}e_i$. This means that $\exists c_i$ such that

  $$A^{1/2}e_i = c_i \cdot B^{1/2}e_i \implies Ae_i = c_i e_i.$$

This means that e_i is an eigenvector of A (with corresponding eigenvalue c_i). Since e_i is a basis vector for A, this means that the ith column of A is only nonzero on the diagonal. Since the choice of i was arbitrary, this will hold for all $i \in \{1, \dots, m\}$, so every column of A is only nonzero on its diagonal. Therefore, $a_{ij} = 0$ for all $i \neq j$. ■

- **Example**: Show that If A is positive-definite, that A^{-1} is also positive-definite.

  Since A is positive-definite, $x^T Ax > 0$ for all x. Define $y = A^T x$.

  $$0 < x^T Ax = x^T A(A^{-1}A)x = (A^T x)^T A^{-1}(A^T x) = y^T A^{-1} y;$$

  Therefore, by definition, A^{-1} is positive-definite. ■

- Suppose we want to find $x = (I - B)^{-1}$, where $I - B$ is nonsingular. We can define $x_{n+1} = Bx_n + w$, which converges if $\|B\| < 1$.
  - **Example**: Show that $x_{n+1} = Bx_n + w$ converges if $\|B\| < 1$.

    $$\|x_{n+1} - x_n\| = \|B(x_n - x_{n-1})\| \leq \|B\| \|x_n - x_{n-1}\| \leq \|B\|^2 \|x_{n-1} - x_{n-2}\| \leq \cdots \leq \|B\|^n \cdot \|x_1 - x_0\|.$$

    Thus, if $\|B\| < 1$, then $\|B\|^n \to 0$. Since $x_1 - x_0$ is finite, $\|x_{n+1} - x_n\|$ must converge. ■

- **Jacobi's Method**: Solve $Ax = b$ by Defining $B = I - D^{-1}A$, where $D = \text{diag}(A)$.
  - Requires a diagonally dominant A; that is, $|A_{ii}| > \sum_{j \neq i} |A_{ij}|$ for all i.
  - This method converges, since

    $$\|B\|_\infty = \max_i \sum_{j=1}^{m} |A_{ij}| = \max_i \sum_{j \neq i} \left| \frac{A_{ij}}{A_{ii}} \right| < 1.$$

- **Landweber's Iteration Scheme**: We are solving $Ax = b$. For some small and positive ϵ, we can set $B = I - \epsilon A^T A$.
  - As long as $1 - \epsilon \cdot \rho(A) > -1$, $\lambda_i \in (-1, 1) \implies \left\| I - \epsilon A^T A \right\|_2 < 1$.

- **Power Method**: Consider an MC $(X_0, X_1, \dots)$ with m states $s_1, \dots, s_m$. The transition matrix P governs the evolution of the chain, and we often want to find the equilibrium distribution x that satisfies $P^T x = x$. Start with some initial point x_0, and use the recurrence relation $x_{n+1} = P^T x_n$.

- **Sweep Operator**: For symmetric A with non-zero diagonals, we sweep on the kth diagonal a_{kk} to create a new matrix $\hat{A}$, where for $i, j \neq k$,:

  $$\hat{a}_{kk} = -\frac{1}{a_{kk}}, \; \hat{a}_{ik} = \frac{a_{ik}}{a_{kk}},$$
  $$\hat{a}_{kj} = \frac{a_{kj}}{a_{kk}}, \; \hat{a}_{ij} = a_{ij} - \frac{a_{ik}a_{kj}}{a_{kk}}.$$

  - Sweeping exploits symmetry in symmetrical matrices.
  - We can "reverse sweep" on the kth diagonal of $\hat{A}$ to obtain A.
  - Sweep operator is $O(n^2)$.
    - Exploiting symmetry can mitigate this, but won't change the complexity.
  - Suppose we have a U, V such that $V = UA$. After sweeping on a_{kk}, we form $\hat{V} = \hat{U}\hat{A}$, where $\hat{U} = U$ except for $\hat{U}_{.k} = V_{.k}$, and $\hat{V} = V$ except for $\hat{V}_{.k} = -U_{.k}$.
    - **Example**: Prove the above claim.

Consider $1 \leq j$ and $\ell \leq m$.

$$v_{j\ell} = \sum_{i=1}^{m} u_{ji}A_{i\ell} = u_{jk}A_{k\ell} + \sum_{i\neq k} ujiA_{i\ell} \implies u_{jk}A_{k\ell} = v_{j\ell} - \sum_{i\neq k} u_{ji}A_{i\ell};$$

Suppose $\ell = k$;

$$\begin{aligned}\hat{v}_{jk} = &- u_{jk} = -\frac{1}{a_{kk}}\left(v_{jk} - \sum_{i\neq k} u_{ji}A_{ik}\right)\\ =&\hat{A}_{kk}\hat{u}_{jk} + \sum_{i\neq k}\hat{u}_{ji}\hat{A}_{ik} = \sum_{i=1}^{m}\hat{u}_{ji}\hat{A}_{ik};\end{aligned}$$

Now, set $\ell \neq k$;

$$\begin{aligned}\hat{v}_{j\ell} =&u_{jk}A_{k\ell} + \sum_{i\neq k} u_{ji}A_{i\ell} = \left(v_{jk} - \sum_{i\neq k} u_{ji}A_{ik}\right)\cdot\frac{A_{k\ell}}{A_{kk}} + \sum_{i\neq k} u_{ji}A_{i\ell}\\ =&v_{jk}\cdot\frac{A_{k\ell}}{A_{kk}} + \sum_{i\neq k} u_{ji}\left(A_{i\ell} - \frac{A_{ik}A_{k\ell}}{A_{kk}}\right)\\ =&\hat{u}_{jk}\hat{A}_{k\ell} + \sum_{i\neq k}\hat{u}_{ji}\hat{A}_{i\ell} = \sum_{i=1}^{m}\hat{u}_{ji}\hat{A}_{i\ell}.\end{aligned}$$

Therefore, $\hat{v}_{j\ell} = \sum_{i=1}^{m}\hat{u}_{ji}\hat{A}_{i\ell}$, regardless of whether or not $\ell = k$, so $\hat{V} = \hat{U}\hat{A}$. ■

– Partition $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22}\end{pmatrix}$. Sweeping on A_{11} yields $\hat{A} = \begin{pmatrix} -A_{11}^{-1} & A_{11}^{-1}A_{12} \\ A_{21}A_{11}^{-1} & A_{22} - A_{21}A_{11}^{-1}A_{12}\end{pmatrix}$.

  * We sweep on several columns by recursively sweeping on one column at a time.

– A symmetric matrix A is positive definite iff each diagonal entry can be swept in succession and is positive until swept.

– We can obtain $\hat{\beta}$ in MLR. Let $A = \begin{pmatrix} X^TX & X^Ty \\ y^TX & y^Ty\end{pmatrix}$. Sweeping on X^TX yields $\hat{A} = \begin{pmatrix} -\frac{1}{\sigma^2}Var(\hat{\beta}) & \hat{\beta} \\ \hat{\beta}^T & \text{SSE}\end{pmatrix}$.

  * This operation is $O(p^3)$.

– **Example**: The jackknife method of regression analysis can be implemented by replacing

$$\begin{pmatrix} X^TX & X^Ty \\ y^TX & y^Ty\end{pmatrix} \text{ with } \begin{pmatrix} X^TX & X^T & X^Ty \\ X & I_p & y \\ y^TX & y^T & y^Ty\end{pmatrix},$$

and then sweeping on X^TX, and then the $(q+k)$th diagonal entry for some $k \in \{1,\ldots,p\}$.

1. Let h_{kk} be the kth diagonal entry of P_X. Show that $y_k - \hat{y}^{-k} = \frac{y_k - \hat{y}_k}{1-h_{kk}}$.
2. Show that this action yields the necessary ingredients for regression that omits the kth observation. Use the fact that the additional sweep is equivalent to replacing the kth regression equation $y_k = \sum_{\ell=1}^{q} x_{k\ell}\beta_\ell + e_k$ by $y_k = \sum_{\ell=1}^{q} x_{k\ell}\beta_\ell + \beta_{q+k} + e_k$, where the other β's are untouched. This means we can adjust β_{q+k} to obtain a perfect fit for y_k, so the estimates $\hat{\beta}_1,\ldots,\hat{\beta}_q$ after the additional sweep depend only on y_i for $i \neq k$.

1. $$y_k - \hat{y}_k^{-k} = y_k - X_{k\cdot}\hat{\beta}^{-k}, \text{ where } \hat{\beta}^{-k} = (X_{-k}^TX_{-k})^{-1}X_{-k}^Ty_{-k};$$

We first apply Woodbury's formula to obtain $(X_{-k}^T X_{-k})^{-1}$.

$$
\begin{aligned}
X_{-k}^T X_{-k} =& \underbrace{X^T X}_{A} \underbrace{- X_{k.}^T}_{u} \underbrace{X_{k.}}_{v^T};\\
(X^T X - X_{k.}^T X_{k.})^{-1} =& (X^T X)^{-1} - \frac{1}{1 + X_{k.}(X^T X)^{-1}(-X_{k.}^T)} (X^T X)^{-1}(-X_{k.}^T) X_{k.} (X^T X)^{-1}\\
=& (X^T X)^{-1} + \frac{(X^T X)^{-1} X_{k.}^T X_{k.} (X^T X)^{-1}}{1 - h_{kk}}\\
=& (X^T X)^{-1} + \frac{(X^T X)^{-1} X_{k.}^T X_{k.} (X^T X)^{-1}}{1 - h_{kk}};\\
y_k - \hat{y}_k^{-k} =& y_k - X_{k.} \left\{ (X^T X)^{-1} + \frac{(X^T X)^{-1} X_{k.}^T X_{k.} (X^T X)^{-1}}{1 - h_{kk}} \right\} X_{-k}^T y_{-k}\\
=& y_k - X_{k.}(X^T X)^{-1}(X^T y - X_{k.}^T y_k) - \frac{X_{.k}(X^T X)^{-1} X_{k.}^T X_{k.} (X^T X)^{-1}(X^T y - X_{k.}^T y_k)}{1 - h_{kk}}\\
=& y_k - X_{k.}\hat{\beta} + h_{kk} y_k - \frac{h_{kk} X_{k.}\hat{\beta} - h_{kk}^2 y_k}{1 - h_{kk}}\\
=& \frac{y_k - \hat{y}_k + h_{kk} y_k - h_{kk} y_k + h_{kk}\hat{y}_k - h_{kk}^2 y_k - h_{kk}\hat{y}_k + h_{kk}^2 y_k}{1 - h_{kk}}\\
=& \frac{y_k - \hat{y}_k}{1 - h_{kk}}. \ \blacksquare
\end{aligned}
$$

2. Define $A := \begin{pmatrix} X^T X & \begin{pmatrix} X^T & X^T y \end{pmatrix} \\ \begin{pmatrix} X \\ y^T X \end{pmatrix} & \begin{pmatrix} I_p & y \\ y^T & y^T y \end{pmatrix} \end{pmatrix}$. Similar to OLS, we can then let $V = A$ and $U = I_{2p+1}$.

First, we sweep on the columns of X^TX.

$$\begin{pmatrix} X^TX & \begin{pmatrix} X^T & X^Ty \end{pmatrix} \\ \begin{pmatrix} X \\ y^TX \end{pmatrix} & \begin{pmatrix} I_p & y \\ y^T & y^Ty \end{pmatrix} \end{pmatrix} = \begin{pmatrix} I_p & 0 \\ 0 & I_{p+1} \end{pmatrix} \begin{pmatrix} X^TX & \begin{pmatrix} X^T & X^Ty \end{pmatrix} \\ \begin{pmatrix} X \\ y^TX \end{pmatrix} & \begin{pmatrix} I_p & y \\ y^T & y^Ty \end{pmatrix} \end{pmatrix}$$

$$\rightarrow \hat{A} = \begin{pmatrix} -(X^TX)^{-1} & (X^TX)^{-1}\begin{pmatrix} X^T & X^Ty \end{pmatrix} \\ \begin{pmatrix} X \\ y^TX \end{pmatrix}(X^TX)^{-1} & \begin{pmatrix} I_p & y \\ y^T & y^Ty \end{pmatrix} - \begin{pmatrix} X \\ y^TX \end{pmatrix}(X^TX)^{-1}\begin{pmatrix} X^T & X^Ty \end{pmatrix} \end{pmatrix}$$

$$= \begin{pmatrix} -(X^TX)^{-1} & \begin{pmatrix} (X^TX)^{-1}X^T & (X^TX)^{-1}X^Ty \end{pmatrix} \\ \begin{pmatrix} X(X^TX)^{-1} \\ y^TX(X^TX)^{-1} \end{pmatrix} & \begin{pmatrix} I_p & y \\ y^T & y^Ty \end{pmatrix} - \begin{pmatrix} X(X^TX)^{-1} \\ y^TX(X^TX)^{-1} \end{pmatrix}\begin{pmatrix} X^T & X^Ty \end{pmatrix} \end{pmatrix}$$

$$= \begin{pmatrix} -(X^TX)^{-1} & \begin{pmatrix} (X^TX)^{-1}X^T & \hat{\beta} \end{pmatrix} \\ \begin{pmatrix} X(X^TX)^{-1} \\ \hat{\beta}^T \end{pmatrix} & \begin{pmatrix} I_p & y \\ y^T & y^Ty \end{pmatrix} - \begin{pmatrix} P_X & X\hat{\beta} \\ (X\hat{\beta})^T & y^TP_Xy \end{pmatrix} \end{pmatrix}$$

$$= \begin{pmatrix} -(X^TX)^{-1} & \begin{pmatrix} (X^TX)^{-1}X^T & \hat{\beta} \end{pmatrix} \\ \begin{pmatrix} X(X^TX)^{-1} \\ \hat{\beta}^T \end{pmatrix} & \begin{pmatrix} I_p - P_X & y - X\hat{\beta} \\ (y - X\hat{\beta})^T & y^Ty - y^TP_Xy \end{pmatrix} \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{1}{\sigma^2}Var(\hat{\beta}) & (X^TX)^{-1}X^T & \hat{\beta} \\ X(X^TX)^{-1} & I_p - P_X & \hat{e} \\ \hat{\beta}^T & \hat{e}^T & \text{SSE} \end{pmatrix}$$

Now, we sweep on the $(q+k)$th diagonal. Applying the fact, sweeping on this diagonal yields $\hat{\beta}$ estimates that do not depend on k. Therefore, sweeping on X^TX, and then the $(q+k)$th diagonal, we are able to obtain $\hat{\beta}^{-k}$, which are the β estimates that omit the kth observation.
Sweeping on the $(q+k)$th diagonal will also let us obtain $X_{-k}^TX_{-k}$. Lastly, we use the formula in the previous part to obtain $\hat{y}_k^{-k}$.

- **Gaussian Elimination**: Suppose A is triangular, and we want to solve $Ax = b$. Decompose $A = LU$, and express $LUx = b$. Solve for y in $Ly = b$, then solve for x where $Ux = y$.
  - If A is not triangular, then for any non-singular P we have $Ax = b \equiv PAx = Pb$, which we then solve $PAx = Pb$ as before.
  - Complexity is $O(p^3)$.
- **Cholesky Decomposition**: Suppose A is positive definite. Then, there exists a unique lower-triangular matrix L such that $A = LL^T = LU$.
  - **Example**: Prove the above claim.

    Proceed with induction. Base case: $m = 1$. $a = \sqrt{a} \cdot \sqrt{a}$ ✓

Inductive Step: Suppose $A = LL^T$ for some m.

$$A_{(m+1)\times(m+1)} = \begin{pmatrix} \ell_{11} & 0^T \\ \ell & L_{22} \end{pmatrix} \begin{pmatrix} \ell_{11} & \ell^T \\ 0 & L_{22}^T \end{pmatrix};$$
$$A_{11} = \ell_{11}^2;\ \ell = \frac{a}{\sqrt{A_{11}}};$$
$$A_{22} = \ell\ell^T + L_{22}L_{22}^T \implies L_{22}L_{22}^T = A_{22} - a_{11}^T aa^T;$$

This means that we have a positive-definite A. ■

– We can obtain $\hat{e}^2$ in MLR.

$$(X,y)^T(X,y) = \begin{pmatrix} L & 0 \\ \ell^T & d \end{pmatrix} \begin{pmatrix} L^T & \ell \\ 0^T & d \end{pmatrix} \implies d^2 = \|y - \hat{y}\|_2^2.$$

- **Gram-Schmidt Orthogonalization**: Let $u_1 = \frac{x_1}{\|x_1\|_2}$ and $u_k = \frac{v_k}{\|v_k\|_2}$, where $v_k = x_k - \sum_{j=1}^{k-1}(u_j^T X_k)u_j$. Then, let $Q = (u_1, \ldots, u_p)$, and $R_{jk} = u_j^T X_k$, with $R_{kk} = \|v_k\|_2$.

  – Operates directly on X, as opposed to $X^T X$.

  – **Example**: Show that $X = QR$.

$$\begin{aligned} X_{ik} =& \sum_{j=1}^{k} u_{ij}R_{jk}; \\ v_k =& X_k - \sum_{j=1}^{k-1}(u_j^T X_k)u_j, \text{ and } = u_k \|v_k\|_2 \\ \implies X_k =& v_k + \sum_{j=1}^{k-1}(u_j^T X_k)u_j = u_k R_{kk} + \sum_{j=1}^{k-1} u_j R_{kk} \\ =& \sum_{j=1}^{m} u_j R_{jk} = QR. \ \blacksquare \end{aligned}$$

  – For MLR, we would use $X = QR$, and $X^T X = R^T Q^T QR = R^T R$, and $X^T = R^T Q^T$.

- **Householder Reflections**: To construct the QR decomposition of X, we can carry out a sequence of Householder reflections $H_1, \ldots, H_{n-1}$ to form $H_{n-1} \ldots H_1 X = OX = (R^T, \mathbf{0}^T)^T$, where $H = I - 2uu^T$ is orthogonal and symmetric.

  – **Example**: Define $H(u) = Q = 1 - 2\frac{uu^T}{u^T u}$, and let $Q = H(x)$, where $\|x\|_2 = 1$.

    1. Show that Q is symmetric and orthogonal.
    2. Determine the eigenvalues of Q.
    3. Now, suppose $Q = I - 2\frac{uu^T}{u^T u}$ for arbitrary $u \neq \mathbf{0}$. Show that $H(u)z = z$ for z that is perpendicular to u.
    4. Let $x = z + u^T xu$, where z is perpendicular to u. Show that $H(u)x = z - u^T xu$.

    1. First, show that Q is symmetric, or $Q^T = Q$.

$$Q^T = (I - 2xx^T)^T = I^T - 2(xx^T)^T = I = 2xx^T = Q.$$

    Next, show Q is orthogonal, or $QQ^T = QQ = Q^2 = I$.

$$\begin{aligned} QQ =& \left(I - 2xx^T\right)\left(I - 2xx^T\right) = I - I(2xx^T) - 2xx^T(I) + 2xx^T(2xx^T) \\ =& I - 4xx^T + 4x(x^T x)x = I - 4xx^T + 4xx^T = I \implies Q^T = Q^{-1}. \end{aligned}$$

    2. Since Q is symmetric and orthogonal,

$$Qv = \lambda v \implies Q^2 v = Iv = v = \lambda^2 v.$$

    $\lambda = \pm 1$ satisfies $\lambda^2 v = v$, so the eigenvalues of Q must be ± 1.

3.
$$H(u)z = \left(I - 2\frac{uu^T}{u^Tu}\right)z = z - 2\frac{u(u^Tz)}{u^Tu} = z - 2\frac{u(0)}{u^Tu} = z - \mathbf{0} = z.$$

4. Note that $u^Tx \in \mathbb{R}$.

$$\begin{aligned} H(u)x = \left(I - 2\frac{uu^T}{u^Tu}\right)(z + u^Txu) &= z + u^Txu - 2\frac{uu^T(u^Tx)u}{u^Tu} \\ &= z + u^Txu - 2\frac{(u^Tx)u(u^Tu)}{u^Tu} = z + u^Txu - 2u^Txu \\ &= z - u^Txu. \ \blacksquare \end{aligned}$$

- **Sherman-Morrison Formula**: Suppose $A_{m\times m}$ is invertible. Then, $A+uv^T$ is invertible iff $1+v^TA^{-1}u \neq 0$, and
$$(A + uv^T)^{-1} = A^{-1} - \frac{1}{1 + v^TA^{-1}u}A^{-1}uv^TA^{-1}.$$
  - **Example**: Prove the Sherman-Morrison formula.

    Suppose $(A + uv^T)^{-1} \neq 0$.

$$\begin{aligned} &(A + uv^T)\left(A^{-1} - \frac{1}{1 + v^TA^{-1}u}A^{-1}uv^TA^{-1}\right) \\ &\quad = AA^{-1} + uv^TA^{-1} - \frac{1}{1 + v^TA^{-1}u}\left(AA^{-1}uv^TA^{-1} + uv^TA^{-1}uv^TA^{-1}\right) \\ &\quad = AA^{-1} + uv^TA^{-1} - \frac{1}{1 + v^TA^{-1}u}\left[u(1 + v^TA^{-1}u)v^TA^{-1}\right] \\ &\quad = AA^{-1} + uv^TA^{-1} - uv^TA^{-1} = AA^{-1} = I; \end{aligned}$$

    Therefore, $A + uv^T$ is invertible. Next, suppose $1 + v^TA^{-1}u = 0$;

$$\det(A + uv^T) = (1 + v^TA^{-1}u)\det(A) = 0 \implies A + uv^T \text{ is not invertible. } \blacksquare$$

- **Jacobi's Method**: Given a symmetric matrix A, rotate row k and column ℓ. WLOG, this will be row 1 and column 2, which forms $U = \begin{pmatrix} R & 0 \\ 0^T & I_{m-2} \end{pmatrix}$, where R_1 is a rotation matrix, and U is orthogonal.
  - $B = U^TAU = \begin{pmatrix} a_{11}\cos^2(\theta) - 2a_{12}\cos\theta\sin\theta + a_{22}\sin^2(\theta) & (a_{11} - a_{22})\cos\theta\sin\theta + a_{12}(\cos^2(\theta)\sin^2(\theta)) \\ (a_{11} - a_{22})\cos\theta\sin\theta + a_{12}(\cos^2(\theta)\sin^2(\theta)) & a_{11}\sin^2(\theta) + 2a_{12}\cos\theta\sin\theta + a_{22}\cos^2(\theta) \end{pmatrix}$.
    - $\sum_{i=1}^m a_{ii}^2 \leq \sum_{i=1}^m b_{ii}^2$.
    - Jacobi's method ensures that $b_{12} = 0$.
  - Iteratively obtain eigenvalues/vectors of a symmetric matrix A by transforming into a diagonal matrix.
  - **Rotation Matrix**: $R = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$.
    - R is an orthogonal transformation.

$$R \times R = \begin{bmatrix} \cos^2(\theta) + (-\sin\theta)^2 & \cos\theta\sin\theta - \sin\theta\cos\theta \\ \cos\theta\sin\theta - \sin\theta\cos\theta & (-\sin\theta)^2 + \cos^2(\theta) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

  - Two square matrices related by an orthogonal similarity transformation $B = U^TAU$ have equivalent traces and Frobenius norms.
  - **Similarity Transformation**: If λ and x are eigenvalue-vector pairs for matrix A, then λ and Bx are a pair for BAB^{-1}.

- A classical approach is to select the row i and column j such that we maximize the increase in the sum of squares in the diagonal.
  - However, this requires us to search all off-diagonal entries, which once the algorithm is complete, is $O(n^2 \log n)$.

- **Example**: Suppose λ is an eigenvalue of orthogonal matrix $\boldsymbol{O}$ with corresponding eigenvector v. Show that v has real entries only if $\lambda = \pm 1$.

  Recall that orthogonal transformations are norm-preserving. Therefore, using the definition of eigenvalues/vectors,

  $$|\lambda| \cdot \|v\| = \|\lambda v\| = \|Av\| = \|v\|;$$

  This means that $|\lambda| \cdot \|v\| \, \|v\| \implies |\lambda| = 1$. Since v is real, then $Av = \lambda v \implies \lambda v$ is real as well. Therefore, λ must also be a real number, which means that $\lambda = \pm 1$. ■

- **Rayleigh Quotient**: $R(x) = \frac{x^T A x}{x^T x}$, where $x \neq 0$.
  - If A has eigenvalues $\lambda_1 \leq \lambda_m$, then $R(x) \leq \lambda_m$, and $R(u_m) = \lambda_m$, where u_m is an orthonormal eigenvector.
    - **Example**: Prove the above statement.

      $$R(x) = \frac{x^T A x}{x^T x} = \frac{\left(\sum_{i=1}^m c_i u_i^T\right)\left(\sum_{i=1}^m \lambda_i u_i u_i^T\right)\left(\sum_{i=1}^m c_i u_i^T\right)}{\sum_{i=1}^m c_i^2} = \frac{\sum_{i=1}^m \lambda_i c_i^2}{\sum_{i=1}^m c_i^2} \leq \frac{\lambda_m \sum_{i=1}^m c_i^2}{\sum_{i=1}^m c_i^2} = \lambda_m;$$

      When $c_m = u_m$, $\frac{\sum_{i=1}^m \lambda_i c_i^2}{\sum_{i=1}^m c_i^2} = \frac{\lambda_m u_m^T u_m}{u_m^T u_m} = \lambda_m$. ■
  - **Courant-Fischer Theorem**: $\lambda_k = \min_{V_k} \max_{x \in V_k, x \neq 0} R(x) = \max_{V_{m-k+1}} \min_{x \in V_{m-k+1}, x \neq 0} R(x)$.
    - The Courant-Fischer theorem lets us bound how much the eigenvalues of a symmetric matrix change under a symmetric perturbation of the matrix. That is, $|\lambda_k - \mu_k| \leq \|\Delta A\|_2$.
    - **Example**: Prove the above claim.

      Let $\eta_1 \leq \cdots \leq \eta_m$ be the *magnitudes* of the eigenvalues of ΔA. In addition, let

      $$R_A(x) = \frac{x^T A x}{x^T x},\ R_B(x) = \frac{x^T (A + \Delta A) x}{x^T x} = R_A(x) + R_{\Delta A}(x).$$

      Let U_k be the span of the first kth eigenvectors of A.

      $$\begin{aligned}\mu_k &= \max_{V_k} \max_{x \in V_k, x \neq 0} R_B(x) \leq \max_{x \in U_k, x \neq 0} R_B(x) = \max_{x \in U_k, x \neq 0} (R_A(x) + R_{\Delta A}(x)) \\ &\leq \max_{x \in U_k, x \neq 0} R_A(x) + \max_{x \in U_k, x \neq 0} R_{\Delta A}(x) \leq \lambda_k + \max(\eta_i) = \lambda_k + \|\Delta A\|_2;\end{aligned}$$

      Reversing the roles of A and B, and combining with above, we can get that $\lambda_k - \mu_k \leq \eta_m$. ■
  - **Weyl's Inequality**: $\lambda_i + \rho_1 \leq \mu_i \leq \lambda_i + \rho_m$, where $\lambda_1, \ldots, \lambda_m$ are eigenvalues of A, $\rho_1 \leq \cdots \leq \rho_m$ are eigenvalues of ΔA, and $\mu_1, \ldots, \mu_m$ are eigenvalues of $B = A + \Delta A$.

- **Power Iteration Method**: Given diagonalizable and square matrix A with eigenvalues $|\lambda_1| \leq \cdots \leq |\lambda_m|$:
  1. Start with random $u^{(0)}$.
  2. Normalize the recurrence relation $u^{(n)} = \frac{Au^{(n-1)}}{\left\|Au^{(n-1)}\right\|_2}$. Repeat until convergence to $u^{(m)}$.
  3. Obtain $\lambda_m = R_A(u_m)$.
  - Lets us obtain specific eigenvalues of A.
  - The rate of convergence depends on $\frac{|\lambda_{m-1}|}{|\lambda_m|}$ and c_m (c_m defined in example below).
    - If $\lambda_{m-1} \approx \lambda_m$, then this algorithm converges very slowly, and if $\lambda_{m-1} = \lambda_m$, then the algorithm will not work at all.
    - If $c_m \approx 0$, convergence is slow, and if $c_m = 0$, then the algorithm will not converge.
  - If λ_m is negative, then $u^{(2n)}$ will still converge.
  - The default method obtains λ_m. To obtain λ_1, apply the algorithm to A^{-1}, and to obtain λ_k, apply the algorithm to $(A - \mu \cdot I)^{-1}$, where $\mu \approx \lambda_k$.

∗ **Example**: Obtain the eigenvalues of $(A-\mu I)^{-1}$.

Suppose $\lambda_1 \leq \cdots \leq \lambda_m$ are the eigenvalues of A, with corresponding eigenvectors $v_1, \ldots, v_m$.

$$\begin{aligned} Av_i = \lambda_i v_i \implies (A-\mu I)v_i &= \lambda_i v_i - \mu I v_i; \\ v_i &= (A-\mu I)^{-1}(\lambda_i - \mu)v_i; \\ (A-\mu I)^{-1} v_i &= \left(\frac{1}{\lambda_i - \mu}\right) v_i; \end{aligned}$$

Therefore, $\frac{1}{\lambda_i - \mu}$ are the eigenvalues of $(A-\mu I)^{-1}$. ■

– **Example**: Demonstrate that the power iteration method obtains an eigenvector.

Note that $\left(\frac{\lambda_i}{\lambda_m}\right)^m \to 0$ for $i < m$, and is 1 when $i = m$.

$$\begin{aligned} u^{(0)} &= \sum_{i=1}^m c_i v_i; \\ u^{(1)} &= \frac{Au^{(0)}}{\left\|Au^{(0)}\right\|_2} = \frac{A\sum_{i=1}^m c_i v_i}{\left\|A\sum_{i=1}^m c_i v_i\right\|_2} = \frac{\sum_{i=1}^m c_i \lambda_i v_i}{\left\|\sum_{i=1}^m c_i \lambda_i v_i\right\|_2} = \frac{\lambda_m \sum_{i=1}^m c_i \frac{\lambda_i}{\lambda_m} v_i}{|\lambda_m| \cdot \left\|\sum_{i=1}^m c_i \frac{\lambda_i}{\lambda_m} v_i\right\|_2}; \\ u^{(2)} &= \frac{\lambda_m^2}{|\lambda_m|^2} \cdot \frac{\sum_{i=1}^m c_i \left(\frac{\lambda_i}{\lambda_m}\right)^2 v_i}{\left\|\sum_{i=1}^m c_i \left(\frac{\lambda_i}{\lambda_m}\right)^2 v_i\right\|_2} = \frac{\lambda_m^2}{|\lambda_m|^2} \cdot \left(\frac{c_m v_m}{\|c_m v_m\|_2} + O(1)\right) \to v_m. \blacksquare \end{aligned}$$

• **Gerschgorin's Circle Theorem**: Every eigenvalue λ of $A_{m\times m}$ must satisfy $|\lambda - a_{ii}| \leq \sum_{j\neq i} |a_{ij}|$.

– **Example**: Prove Gerschgorin's Circle Theorem.

Suppose λ is an eigenvalue of A, and fix $i \in \{1, \ldots, m\}$.

$$\begin{aligned} Av = \lambda v \implies \lambda_v &= \sum_{i=1}^m \langle A_i, v\rangle = \sum_{i=1}^m \sum_{j=1}^m A_{ij} v_j; \\ (\lambda - A_{ii}) v_i &= \sum_{j\neq i} A_{ij} v_j; \\ &\text{Now, let } i = \arg\max_{j=1,\ldots,m} |v_j|. \\ |(\lambda - A_{ii})v_i| &= \left|\sum_{j\neq i} A_{ij} v_j\right| \leq \sum_{j\neq i} |A_{ij}| \cdot |v_j| \\ \implies |\lambda - A_{ii}| &\leq \sum_{j\neq i} |A_{ij}| \cdot \frac{|v_j|}{|v_i|} \leq \sum_{j\neq i} |A_{ij}|. \blacksquare \end{aligned}$$

– If A is symmetric and we want λ_m, then a plausible guess is $\mu = \max_i a_{ii}$.

∗ A more conservative guess would be $\mu = \max_{1\leq i\leq m} \left\{a_{ii} + \sum_{j\neq i} |a_{ij}|\right\}$.

– **Example**:

1. Suppose that the rows of matrix $\boldsymbol{A}$ satisfy $\sum_j |\boldsymbol{A}_{ij}| < 1$. Show that all eigenvalues of $\boldsymbol{A}$ satisfy $|\lambda| < 1$.
2. Now, suppose that the rows of matrix $\boldsymbol{A}$ satisfy $\sum_j |\boldsymbol{A}_{ij}| < |\boldsymbol{A}_{ii}|$. Show that $\boldsymbol{A}$ is non-singular.

1. Let λ be some arbitrary eigenvalue of A.

$$\sum_j |A_{ij}| < 1 \implies 1 - |A_{ii}| > \sum_{j\neq i} |A_{ij}| \geq |\lambda - A_{ii}|;$$

$|\lambda - A_{ii}| < 1 - |A_{ii}| \implies |\lambda| < 1$. Since λ is an arbitrary eigenvalue, this must hold for all eigenvalues of A.

2. We need to show that $\lambda \neq 0$ for all λ. Let λ be some eigenvalue of A. Proceed with contradiction. Suppose $\lambda = 0$.

$$|\lambda - A_{ii}| = |A_{ii}| \leq \sum_{j\neq i} |A_{ij}| < |A_{ii}|;$$

$|A_{ii}| < |A_{ii}|$ is a contradiction, so $\lambda \neq 0$ by contradiction. Therefore, all of the eigenvalues of A are non-zero, so A is non-singular by definition. ■

- **SVD**: Decompose rank-k matrix A into $A = U\Sigma V^T$, where $u_1, \ldots, u_k$ and $v_1, \ldots, v_k$ are orthonormal.
  - **Singular Values**, or σ_j: The eigenvalues of $A^T A$.
  - SVD is a one-sided Jacobi transformation, where in SVD we right-multiply the matrix by a transformation matrix, and continue operating on AV. Once AV stabilizes, we normalize nontrivial columns that results in $U\Sigma$, where columns of U are either orthogonal or zero. Discarding the zero columns of U results in $AV = U\Sigma \equiv A = U\Sigma V^T$, since V is orthogonal.
  - $U\Sigma V^T = \sum_{i=1}^{k} \sigma_i u_j v_j^T$, where $\sigma_i > 0$.
  - $\|A\|_2 = \|\Sigma\|_2 = \max_i \sigma_i$.
  - $\|A\|_F = \|\Sigma\|_F = \sqrt{\sum_i \sigma_i^2}$.
  - **Example**: Suppose $\boldsymbol{A}$ has SVD $\boldsymbol{U\Sigma V^T}$, with σ_{ii} appearing in decreasing order. Prove that the best rank-k approximation of $\boldsymbol{A}$ for $\|\boldsymbol{A}\|_F$ is $\boldsymbol{B} = \sum_{j=1}^{k} \sigma_j u_j v_j^T$.

    $$\|\boldsymbol{A} - \boldsymbol{B}\|_F^2 = \left\|\boldsymbol{U\Sigma V}^T - \boldsymbol{B}\right\|_F^2 = \left\|\boldsymbol{U}^T(\boldsymbol{U\Sigma V}^T - \boldsymbol{B})\boldsymbol{V}\right\|_F^2 = \left\|\boldsymbol{\Sigma} - \boldsymbol{U}^T\boldsymbol{BV}\right\|_F^2;$$

    Since $\boldsymbol{\Sigma}$ is a diagonal matrix with σ_{ii} appearing in decreasing order, a good rank-k approximation of $\boldsymbol{A}$ would use a modified Σ that sets $\sigma_{(k+1)(k+1)}, \ldots, \sigma_{mm} = 0$. Under this approximation, $\boldsymbol{B} = \sum_{j=1}^{k} \sigma_{jj} u_j v_J^T$. ■
    - Under this approximation, $\|A - B\|_F = \sqrt{\sum_{i \geq k} \sigma_i}$, and $\|A - B\|_2 = \sigma_{k+1}^2$.
  - **Ridge Regression**: Introduce a $\lambda > 0$ term such that $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$.
    - Using SVD on X, we get that $X^T y = \sum_j \sigma_j v^T y v_j$, and $X^T X + \lambda I = \sum_j (\sigma_j^2 + \lambda) v_j v_j^T$.
    - Using SVD on X, we obtain

      $$\hat{\beta} = \sum_j \frac{\sigma_j}{\sigma_j^2 + \lambda} u_j^T j v_j, \text{ and } \hat{y} = \sum_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} u_j^T y v_j.$$

      This means that increasing λ shrinks $\hat{\beta}$ and $\hat{y}$ towards zero, and shrinkage is most pronounced in directions v_j with smaller σ_j.
  - **Example**: Show that the singular values of $\boldsymbol{A}$ are equal to the singular values of $\boldsymbol{WA}$, where $\boldsymbol{W}$ is orthogonal.

    Using SVD, let $\boldsymbol{A} = \boldsymbol{U\Sigma V}^T$. $\boldsymbol{WA} = \boldsymbol{WU\Sigma V}^T = (\boldsymbol{WU})\boldsymbol{\Sigma V}^T$. Since $\boldsymbol{\Sigma}$ is unchanged, and we have found the SVD of $\boldsymbol{WA}$, $\boldsymbol{WA}$ must have the same singular values as $\boldsymbol{A}$. ■
  - **Example**: Show that $\|\boldsymbol{A}\|_F \leq \text{rank}(\boldsymbol{A}) \|\boldsymbol{A}\|_2$.

    Let $r = \text{rank}(A) \leq m$, and σ_i be the ith smallest singular value of A. If $r \leq m$, then there are $(m - r)$ singular values equal to zero.

    $$\|\boldsymbol{A}\|_F^2 = \text{tr}\boldsymbol{A}^T\boldsymbol{A} = \text{tr}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} = \sum_{i=1}^{m} \sigma_i^2 \leq \sum_{i=1}^{m} \sigma_m^2 = \sum_{i=(m-r)+1}^{m} \sigma_i^2 = r\sigma_m^2 = r\|\boldsymbol{A}\|_2^2. \blacksquare$$

8.3 Optimization

Return to Table of Contents

- **Unconstrained Optimization**: Solving $\min_x f(x) = f(x^*) = p^*$ for convex and twice-differentiable $f : \mathbb{R}^n \to \mathbb{R}$.
  - **Convex Function**: A function $f : \mathbb{R}^n \to \mathbb{R}$ iff for all x, y and any $\theta \in (0, 1)$, $f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta) f(y)$.
    - For this class, we assume the domain is $\mathbb{R}^n$.
    - x, y are vectors.
  - **Strictly Convex**; A function $f : \mathbb{R}^n \to \mathbb{R}$ iff for all x, y and any $\theta \in (0, 1)$, $f(\theta x + (1 - \theta)y) < \theta f(x) + (1 - \theta) f(y)$.

* Algorithms that converge to the global minimum on a strictly convex function will converge to a global minimum on a convex function.
* We extend this via Jensen's inequality to get

$$f\left(\sum_{i=1}^{n}\theta_i x_i\right) \leq \sum_{i=1}^{n}\theta_i f(x_i) \text{ for } \sum_{i=1}^{n}\theta_i = 1.$$

– **Example**: Show that the vector norm function f is convex.

We need to show that $\|\theta x + (1-\theta)y\| \leq \theta\|x\| + (1-\theta)\|y\|$ for vectors x and y. Applying the sub-additivity and scalar multiplicity properties of vector norms,

$$\|\theta x + (1-\theta)y\| \leq \|\theta x\| + \|(1-\theta)y\| = |\theta| \cdot \|x\| + |1-\theta| \cdot \|y\| = \theta\|x\| + (1-\theta)\|y\|. \blacksquare$$

– **Example**: Show that $f(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$ is convex.

We need to show that $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$ for vectors x and y.

$$\begin{aligned} f(\theta x + (1-\theta)y) =& \max\{\theta x_1 + (1-\theta)y_1, \ldots, \theta x_n + (1-\theta)y_n\} \\ \leq& \max\{\theta x_1, \ldots, \theta x_n\} + \max\{(1-\theta)y_1, \ldots, (1-\theta)y_n\} \\ =& \theta \cdot \max\{x_1, \ldots, x_n\} + (1-\theta) \cdot \max\{y_1, \ldots, y_n\} = \theta f(x) + (1-\theta)f(y). \blacksquare \end{aligned}$$

– **Example**: Show that the K-L Divergence function $f(x_1, x_2) = x_1 \log\left(\frac{x_1}{x_2}\right) + x_2 - x_1$ is convex.

Suppose $x = (x_1, x_2)$ and $y = (y_1, y_2)$.

$$\begin{aligned} f(\theta x + (1-\theta)y) =& (\theta x_1 + (1-\theta)y_1)\log\left(\frac{\theta x_1 + (1-\theta)y_1}{\theta x_2 + (1-\theta)y_2}\right) + [\theta x_2 + (1-\theta)y_2] - [\theta x_1 + (1-\theta)y_1] \\ =& \theta[x_2 - x_1] + (1-\theta)[y_2 - y_1] + (\theta x_1 + (1-\theta)y_1)\log\left(\frac{\theta x_1 + (1-\theta)y_1}{\theta x_2 + (1-\theta)y_2}\right); \end{aligned}$$

Applying the log-sum inequality with $n = 2$, $\log\left(\frac{\theta x_1 + (1-\theta)y_1}{\theta x_2 + (1-\theta)y_2}\right) \leq \theta x_1 \log\left(\frac{x_1}{x_2}\right) + (1-\theta)y_1 \log\left(\frac{y_1}{y_2}\right)$. Therefore,

$$\begin{aligned} f(\theta x + (1-\theta)y) =& \theta[x_2 - x_1] + (1-\theta)[y_2 - y_1] + (\theta x_1 + (1-\theta)y_1)\log\left(\frac{\theta x_1 + (1-\theta)y_1}{\theta x_2 + (1-\theta)y_2}\right) \\ \leq& \theta[x_2 - x_1] + (1-\theta)[y_2 - y_1] + \theta x_1 \log\left(\frac{x_1}{x_2}\right) + (1-\theta)y_1 \log\left(\frac{y_1}{y_2}\right) \\ =& \theta\left[x_2 - x_1 + x_1 \log\left(\frac{x_1}{x_2}\right)\right] + (1-\theta)\left[y_2 - y_1 + y_1 \log\left(\frac{y_1}{y_2}\right)\right] = \theta f(x) + (1-\theta)f(y). \blacksquare \end{aligned}$$

– **Example**: Suppose f is convex.

1. Show that $f(x) \leq \frac{b-x}{b-a}f(a) + \frac{x-a}{b-a}f(b)$ for $x \in [a, b]$.
2. Show that $\frac{f(x)-f(a)}{x-a} \leq \frac{f(b)-f(a)}{b-a} \leq \frac{f(b)-f(x)}{b-x}$ for $x \in (a, b)$.
3. Suppose f is differentiable. Show that $f'(a) = \frac{f(b)-f(a)}{b-a} \leq f'(b)$.
4. Suppose f is twice-differentiable. Show that $f''(a)$ and $f''(b) \geq 0$.

1. Since f is convex iff for any $\theta \in (0, 1)$, $f(\theta a + (1-\theta)b) \leq \theta f(a) + (1-\theta)f(b)$. Let $x = \theta a + (1-\theta)b$, which means that $x \in [a, b]$, and let $\theta = \frac{b-x}{b-a} \in [0, 1]$ for $x \in [a, b]$.

$$\begin{aligned} f(x) \leq& \theta f(a) + (1-\theta)f(b) = \frac{b-x}{b-a}f(a) + \left(1 - \frac{b-x}{b-a}\right)f(b) \\ =& \frac{b-x}{b-a}f(a) + \left(\frac{b-a-(b-x)}{b-a}\right)f(b) = \frac{b-x}{b-a}f(a) + \frac{x-a}{b-a}f(b). \end{aligned}$$

2. Using the previous part,

$$\begin{aligned}
f(x) \leq& \frac{b-x}{b-a}f(a) + \frac{x-a}{b-a}f(b);\\
(b-a)f(x) \leq& (b-x)f(a) + (x-a)f(b)\\
(b-a)f(x) - (b-x)f(a) \leq& (x-a)f(b)\\
(x-a)f(x) + (b-x)(f(x)-f(a)) \leq& (x-a)f(b)\\
(b-x)(f(x)-f(a)) \leq& (x-a)f(b) - (x-a)f(x) = (x-a)(f(b)-f(x));\\
\frac{f(x)-f(a)}{x-a} \leq& \frac{f(b)-f(x)}{b-x}.
\end{aligned}$$

Next,

$$\begin{aligned}
f(b)-f(a) =& f(b)-f(x)-(f(a)-f(x));\\
\frac{f(b)-f(a)}{b-a} =& \frac{b-x}{b-x}\cdot\frac{f(b)-f(x)}{b-a} - \frac{x-a}{x-a}\cdot\frac{f(x)-f(a)}{b-a}\\
=& \frac{b-x}{b-a}\cdot\frac{f(b)-f(x)}{b-x} - \frac{x-a}{b-a}\cdot\frac{f(x)-f(a)}{x-a};
\end{aligned}$$

Combining these two results,

$$\begin{aligned}
\frac{f(b)-f(a)}{b-a} \leq& \frac{b-x}{b-a}\cdot\frac{f(b)-f(x)}{b-x} - \frac{x-a}{b-a}\frac{f(b)-f(x)}{b-x}\\
=& \frac{f(b)-f(x)}{b-x}\left(\frac{b-x}{b-a} - \frac{x-a}{b-a}\right) = \frac{f(b)-f(x)}{b-x}\left(\frac{b+a-2x}{b-a}\right)\\
\leq& \frac{f(b)-f(x)}{b-x}\left(\frac{b-a}{b-a}\right) = \frac{f(b)-f(x)}{b-x}
\end{aligned}$$

Next, note that we can express $\frac{f(b)-f(a)}{b-a}$ as a weighted average of $\frac{f(b)-f(x)}{b-x}$ and $\frac{f(x)-f(a)}{x-a}$. This means that $\frac{f(b)-f(a)}{b-a} \geq \min\left\{\frac{f(b)-f(x)}{b-x}, \frac{f(x)-f(a)}{x-a}\right\} = \frac{f(x)-f(a)}{x-a}$. Therefore,

$$\frac{f(x)-f(a)}{x-a} \leq \frac{f(b)-f(a)}{b-a} \leq \frac{f(b)-f(x)}{b-x}, \text{ as desired.}$$

3. Note that, by the definition of a derivative, $f'(b) = \lim_{x\to b^-}\frac{f(b)-f(x)}{b-x}$ and $f'(a) = \lim_{x\to a^+}\frac{f(x)-f(a)}{x-a}$. So,

$$\begin{aligned}
\frac{f(x)-f(a)}{x-a} \leq& \frac{f(b)-f(a)}{b-a} \leq \frac{f(b)-f(x)}{b-x};\\
f'(a) = \lim_{x\to a^+}\frac{f(x)-f(a)}{x-a} \leq& \frac{f(x)-f(a)}{x-a} \leq \frac{f(b)-f(a)}{b-a} \leq \frac{f(b)-f(x)}{b-x} \leq \lim_{x\to b^-}\frac{f(b)-f(x)}{b-x} = f'(b).
\end{aligned}$$

4. Let $\epsilon > 0$ be small. Then,

$$f'(a) = \lim_{\epsilon\to 0^+}\frac{f(x+\epsilon)-f(a)}{\epsilon} \leq \frac{f(x+\epsilon)-f(a)}{\epsilon} \leq f'(a+\epsilon).$$

This means that

$$0 \leq \frac{f(x+\epsilon)-f(a)}{\epsilon} - f'(a) \leq f'(a+\epsilon) - f'(a).$$

Dividing by ϵ and taking the limit, we get that

$$0 \leq \lim_{\epsilon\to 0^+}\frac{\frac{f(x+\epsilon)-f(a)}{\epsilon} - f'(a)}{\epsilon} \leq \lim_{\epsilon\to 0^+}\frac{f'(a+\epsilon)-f'(a)}{\epsilon} = f''(a).$$

Now, we apply similar logic to get that $f''(b) \geq 0$.

$$f'(b) = \lim_{\epsilon\to 0^-}\frac{f(b)-f(b-\epsilon)}{\epsilon} \geq \frac{f(b)-f(b-\epsilon)}{\epsilon} \geq f'(b-\epsilon).$$

This means that

$$0 \leq f'(b) - \frac{f(b)-f(b-\epsilon)}{\epsilon} \leq f'(b) - f'(b-\epsilon).$$

Dividing by ϵ and taking the limit, we get that

$$0 \leq \lim_{\epsilon\to 0^-}\frac{f'(b) - \frac{f(b)-f(b-\epsilon)}{\epsilon}}{\epsilon} \leq \lim_{\epsilon\to 0^-}\frac{f'(b)-f'(b\epsilon)}{\epsilon} = f''(b). \blacksquare$$

- Properties of convex functions:
  1. Every local minimum of a convex function is a global minimum. If f is strictly convex, then it has a unique global minimum.
     - *Proof*: Suppose x^* is a local minimum. This means that $f(x^*) \leq f(x)$ for all $x \ni |x - x^*| < \epsilon$. Consider any $x^{**} \in \mathbb{R}$, and find $\theta \ni x = x^* + \theta(x^{**} - x^*)$ and $|x - x^*| < \epsilon$. Then,

$$f(x^*) \leq f(x) = f\left((1-\theta)x^* + \theta x^{**}\right) \leq (1-\theta)f(x^*) + \theta f(x^{**}) \implies f(x^*) \leq f(x^{**}). \blacksquare$$

  2. If f is convex and differentiable, then $f(y) \geq f(x) + \nabla f(x)^T(y - x)$.
     - This inequality is strict when f is strictly convex.
     - *Proof*: Start with $n = 1$ (univariate function).

$$\begin{aligned}\forall t \in (0,1],\, f\left(x + t(y-x)\right) \leq & (1-t)f(x) + tf(y) \implies f(y) \geq f(x) + \frac{f\left(x + t(y-x) - f(x)\right)}{t} \\ = & f(x) + \frac{f\left(x + t(y-x) - f(x)\right)}{t(y-x)}(y-x) = f(x) + f'(x)(y-x).\end{aligned}$$

       Now, suppose $n \geq 2$. Define $g : [0,1] \to \mathbb{R}$ as $g(t) = f\left(ty + (1-t)x\right)$. First, show g is convex. Fix $t_1, t_2 \in [0,1]$ and $\theta \in [0,1]$. Let $z_1 = t_1 y + (1-t_1)x$ and $z_2 = t_2 y + (1-t_2)x$. Then, $g(t_1) = f(z_1)$ and $g(t_2) = f(z_2)$.

$$\begin{aligned}g\left(\theta t_1 + (1-\theta)t_2\right) = & f\left(\left[\theta t_1 + (1-\theta)t_2\right]y + \left[1 - \left(\theta t_1 + (1-\theta)t_2\right)\right]x\right) \\ = & f\left(\theta z_1 + (1-\theta)z_2\right) \leq \theta f(z_1) + (1-\theta)f(z_2) \\ = & \theta g(t_1) + (1-\theta)g(t_2).\end{aligned}$$

       Since g is convex and univariate, $g(y) \geq g(x) + g'(x)(y-x)$ for any x, y in the domain of g. So, $g(t_2) \leq g(t_1) + g'(t_1)(t_2 - t_1)$. Now set $t_1 = 0$, $t_2 = 1$.

$$f(y) = g(1) \geq g(0) + g'(0) = f(x) + g'(0) = f(x) + \nabla f(x)^T(y-x). \blacksquare$$

     - This implies that if $\nabla(x^*) = 0$, then x^* is a global minimum of f.
     - $f(x) + \nabla f(x)^T(y - x)$ is the first-order Taylor approximation of f near x.
       - For a convex function, this approximation is a global under-estimator of f.
- A necessary and sufficient condition for x^* to be optimal is $\nabla f(x^*) = 0$.
  - In a few special cases, we can find a solution analytically. However, in most cases, we must iteratively solve for x^* within some tolerance $\epsilon > 0$.
    - We terminate the algorithm when $f(x^{(k)}) - p^* \leq \epsilon$.
- **Section Search**: Suppose we have $f(x) = \sum_{i=1}^m |x_i - x|$. f does not have a derivative, so it does not have a closed-form solution. However, we do know that f is convex, and $x^* \in (x_L, x_U)$, where $x_L = \min\{x_i\}$, and $x_U = \max\{x_i\}$.

Algorithm 1 Univariate Section Search

Require: Initial bounds (l, u), $x^* = \frac{l+u}{2}$ and some $\epsilon > 0$.
while $ub - lb > \epsilon$ **do**
  Generate candidate $x_c \sim Unif(lb, ub)$, and calculate $f(x_c)$.
  if $f(x_c) < f(x^*)$ **then**
    Set $x^* = x_c$ and $f(x^*) = f(x_c)$.
  end if
  if $x_c < x^*$ **then**
    Set $l = x_c$
  else
    Set $u = x_c$
  end if
end while

  - This is a greedy algorithm.

- Uniform sampling might not be efficient (especially for asymmetric convex functions). Rather, we should shrink the larger of the intervals (x_L, x^*) and (x^*, x_U). In other words, we should pick points such that the new search interval is self-similar to the previous interval.
    - Symmetric functions are naturally self-similar.
- **Golden Section Search**:

Algorithm 2 Golden Section Search

Require: Initial bounds (l, u), $x^* = \frac{l+u}{2}$ and some $\epsilon > 0$, $\phi = \frac{\sqrt{5}-1}{2}$, and $\Delta x = ub - lb$.
while $ub - lb > \epsilon$ **do**
  Generate candidate $x_{cl} = u - \phi \cdot \Delta x$ and $x_{cu} = l + \phi \cdot \Delta x$, and calculate $f(x_{cl})$ and $f(x_{cu})$.
  if $f(x_{cl}) < f(x_{cu})$ **then**
    Set $u = x_{cu}$.
    Set $x_{cu} = x_{cl}$ and $f(x_{cu}) = f(x_{cl})$.
    Set $x_{cl} = u - \phi \cdot \Delta x$, and obtain $f(x_{cl})$.
  else
    Set $l = x_{cl}$.
    Set $x_{cl} = x_{cu}$ and $f(x_{cl}) = f(x_{cu})$.
    Set $x_{cu} = l + \phi \cdot \Delta x$, and obtain $f(x_{cu})$.
  end if
end while
return $\frac{l+u}{2}$.

- (x_L, x_U) converges to zero, but how fast?
    - Assuming f is locally smooth, we can fit a locally smooth curve to f, and used the closed-form solution for that curve.

- **Parabolic Interpolation Method**: Suppose we have $a = x_L$, $c = x_U$, and $x_{new} = b \in (a, c)$ such that $f(c) < f(b) < f(a)$. Then, we fit a parabola to the points $\{(a, f(a)), (b, f(b)), (c, f(c))\}$. This returns a closed form expression for the minimizer x_{min}, which ensures $f(x_{min}) < f(b)$. Then, repeat the process with a, x_{min}, b, or b, x_{min}, c.

Algorithm 3 Parabolic Interpolation

Require: Initial bounds (l, u), $x^* = m = \frac{l+u}{2}$, $f(x^*)$, and some $\epsilon > 0$.
while $ub - lb > \epsilon$ **do**
  Generate $f(l)$, $f(m)$, and $f(u)$.
  Calculate $x_c = m + \frac{1}{2} \cdot \frac{[f(l)-f(m)](u-m)^2 - [f(u)-f(m)](m-l)^2}{[f(l)-f(m)](u-m) + [f(u)-f(m)](m-l)}$, and $f(x_c)$.
  if $f(x_c) < f(x^*)$ **then**
    Set $x^* = x_c$ and $f(x^*) = f(x_c)$.
  else if $x^* < x_c$ **then**
    Set $u = x_c$
  else
    Set $l = x_c$.
  end if
end while

- The minimum of the quadratic function is on the slides, as is a more formal algorithm. Once again, there is a typo.
- Is not very robust, and may get stuck in certain forms.
    - **Brent's Method**: If this happens, switch to Golden section search.
        - Brent's method is the fastest when we don't have access to readily computable first and second derivatives.

- **Newton's Method**: If f is twice-differentiable, and we know how to compute the derivatives, then we can make the optimizer much more efficient.

Algorithm 4 Univariate Newton's Method

Require: Initial $x^{(0)}$ and some $\epsilon > 0$.
1: **while** $|x^{(k+1)} - x^{(k)}| > \epsilon$ **do**
2: Generate $x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})}$.
3: **end while**
4: **return** $x^{(k+1)}$.

- Is based on the assumption that f can be approximated in the neighborhood of its minimum by the quadratic Taylor series approximation, $q(x)$. For x close to x_0,

$$f(x) \approx f(x_0) + f(x_0)'(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 = q(x).$$

- Need to compute derivatives at each step, which may be expensive.
  - However, since we may need fewer iterations, this additional cost may not impact overall runtime.
- We construct a convergent sequence of solutions $x^{(k)}$ as opposed to shrinking intervals.
- We are solving for $f'(x^*) = 0$ to solve the optimization problem.
- **Newton's Root-Finding Method**:

Algorithm 5 Newton's Root-Finding Method

Require: Initial $x^{(0)}$ and some $\epsilon > 0$.
1: **while** $|x^{(k+1)} - x^{(k)}| > \epsilon$ **do**
2: Generate $x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}$.
3: **end while**
4: **return** $x^{(k+1)}$.

- Is the fastest (in the number of iterations needed) if we have access to the first and second derivatives.

- **Strong Convexity**: A differentiable function f with parameter $m > 0$ such that for all x, y, $(\nabla f(x) - \nabla f(y))^T (x - y) \geq m \|x - y\|_2^2$, or more generally, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq m \|x - y\|^2$.
  - We will default to ℓ_2 norm, although there is no real difference than using some other norm.
  - If $m = 0$, then all convex functions are strongly convex.
  - An equivalent condition is that $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|y - x\|_2^2$.
  - It is not necessary for f to be differentiable to be strongly convex. We can also define strongly convex as, for $t \in [0, 1]$, $f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta) f(y) - \frac{1}{2} m \theta (1 - \theta) \|x - y\|_2^2$.
- Strong convexity extends and parameterizes strict convexity. Strongly convex $\implies$ strictly convex.
  - For twice-continuously differentiable $f : \mathbb{R} \to \mathbb{R}$:
    - f is convex iff $f''(x) \geq 0$ for all x.
    - **Strictly Convex**: f is strictly convex if $f''(x) > 0$ for all x.
      - This is sufficient, but not necessary.
    - f is strongly convex iff $f''(x) \geq m > 0$ for all x.
- What if we are in a multivariate setting?
  - If f is twice continuously-differentiable, then it is strongly convex with parameter m iff $\nabla^2 f(x) \succeq mI$ for all $x \in \mathbb{R}^m$, where I is the identity matrix and $\nabla^2 f$ is the Hessian matrix, and $\succeq$ means that $\nabla^2 f(x) - mI$ is non-negative definite.
    - When $n = 1$, this reduces to the univariate definition given earlier, where $f''(x) \geq m > 0$.
  - We create a global under-estimator of f as

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|y - x\|_2^2 .$$

- **Descent Methods**: The algorithms $x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)}$, where $\Delta x^{(k)}$ is a descent direction, iteratively updating Δx and $t > 0$, and $t^{(k)} > 0$ is a step size.

- These are descent methods, since $f(x^{(k+1)}) < f(x^{(k)})$.
- These algorithms involve choosing initial $x^{(0)} \in \mathbb{R}^n$, and updating $x \leftarrow x + t\Delta x$, stopping when some criterion is satisfied.

- **Line Search**: Choosing $t > 0$.
  - t is chosen to minimize f along the ray $\{x + t\Delta x | t \geq 0\}$, or $t = \arg\min_{s \geq 0} f(x + s\Delta x)$.
  - Line search is a univariate convex optimization problem, represented as $\phi(s) = f(x + s\Delta x)$ which is convex in s.
    * $\phi(s)$ is convex since it is a composition of a convex function f and $x + t\Delta x$.
  - In practice, it might be inefficient to optimize t at each iteration.
    * In practice, line searches typically *sufficiently* reduce f along the aforementioned ray.
  - **Backtracking Line Search**: Given descent direction Δx, $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$, first set $t := 1$. While $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$, $t := \beta t$.
    * This is backtracking because it starts with a unit step size and reduces by β until the stopping condition is satisfied.
    * Since Δx is a descent direction, $\nabla f(x)^T x < 0$, so for small t,
    $$f(x + t\Delta x) \approx f(x) + f\triangle f(x)^T \Delta x < f(x) + \alpha t \nabla f(x)^T \Delta x.$$
    * α can be interpreted as the fraction of the decrease in f predicted by linear extrapolation that we will accept.

- **Gradient Descent**: $\Delta x = -\nabla f(x)$.

Algorithm 6 Gradient Descent

Require: Initial $x^{(0)}$ and some $\epsilon > 0$.
1: **while** $|x^{(k+1)} - x^{(k)}| > \epsilon$ **do**
2: Calculate $\Delta x = -\nabla f(x)$.
3: Choose step size t via line search or backtracking line search.
4: Update $x^{(k+1)} = x^{(k)} + t\Delta x$.
5: **end while**
6: **return** $x^{(k+1)}$.

  - From convexity, $\nabla f(x^{(k)})^T (y - x^{(k)}) \geq 0 \implies f(y) \geq f(x^{(k)})$, so $\nabla f(x^{(k)})^T \Delta x^{(k)} < 0$.
  - $x^+ = x - t\nabla f(x)$. Define $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$ as $f(t) = f(x - t\nabla f(x))$. Since f is strongly convex, $\exists M \ni x, y$ for all x and y such that $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|_2^2$.
    Then, $\tilde{f}(t) \leq f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2} \cdot \|\nabla f(x)\|_2^2$. For exact line search, $\tilde{f}(t) \leq f(x) - t\|\nabla f(x)\|^2 + \frac{Mt^2}{2} \|\nabla f(x)\|_2^2 =: g(t)$, so $g'(t) = -\|\nabla f(x)\|_2^2 + Mt\|\nabla f(x)\|_2^2$. Setting $g'(t) = 0 \implies t = \frac{1}{M}$. Plugging this into the previous formula, $\tilde{f}(t) \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$.
    So, $f(t_{exact}) \leq f(\frac{1}{M}) \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$. So,
    $$f(x^+) \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2 \implies f(x^+) - p^* \leq (f(x) - p^*) - \frac{1}{2M}\|\nabla f(x)\|_2^2,$$
    where p^* is the minimum of f.
    We know that $\|\nabla f(x)\|_2^2 \geq 2m(f(x) - p^*)$ (shown in next class), where m is the parameter in the strong convexity condition.
    Therefore, $f(x^+) - p^* \leq \left(1 - \frac{m}{M}\right)(f(x) - p^*)$. Thus, we are given the rate of convergence and strength. That is, if $\frac{m}{M} \approx 1$, then the rate of convergence is very fast. So,
    $$f(x^{(k+1)}) - p^* \leq \left(1 - \frac{m}{M}\right)^{k+1} \left(f(x^{(0)}) - p^*\right).$$
    So, to achieve a ε-approximation $f(x^{(k)}) - p^* \leq \varepsilon$, we need $k = \frac{-\log\left\{\frac{f(x^{(0)}) - p^*}{\varepsilon}\right\}}{\log\left(1 - \frac{m}{M}\right)}$.

- **Strongly Convex**: f such that there exists positive constants m and M such that $m\boldsymbol{I} \preceq \nabla^2 f(x) \preceq M\boldsymbol{I}$.

– This means that for any x, the Hessian matrix is positive definite, the smallest eigenvalue is no smaller than m, and the largest eigenvalue is no larger than M.

- Convergence analysis of gradient descent: Start with the definition of strongly convex. $\exists m > 0$ such that $\nabla^2 f(x) \succeq m\boldsymbol{I}$ for all $x \in \mathbb{R}^n$.
  Fix any $x, y \in \mathbb{R}^n$. We assume f is twice-differentiable to obtain the following equation. Then, $\exists z = x + \theta(y - x)$ where $\theta \in (0, 1)$ such that

$$\begin{aligned} f(y) =& f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x) \\ \geq& f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|y - x\|_2^2 . \end{aligned}$$

  Consider $g(y) = f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}(y - x)^T (y - x)$, which is a function of y with x fixed. Then,

$$\begin{aligned} \nabla g(y) =& \nabla f(x) + m(y - x); \\ \nabla g(y) = 0 \implies& \tilde{y} = x - \frac{1}{m}\nabla f(x); \\ f(y) \geq& f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|y - x\|_2^2 \\ \geq& f(x) - \frac{1}{m}\nabla f(x)^T \nabla f(x) + \frac{1}{2m}\nabla f(x)^T \nabla f(x) \\ =& f(x) - \frac{1}{2m}\nabla f(x)^T f(x) \implies p^* \geq f(x) - \frac{1}{2m} \|\nabla f(x)\|_2^2 ; \end{aligned}$$

  If $\nabla f(x)$ has small entries, then we bound p^* below by $f(x)$.
  Now, setting $y = x^*$ yields

$$\begin{aligned} p^* =& f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{m}{2} \|x^* - x\|_2^2 \\ \overset{\text{CS}}{\geq}& f(x) - \|\nabla f(x)\|_2 \|x^* - x\|_2 + \frac{m}{2} \|x^* - x\|_2^2 \\ \implies& 0 \geq - \|\nabla f(x)\|_2 \|x^* - x\|_2 + \frac{m}{2} \|x^* - x\|_2^2 \\ \implies& \frac{m}{2} \|x^* - x\|_2^2 \leq \|\nabla f(x)\|_2 \|x^* - x\|_2 \\ \implies& \|x^* - x\|_2 \leq \frac{2}{m} \|\nabla f(x)\|_2 . \end{aligned}$$

  In addition, strongly convex functions satisfy $\nabla^2 f(x) \preceq M\boldsymbol{I}$ for all $x \in \mathbb{R}^n$ and some $M > m$. This means that

$$\begin{aligned} f(y) =& f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x) \\ \leq& f(x)\nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|_2^2 . \end{aligned}$$

  Similar math yields $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|_2^2$.

  – Recall gradient descent uses $x^+ = x - t\nabla f(x)$. Define $\tilde{f} : \mathbb{R} \to \mathbb{R}$ as $\tilde{f}(t) = f\left(x - t\nabla f(x)\right)$.

$$\begin{aligned} f(y) \leq& f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} \|y - x\|_2^2 ; \\ & \text{Set } y = x - t\nabla f(x); \\ \implies \tilde{f}(t) \leq& f(x) - t \|\nabla f(x)\|_2^2 + \frac{Mt^2}{2} \|\nabla f(x)\|_2^2 . \end{aligned}$$

    Suppose we use exact line search. Let $f(x^+) = \tilde{f}(t_{exact})$. $f(x) - t \|\nabla f(x)\|_2^2 + \frac{Mt^2}{2} \|\nabla f(x)\|_2^2$ is minimized when $t = \frac{1}{M}$. So,

$$\begin{aligned} f(x^+) \leq& f(x) - \frac{1}{2M} \|\nabla f(x)\|_2^2 \\ \implies \left(f(x^+) - p^*\right) =& \left(f(x) - p^*\right) - \frac{1}{2M} \|\nabla f(x)\|_2^2 \\ \implies \|\nabla f(x)\|_2^2 \geq& 2m \left(f(x) - p^*\right) \\ \implies f(x^+) - p^* \leq& \left(1 - \frac{m}{M}\right) \left(f(x) - p^*\right) \\ \implies f(x^{(k)}) - p^* \leq& \left(1 - \frac{m}{M}\right)^k \left(f(x^{(0)}) - p^*\right) . \end{aligned}$$

This means we have geometric convergence.
Now, suppose we use backtracking line search (BLS). $\tilde{f}(t) \leq f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2$. The general BLS exit condition is when $f(x+t\Delta x) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$. In a gradient descent framework, it is when $\tilde{f}(t) \leq f(x) - \alpha t \|\nabla f(x)\|_2^2$.

$$\begin{aligned} 0 \leq t \leq \frac{1}{M} \implies & -t + \frac{Mt^2}{2} \leq -\frac{t}{2} \\ \implies \tilde{f}(t) \leq & f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2 \\ \leq & f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2 \leq f(x) - \alpha t \|\nabla f(x)\|_2^2 \text{ when } \alpha < \frac{1}{2}. \end{aligned}$$

Case 1: $t = 1$. $f(x^+) \leq f(x) - \alpha\|\nabla f(x)\|_2^2$.
Case 2: $t \geq \frac{\beta}{M}$. $f(x^+) \leq f(x) - \frac{\beta\alpha}{M}\|\nabla f(x)\|_2^2$.
Under case 1,

$$f(x^+) - p^* \leq f(x) - p^* - \min\left\{\alpha, \frac{\beta\alpha}{M}\right\}\|\nabla f(x)\|_2^2 \implies f(x^{(k)}) - p^* \leq c^k\left(f(x^{(0)}) - p^*\right),$$

where $c = 1 - \min\left\{2m\alpha, \frac{2\beta\alpha m}{|x|}\right\} < 1$. Once again, we have a geometric decrease.

- The first-order Taylor approximation of $f(x+v)$ around x is $f(x+v) \approx f(x) + \nabla f(x)^T v$.
  - $\nabla f(x)^T v$ is the **directional derivative** of f at x in the direction v.
    - Gives the approximate change in f for a small step v.
  - v is a **descent direction** if $\nabla f(x)^T v < 0$.
- How do we choose v to make the directional derivative as negative as possible?
  - We have to constrain v, or else we could simply make v arbitrarily large.
  - **Normalized Steepest Descent Direction**:

    $$\Delta x_{nsd} = \arg\min_{v:\|v\|=1}\left\{\nabla f(x)^T v\right\} = \arg\min_{v:\|v\|\leq 1}\left\{\nabla f(x)^T v\right\}$$

    with respect to some norm $\|\cdot\|$.
    - A normalized steepest descent direction is a step of unit norm that gives the largest decrease in the linear approximation of f.
    - Suppose $\|v\| < 1$, and $\nabla f(x)^T v < 0$. Consider $\tilde{v} = v \cdot \frac{1}{\|v\|} \implies \nabla f(x)^T \tilde{v} < \nabla f(x)^T v$. Now, suppose $\nabla f(x)^T v = 0$. $\tilde{v} = v \cdot \frac{1}{\|v\|} \implies \nabla f(x)^T \tilde{v} = \nabla f(x)^T v$. Lastly, suppose $\nabla f(x)^T v > 0$. This is not a minimizer, since we could always choose $v = -\nabla f(x)$. Therefore, we won't have this case.
- **Dual Norm**: $\|x\|_* = \sup_{\|v\|\leq 1} x^T v$.
  - $\nabla f(x)^T \Delta x_{nsd} = \min_{v:\|v\|\leq 1}\left\{\nabla f(x)^T v\right\} = -\sup_{\|v\|\leq 1} x^T v = -\|\nabla f(x)\|_*$.
  - If $\|x\|$ is an ℓ_p norm, then $\|x\|_*$ is an ℓ_q norm, where $\frac{1}{p} + \frac{1}{q} = 1$.
- For an unnormalized steepest descent, $\Delta x_{sd} = \|\nabla f(x)\|_* \Delta x_{nsd}$. So, for the steepest descent step,

  $$\nabla f(x)^T \Delta x_{sd} = \|\nabla f(x)\|_* \nabla f(x)^T \Delta x_{nsd} = -\|\nabla f(x)\|_*^2.$$

  Δx_{sd} is used as the step direction.

- Convergence analysis of steepest descent. Start with the fact that vector norms are equivalent in the sense that $\exists \gamma, \tilde{\gamma} \ni \|x\| \geq \gamma\|x\|_2, \|x\|_* \geq \tilde{\gamma}\|x\|_2$. In addition, we will use the fact that $f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{M}{2}\|y-x\|_2^2$.
  For steepest descent, we use $x^+ = x + t\Delta x_{sd}$ as our step.

$$\begin{aligned} f\left(x + t\Delta x_{sd}\right) \leq & f(x) + t\underbrace{\nabla f(x)^T \Delta x_{sd}}_{=-\|\nabla f(x)\|_*^2} + \frac{Mt^2}{2} \cdot \|\Delta x_{sd}\|_2^2 \\ \leq & f(x) - t\|\nabla f(x)\|_*^2 + \frac{Mt^2}{2\gamma^2}\|\nabla f(x)\|_*^2. \end{aligned}$$

Algorithm 7 Steepest Descent

Require: Initial $x^{(0)}$ and some $\epsilon > 0$.
1: **while** $|x^{(k+1)} - x^{(k)}| > \epsilon$ **do**
2: Compute $\nabla f(x^{(k)})$ and $\Delta x_{sd} = \|\nabla f(x)\|_* \Delta x_{nsd}$
3: Compute $t^* = \arg\min_{t\in\mathbb{R}} f\left(x^{(k)} - t\Delta x_{nsd}^{(k)}\right)$.
4: Update $x^{(k+1)} = x^{(k)} - t^*\Delta x_{sd}^{(k)}$.
5: **end while**
6: **return** $x^{(k+1)}$.

It can be shown that the right hand side is minimized when $t = \frac{\gamma^2}{M}$. When this is the case,

$$f(x) - t\|\nabla f(x)\|_*^2 + \frac{Mt^2}{2\gamma^2}\|\nabla f(x)\|_*^2 = f(x) - \frac{\gamma^2}{2M}\|\nabla f(x)\|_*^2.$$

Using the backtracking line search condition,

$$f(x) - \frac{\gamma^2}{2M}\|\nabla f(x)\|_*^2 \le f(x) + \frac{\alpha\gamma^2}{M}\nabla f(x)^T\Delta x_{sd}.$$

So,

$$\begin{aligned}
f(x^+) \le& f\left(x + t\Delta x_{sd}\right) \le f(x) - \alpha\cdot\min\left\{1, \frac{\beta\gamma^2}{M}\right\}\cdot\|\nabla f(x)\|_*^2 \\
\le& f(x) - \alpha\tilde{\gamma}^2\cdot\min\left\{1, \frac{\beta\gamma^2}{M}\right\}\cdot\|\nabla f(x)\|_2^2; \\
(f(x^+) - p^*) \le& (f(x) - p^*) - \alpha\tilde{\gamma}^2\cdot\min\left\{1, \frac{\beta\gamma^2}{M}\right\}\cdot\|\nabla f(x)\|_2^2 \\
\le& c(f(x) - p^*), \text{ where } c = 1 - 2m\alpha\tilde{\gamma}^2\cdot\min\left\{1, \frac{\beta\gamma^2}{M}\right\} < 1.
\end{aligned}$$

This means that $f(x^{(k)}) - p^* \le c^k(f(x^{(0)}) - p^*)$, which is a geometric convergence rate.

- For a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, the steepest descent method for the ℓ_2 norm coincides with the gradient descent method.
  - *Proof*: $\Delta x_{nsd} = \arg\min_{v:\|v\|_2=1} \nabla f(x)^T v$. Note that

$$\nabla f(x)^T v \ge -\left|\nabla f(x)^T v\right| \overset{\text{CS}}{\ge} -\|\nabla f(x)\|_2\|v\|_2 = -\|\nabla f(x)\|_2.$$

    This means that the minimum value cannot be smaller than $-\|\nabla f(x)\|_2$. In addition, setting $v = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$ satisfies the norm condition, and will form the steepest gradient. Therefore, we get that $v = \arg\min_{v:\|v\|_2=1} \nabla f(x)^T v$.

- **Coordinate Descent**: Improve the function coordinate-by-coordinate. Start with some $x^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)})$. In the kth iteration, update the solution by changing a single coordinate. Consider

$$g_i^{(k)}(y) = f(x_1^{(k)}, \ldots, x_{i-1}^{(k)}, y, x_{i+1}^{(k)}, \ldots, x_n^{(k)}).$$

Algorithm 8 Coordinate Descent

Require: Initial $x^{(0)}$ and some $\epsilon > 0$.
1: **while** $|x^{(k+1)} - x^{(k)}| > \epsilon$ **do**
2: Fix an index i.
3: Compute $y^* = \arg\min_{y\in\mathbb{R}} f\left(x_1^{(k)}, \ldots, y, \ldots, x_n^{(k)}\right)$.
4: Update $x^{(k+1)} = \left(x_1^{(k)}, \ldots, y^*, \ldots, x_n^{(k)}\right)$.
5: **end while**
6: **return** $x^{(k+1)}$.

  - This is a special case of the ordinary algorithm that uses the maximum gradient.

- Choosing y^* is a univariate optimization problem.
- **Cyclical Coordinate Descent**: Cycle through coordinates in some permutation of $\{1, \ldots, n\}$.
- Works well when the multivariate function is smooth. However, it does not work well when the function is not smooth.
  * Choosing a single coordinate is being made in a sub-optimal manner. We don't leverage information about f in this choice.
- *Theorem*: Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. The steepest descent direction v at a point $x \in \mathbb{R}^n$, under the ℓ_1 norm, is a vector that satisfies $v_i = -\text{sgn}(\nabla f x)_i)$ for some index i where $|\nabla f(x)_i|$ is maximized, and $v_j = 0$ for all $j \neq i$.
  * Coordinate descent is much better when we can use the gradient vector of f.
  * *Proof*: $\Delta x_{nsd} = \arg\min_{v:\|v\|_1=1} \nabla f(x)^T v$.

$$\begin{aligned}\nabla f(x)^T v &= \sum_{i=1}^{n} \nabla f(x)_i v_i \geq -\sum_{i=1}^{n} |\nabla f(x)_i| \cdot |v_i| \\ &\geq -\max_i |\nabla f(x)_i|, \text{ since } |v_i| \leq 1 \ \forall i.\end{aligned}$$

    Define $v_i = -\text{sgn}(\nabla f(x)_i) \cdot \mathbb{I}(i = \arg\max_i |\nabla f(x)_i|)$.

- Consider the quadratic norm $\|z\|_P = \sqrt{z^T P z} = \left\|P^{1/2} z\right\|_2$, where P is positive-definite. The dual norm is given by $\|z\|_* = \left\|P^{-1/2} z\right\|_2$, with a steepest descent direction given by $\Delta x_{sd} = -P^{-1} \nabla f(x)$.
  - When $P = \nabla^2 f(x)$, $\Delta x_{sd} = -\nabla^2 f(x)^{-1} \nabla f(x)$. In other words, Newton's method is the steepest descent method for the quadratic Hessian norm.
- **Newton Step**: $\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$.

Algorithm 9 Multivariate Newton's Method

Require: Initial $x^{(0)}$, $\lambda^2 = 1$, and some $\epsilon \in (0, \frac{\lambda^2}{2})$.

1: **while** $\frac{\lambda^2}{2} > \epsilon$ **do**
2: Compute $\Delta x_{nt} = -\nabla f(x)^{-1} \nabla f(x)$.
3: Compute $\lambda^2; \nabla f(x)^T \nabla f(x)^{-1} \nabla f(x)$.
4: Choose t by backtracking line search.
5: Update $x = x + t\Delta x_{nt}$.
6: **end while**

  - Positive-definitenes of $\nabla^2 f(x)$ implies that $\nabla f(x)^T \Delta x_{nt} = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) < 0$, unless $\nabla f(x) = 0$.
  - Interpretations of Δx_{nt}:
    * The Newton step minimizes the second-order approximation of f at x. The second-order Taylor approximation of f at x is

$$\hat{f}(x + v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v,$$

      which is minimized at $v = \Delta x_{nt}$.
      · If f is nearly quadratic, then $x + \Delta x_{nt}$ is the exact minimizer of f, which results in extremely fast convergence.
      · $\hat{f}$ is accurate when x is near x^*.
    * The Newton step is also the steepest descent direction at x for the quadratic norm defined by the Hessian $\nabla^2 f(x)$. In other words,

$$\|u\|_{\nabla^2 f(x)} = \left[u^T \nabla^2 f(x) u\right]^{1/2}.$$

      · The quadratic norm converges very fast when the Hessian has a small condition number.
      · For $x \approx x^*$, a good choice is $\nabla^2 f(x^*) \approx \nabla^2 f(x)$.

* If we linearize the optimality condition $\nabla f(x^*) = 0$ near x, we obtain

$$\nabla f(x + v) \approx \nabla f(x) + \nabla^2 f(x)v = 0,$$

which is solved by $v = \Delta x_{nt}$.

– The Newton step is independent of linear changes of coordinates.

* Suppose $T \in \mathbb{R}^{n\times n}$ is nonsingular, and define $\bar{f}(y) = f(Ty)$. Then,

$$\nabla \bar{f}(y) = T^T \nabla f(x), \text{ and } \nabla^2 \bar{f}(y) = T^T \nabla^2 f(x) T,$$

where $x = Ty$. The Newton step for $\bar{f}$ at y is

$$\nabla y_{nt} = -\left(T^T \nabla^2 f(x) T\right)^{-1} T^T \nabla f(x) = T^{-1} \Delta x_{nt}.$$

Hence, $x + \Delta x_{nt} = T(y + \Delta y_{nt})$.

- Convergence analysis of Newton's method. We show that there are numbers η and γ with $0 \le \eta \le \frac{m^2}{L}$ and $\gamma > 0$, such that if $\left\|\nabla f(x^{(k)})\right\|_2 \ge \eta$, then

$$f\left(x^{(k+1)}\right) - f\left(x^{(k)}\right) \le -\gamma.$$

In addition, if $\left\|\nabla f(x^{(k)})\right\|_2 < \eta$, then the backtracking line search selects $t^{(k)} = 1$, and

$$\frac{L}{2m^2}\left\|\nabla f(x^{(k+1)})\right\|_2 \le \left(\frac{L}{2m^2}\left\|\nabla f(x^{(k)})\right\|_2\right)^2.$$

First, start with the Newton decrement.

$$\hat{f}(x + v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v.$$

The left hand side is minimized when $v = \Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$. This implies that

$$\begin{aligned}
\inf_y \hat{f}(y) =& f(x) - \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) + \frac{1}{2} \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla^2 f(x) \nabla^2 f(x)^{-1} \nabla f(x) \\
=& f(x) - \frac{1}{2} \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \\
=& f(x) - \frac{1}{2}\lambda^2(x), \text{ where } \lambda(x) = \sqrt{\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)}; \\
\lambda(x) =& \sqrt{\Delta x_{nt} \nabla^2 f(x) \Delta x_{nt}} \implies -\lambda^2(x) = \nabla f(x)^T \Delta x_{nt}; \\
-\lambda^2(x) =& \frac{d}{dt} f(x + \Delta x_{nt} t)|_{t=0};\ \lambda^2(x) = \Delta x_{nt} \nabla^2 f(x) \Delta x_{nt} \ge m \left\|\Delta x_{nt}\right\|_2^2; \\
\lambda^2(x) =& \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) \ge \frac{1}{M} \cdot \|\nabla f(x)\|_2^2
\end{aligned}$$

Next, we discuss the damped Newton phase, which is the first result we need to find. Assume $\|\nabla f(x)\|_2 \ge \eta > 0$. By strong convexity,

$$\begin{aligned}
& \nabla^2 f(x) \succ M\boldsymbol{I} \\
\implies & f(x + t\Delta x_{nt}) \le f(x) + t\nabla f(x)^T \Delta x_{nt} + \frac{M}{2} t^2 \left\|\Delta x_{nt}\right\|_2^2 \\
& \le f(x) - t\lambda^2(x) + \frac{M}{2m} t^2 \lambda^2(x).
\end{aligned}$$

Now, set $\tilde{t} = \frac{m}{M}$. Then,

$$f(x + \tilde{t}\Delta x_{nt}) \le f(x) - \frac{m}{2M} \cdot \lambda^2(x) \le f(x) - \alpha \tilde{t} \lambda^2(x).$$

Therefore, $\tilde{t} = \frac{m}{M}$ satisfies the BLS exit condition, so we will result in $t \ge \beta \frac{m}{M}$. This means that

$$f(x^+) - f(x) \le \alpha t \lambda^2(x) \le -\alpha\beta \frac{m}{M} \cdot \lambda^2(x) \le -\alpha\beta \frac{m}{M} \cdot \frac{1}{M} \cdot \|\nabla f(x)\|_2^2 = -\alpha\beta \frac{m}{M^2} \cdot \eta^2 = -\gamma.$$

Next, we show quadratic convergence, which is triggered when $\left\|\nabla f(x^{(k)})\right\|_2 < \eta$. Further assume that $\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_2 \leq L \|x-y\|_2$ (assuming the Hessian is Lipschitz continuous), and that $\eta \leq 3(1-2\alpha)\frac{m^2}{L}$. For $t \geq 0$,

$$\begin{aligned}&\left\|\nabla^2 f(x+t\Delta x_{nt}) - \nabla^2 f(x)\right\|_2 \leq tL \|\Delta x_{nt}\|_2\\ \implies &\left|\Delta x_{nt}^T \left(\nabla f(x+t\Delta x_{nt}) - \nabla^2 f(x)\right) \Delta x_{nt}\right| \leq tL \|\Delta x_{nt}\|_2^3.\end{aligned}$$

Define $\tilde{f}(t) = f(x+t\Delta x_{nt})$. Then,

$$\begin{aligned}\tilde{f}''(t) =&\Delta x_{nt}^T \nabla^2 f(x+t\Delta x_{nt})\Delta x_{nt} \implies \tilde{f}''(t) \leq \tilde{f}''(0) + tL\|\Delta x_{nt}\|_2^3 \leq \lambda^2(x) + t\frac{L}{m^{3/2}}\cdot\lambda^3(x);\\ &\text{Integrating both sides from 0 to } t,\\ \tilde{f}'(t) \leq&\tilde{f}'(0) + t\lambda^2(x) + t^2\frac{L}{2m^{3/2}}\lambda^3(x) \leq -\lambda^2(x) + t\lambda^2(x) + t^2\frac{L}{2m^{3/2}}\lambda^3(x);\\ \tilde{f}'(t) \leq& -\lambda^2(x) + t\lambda^2(x) + \frac{t^2}{2}\cdot\frac{L}{m^{3/2}}\lambda^3(x);\\ &\text{Integrating again from 0 to } t,\\ \implies \tilde{f}(t) \leq&\tilde{f}(0) - t\lambda^2(x) + \frac{t^2}{2}\lambda^2(x) + \frac{t^3}{6}\cdot\frac{L}{m^{3/2}}\lambda^3(x).\end{aligned}$$

Now, set $t = 1$.

$$f(x+\Delta x_{nt}) \leq f(x) - \frac{1}{2}\lambda^2(x) + \frac{L}{6m^{3/2}}\lambda^3(x).$$

Recall that $\|\nabla f(x)\|_2 \leq \eta \leq 3(1-2\alpha)\cdot\frac{m^2}{L}$. Then,

$$\lambda(x) \leq 3(1-2\alpha)\cdot\frac{m^{3/2}}{L} \implies f(x+\Delta x_{nt}) \leq f(x) - \lambda^2(x)\left(\frac{1}{2} - \frac{L\lambda(x)}{6m^{3/2}}\right) \leq f(x) - \alpha\lambda^2(x) = f(x) + \alpha\nabla f(x)^T\Delta x_{nt}.$$

Now,

$$\begin{aligned}\left\|\nabla f(x^+)\right\|_2 =& \left\|\nabla f(x+\Delta x_{nt}) - \nabla f(x) - \nabla^2 f(x)\Delta x_{nt}\right\|_2 = \left\|\int_0^1 \left(\nabla^2 f(x+\Delta x_{nt}) - \nabla^2 f(x)\right)\Delta x_{nt} dt\right\|_2\\ \leq&\frac{L}{2}\|\Delta x_{nt}\|_2^2 \text{ by Lipschitz continuity}\\ =&\frac{L}{2}\left\|\nabla^2 f(x)^{-1}\nabla f(x)\right\|_2^2 \leq \frac{L}{2m^2}\|\nabla f(x)\|_2^2.\end{aligned}$$

This proves the quadratic convergence. The number of iterations needed will be

$$6 + \frac{M^2L^2/m^5}{\alpha\beta\cdot\min\{1, 9(1-2\alpha)^2\}}\left(f(x^{(0)}) - p^*\right).$$

- **Example**: $f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2)$. This is strongly convex, with $m = \min\{1,\gamma\}$, and $M = \max\{1,\gamma\}$. If γ is large, this can make the problem very hard. Note that the Hessian here is very simple, and does not involve x. ■
- **Example**: $f(x) = \exp\{x_1 + 3x_2 - 0.1\} + \exp\{x_1 - 3x_2 - 0.1\} + \exp\{-x_1 - 0.1\}$. Here, BLS will reach the solution, but it will take a lot of iterations. Exact line search converges extremely quickly. Newton's method is in between. ■
- Exact line search is faster, but also can be more computationally intense.
- **Example**: Logistic Regression. Suppose Y_i is binary data, where $P(Y_i = 1|X_{i1},\ldots,X_{id}) = p_i$, and $\log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{X}_i^T\beta$. We want to find $\beta^* = \arg\min_\beta -\ell(\beta)$, where

$$\ell(\beta) = \sum_{i=1}^n \left[y_i\boldsymbol{X}_i^T\beta - \log\left\{1 + \exp\{\boldsymbol{X}_i^T\beta\}\right\}\right].$$

  - $-\ell(\beta)$ is strongly convex.
- **Stochastic Gradient Descent**: Consider minimizing an average of functions $\arg\min_x \frac{1}{n}\sum_{i=1}^n f_i(x)$. Approximate or estimate the "sample mean" by taking a random subsample.

- The full algorithm is given by $x^{(k)} = x^{(k-1)} - t_k \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^{(k-1)})$ (**batch training**), which can be very time-consuming, and unnecessary.
- One option is to use a single random sample from $y_1, \ldots, y_n$. In other words, $x^{(k)} = x^{(k-1)} - t_k \nabla f_{i_k}(x^{(k-1)}$, where i_k follows a Discrete Uniform.
- Another option is **mini-batching**; that is, randomly choose $I_k \subset \{1, \ldots, n\}$, where $|I_k| = b < n$. Then, use $x^{(k)} = x^{(k-1)} - t_k \frac{1}{b} \sum i \in I_k \nabla f_i(x^{(k-1)})$.
  - There is a tradeoff between computation time and accuracy. If $b = n$, then variance decreases and runtime is high, vice versa.

- **Quasi-Newton Methods**: Use

$$f(x^{(k)} + \delta) - f(x^{(k)}) \approx \nabla f(x^{(k)})^T \delta + \frac{1}{2} \delta^T \boldsymbol{Q}^{(k)} \delta,$$

where $\delta \in \mathbb{R}^p$ has small ℓ_2-norm, and $\boldsymbol{Q}^{(k)}$ is some positive-definite matrix.

  -
  - While Newton's method has many advantages, computing and storing the Hessian at every step may not be computationally feasible.
  - Quasi-Newton methods have the same advantages as Newton methods (such as rapid convergence), and are less computationally expensive.
  - The precise values of the gradient don't matter as much when we are far away from the optimal value.
  - The descent direction is $\delta = -(\boldsymbol{Q}^{(k)})^{-1} \nabla f(x^{(k)})$.
    - $\boldsymbol{Q}^{(k)} = \boldsymbol{I}$ would result in ordinary gradient descent, and $\boldsymbol{Q}^{(k)} = \nabla^2 f(x^{(k)})$ results in Newton's method.
  - Let $\boldsymbol{P}^{(k)} = (\boldsymbol{Q}^{(k)})^{-1}$. Update $\boldsymbol{P}$ iteratively, where $\boldsymbol{P}^{(0)} = \boldsymbol{I}$, and $\boldsymbol{P}^{(k+1)} = \boldsymbol{P}^{(k)} + \boldsymbol{U}^{(k)}$.
    - Faster updates to $\boldsymbol{P}$ will result in efficient methods.
  - **BFGS Method**:

Algorithm 10 BFGS Method

Require: Initial x_0, some approximate inverted Hessian $\boldsymbol{B}_0$, and some $\epsilon > 0$.

1: **while** $|x_{k+1} - x_k| > \epsilon$ **do**
2: Calculate $p_k = -\boldsymbol{B}_k \nabla f(x_k)$.
3: Perform a line search to find step size α_k in the direction p_k.
4: Set $s_k = \alpha_k p_k$ and update $x_{k+1} = x_k + s_k$.
5: Compute $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.
6: Update $\boldsymbol{B}_{k+1} = \boldsymbol{B}_k + \frac{(s_k^T y_k + y^T \boldsymbol{B}_k y_k)(s_k s_k^T)}{(s_k^T y_k)^2} - \frac{\boldsymbol{B}_k y_k s_k^T + s_k y_k^T \boldsymbol{B}_k}{s_k^T y_k}$.
7: **end while**

    - Guaranteed to be low-rank.

9 ST 779: Advanced Probability for Statistical Inference

Instructor: Dr. Subhashis Ghoshal
Semester: Spring 2025
Main Textbook: Resnick, *A Probability Path*

9.1 Introduction to Measure Theory

Return to Table of Contents

- **Bolzano-Weierstrass Theorem**: Every bounded sequence has a convergent subsequence.
- **Interior Point**: A point $x \in G$ if $\exists \epsilon > 0$ such that $\{y : \|y - x\| < \epsilon\} \subset G$.
  - G is open if all points are interior points.
- **Outcome**, or $\boldsymbol{\omega}$: A result of an experiment.
- **Sample Space**, or $\boldsymbol{\Omega}$: The collection of all outcomes.
- **Event**: Some subset of Ω.
  - Events $A = B$ if for any $x \in \Omega$, $x \in A$ iff $x \in B$.
  - $A \subset B$ if $x \in A \implies x \in B$.
  - **Exhaustive**: $A_1, \ldots, A_n$ such that $\bigcup_{i=1}^n A_i = \Omega$.
- **Empty Set**, or $\boldsymbol{\varnothing}$: Empty set.
- **Union**, or $\cup$: $A \cup B = \{x \in A \text{ or } x \in B\}$.
  - **Finite Union**: $\bigcup_{i=1}^n A_i = A_1 \cup \cdots \cup A_n$.
  - **Disjoint Union**: $A \sqcup B$, where A and B are disjoint.
- **Intersection**, or $\cap$: $A \cap B = \{x \in A \text{ and } x \in B\}$.
  - **Finite Intersection**: $\bigcap_{i=1}^n A_i = A_1 \cap \cdots \cap A_n$.
- **Complement**, or $\boldsymbol{A^c}$: $A^c = \{x \in \Omega : x \notin A\}$.
- **Difference**, or $\boldsymbol{A \setminus B}$: $A \setminus B = \{x \in A \text{ and } x \notin B\}$.
- **Symmetric Difference**, or $\boldsymbol{A \Delta B \equiv A \oplus B}$: $A \Delta B = (A \cup B) \setminus (A \cap B)$.
- Properties of set operations:
  - **Commutativity**: $A \cup B = B \cup A$ (works for $\cap$ and Δ too).
  - **Associativity**: $A \cup (B \cup C) = (A \cup B) \cup C$ (works for $\cap$ and Δ too).
  - **Distributivity**: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$, and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
  - **De Morgan's Law**: $(A \cup B)^c = A^c \cap B^c$, and $(A \cap B)^c = A^c \cup B^c$.
- **Mutually Exclusive**, or **Disjoint**: A and B are disjoint if $A \cap B = \varnothing$.
  - **Partition**: A collection of disjoint and exhaustive events $A_1, \ldots, A_n$.
  - **Disjointification**: Converting a sequence of events $\{A_n\}$ into disjoint sequence $D_j = A_j \setminus \bigcup_{i=1}^{j-1} A_j$ such that $\bigcup_{i=1}^n A_i = \bigsqcup_{i=1}^n D_i$.
- **lim sup**: $\limsup_{n\to\infty} A_n = \bigcap_{n=1}^\infty \bigcup_{m=n}^\infty A_m = \{x : x \text{ belongs to infinitely many } A_n\}$.
  - lim sup is the largest possible subsequential limit.
  - For all convergent subsequences $x_{n_k} \to x$, there will always be a larger possible limit.
  - Is the collection of all events that appear in the tail of a sequence.
  - Suppose $x \in \limsup n$. This means that $\forall n$, there exists a $k \geq n$ such that $x \in A_k$.
- **lim inf**: $\liminf_{n\to\infty} A_n = \bigcup_{n=1}^\infty \bigcap_{m=n}^\infty A_m = \{x : x \text{ belongs to all but finitely many } A_n\}$.
  - $(\limsup_{n\to\infty} A_n)^c = \liminf_{n\to\infty} A_n^c$.

- x_n converges to x when $\limsup = \liminf$.
- The union of all elements appearing in every set after a point of time.
- Suppose $x \in \liminf n$. This means that $\exists n$ such that $\forall k \geq n$, $x \in A_k$.

- **Example**: Let A_n, A, B_n, B be subsets of Ω. Show that $\limsup_{n\to A_n}(A_n \cup B_n) = \limsup_{n\to\infty} A_n \cup \limsup_{n\to\infty} B_n$. Then, if $A_n \to A$ and $B_n \to B$, does $A_n \cup B_n \to A \cup B$, and/or $A_n \cap B_n \to A \cap B$?

$$\limsup_{n\to\infty}(A_n \cup B_n) = \bigcap_{n=1}^{\infty}\bigcup_{m\geq n}(A_m \cup B_m) = \left(\bigcap_{n=1}^{\infty}\bigcup_{m\geq n} A_m\right) \cup \left(\bigcap_{n=1}^{\infty}\bigcup_{m\geq n} B_m\right) = (\limsup_{n\to\infty} A_n \cup \limsup_{n\to\infty} B_n).$$

If $A_n \to A$ and $B_n \to B$, then $A_n \cup B_n \to A \cup B$, and $A_n \cap B_n \to A \cap B$. The previous result will let us show this. We need to show that $\limsup_{n\to\infty}(A_n \cap B_n) = \liminf_{n\to\infty}(A_n \cap B_n) = A \cap B$, and that $\limsup_{n\to\infty}(A_n \cup B_n) = \liminf_{n\to\infty}(A_n \cup B_n) = A \cup B$.
$\limsup_{n\to\infty}(A_n \cup B_n) = \limsup_{n\to\infty} A_n \cup \limsup_{n\to\infty} B_n = A \cup B$. Using DeMorgan's law,

$$\liminf_{n\to\infty}(A_n \cap B_n) = \left(\limsup_{n\to\infty}(A_n^c \cup B_n^c)\right)^c = \left(\limsup_{n\to\infty} A_n^c \cup \limsup_{n\to\infty} B_n^c\right)^c = (A^c \cup B^c)^c = A \cap B.$$

We know that $\liminf_{n\to\infty} A_n \subset \limsup_{n\to\infty} A_n$ for arbitrary set A_n, so if we can show that $\liminf_{n\to\infty} A_n \supset \limsup_{n\to\infty} A_n$, then we are done.
If $x \in A \cup B$, then $x \in A$ or $x \in B$. If $x \in A$, then $x \in \bigcap_{m\geq n} A_m$ for some n, since $\bigcap_{m\geq n} A_m \uparrow A$. Thus, $x \in \bigcap_{m\geq n} A_m \subset_{m\geq k} A_m \subset \bigcup_{m\geq k}(A_m \cup B_m)$ for any k. Hence, $x \in \bigcap_{k=1}^{\infty}\bigcup_{m\geq k}(A_m \cup B_m) = \liminf_{n\to\infty}(A_n \cup B_n)$. This is the same conclusion if $x \in B$. Therefore, $\liminf_{n\to\infty}(A_n \cup B_n) \supset A \cup B$. ■

- **Example**: Suppose $a_n > 0$, $b_n > 1$, and $a_n \to 0$, $b_n \to 1$. Define $A_n := \{x : a_n \leq x < b_n\}$. Find $\limsup_{n\to\infty} A_n$ and $\liminf_{n\to\infty} A_n$.

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty}\bigcup_{m\geq n}\{x : a_m \leq x < b_m\} = \bigcap_{n=1}^{\infty}\bigcup_{m\geq n}[a_m, b_m);$$

since $a_m > 0$ and $b_m > 1$ strictly, but $a_m \to 0$, then there must exist some point where $\forall m' > m$, $0 < a_{m'} < 1$. Also applying the fact that $b_m \to 1$,

$$\bigcap_{n=1}^{\infty}\bigcup_{m\geq n}[a_m, b_m) = \bigcap_{n=1}^{\infty}(0, b_n) = (0, 1].$$

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty}\bigcap_{m\geq n}\{x : a_m \leq x < b_m\} = \bigcup_{n=1}^{\infty}\bigcap_{m\geq n}[a_m, b_m);$$

Using the same results as discussed above,

$$\bigcup_{n=1}^{\infty}\bigcap_{m\geq n}[a_m, b_m) = \bigcup_{n=1}^{\infty}(a_n, 1] = (0, 1].$$ ■

- **Example**: Suppose $A_n = \{\frac{m}{n} : m \in \mathbb{N}\}$, for $n \in \{1, 2, \dots\}$. Find $\liminf_{n\to\infty} A_n$ and $\limsup_{n\to\infty} A_n$.

Observe that $A_2 \supset A_1$, but $A_3 \not\supset A_2$. More generally, $A_p \subset A_q$ when q is divisible by p. lim sup is the collection of all events that appear in the tail of a sequence. As q increases, more natural numbers before it are divisible by it (in other words, the number of prime numbers decreases as the number itself increases). This means that $\limsup_{n\to\infty} A_n = \mathbb{Q}^+$, the set of positive rational numbers.
Similarly, A_p and A_q for any positive integers $p < q$ will only contain $\mathbb{N}$ in common (think of A_2 and A_3, for example). lim inf is the union of all elements appearing in every set after a point of time, which is only guaranteed to be $\mathbb{N}$ here. Thus, $\liminf_{n\to\infty} A_n = \mathbb{N}$. ■

- **Example**: Let f_n, f be real functions on Ω. Show $\{\omega : f_n(\omega) \not\to f(\omega)\} = \bigcup_{k=1}^{\infty}\bigcap_{N=1}^{\infty}\bigcup_{n=N}^{\infty}\{\omega : |f_n(\omega) - f(\omega)| \geq \frac{1}{k}\}$.

This is asking us to show that convergence fails at ω. Recall that the $\epsilon - \delta$ definition of convergence is that $\forall\epsilon > 0$, $\exists N : \forall n \geq N$, $|f_n(\omega) - f(\omega)| < \epsilon$. So, if convergence fails, then $\exists\epsilon > 0 : \forall N \geq 1$, $\exists n \geq N$ such that $|f_n(\omega) - f(\omega)| \geq \epsilon = \bigcup_{\epsilon>0}\bigcap_{N=1}^{\infty}\bigcup_{n=N}^{\infty}\{\omega : |f_n(\omega) - f(\omega)| \geq \epsilon\}$.

This is almost the answer, except we have an uncountable ϵ term that we need to make countable. Using the denseness of $\mathbb{R}$, $\forall \epsilon > 0$, $\exists \frac{1}{k} : 0 < \frac{1}{k} < \epsilon$. Thus,

$$\bigcup_{\epsilon>0} \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} \{\omega : |f_n(\omega) - f(\omega)| \geq \epsilon\} = \bigcup_{k=1}^{\infty} \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} \{\omega : |f_n(\omega) - f(\omega)| \geq \frac{1}{k}\}. \blacksquare$$

- **Inverse Map**, or $\boldsymbol{f^{-1}}$: $f^{-1}(A) = \{x \in S : f(x) \in A\} \subset S$.
  - Note that f^{-1} returns a set.
  - $f^{-1}(A^c) = f^{-1}(A)^c$ (works for $\cup$ and $\cap$ too).
- **Cartesian Product**: $\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1 \text{ and } \omega_2 \in \Omega_2\}$.
- **$\boldsymbol{\omega_1}$-Section**, or $\boldsymbol{E_{\omega_1}}$: The ω_1-section of $E \subset \Omega_1 \times \Omega_2$ is $E_{\omega_1} = \{\omega_2 \in \Omega_2 : (\omega_1, \omega_2) \in E\}$.
  - ω_2-section is defined similarly, but notated as E^{ω_2} as a visual aide.
  - **Inclusion Map**, or $\boldsymbol{\iota_{\omega_1}}$: $\iota_{\omega_1} : \Omega_2 \to \Omega_1 \times \Omega_2$ defined by $\iota(\omega_2) = (\omega_1, \omega_2)$.
    - $E_{\omega_1} = \iota_{\omega_1}^{-1}(E)$ for any $E \subset \Omega_1 \times \Omega_2$.
- **Indicator Function**, or $\mathbb{1}$: $\mathbb{1}_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A. \end{cases}$
  - A set can be uniquely determined by its indicator function.
  - Properties of indicator functions:
    - $\mathbb{1}_{A^c} = 1 - \mathbb{1}_A$.
    - $\mathbb{1}_{A \setminus B} = \mathbb{1}_A - \mathbb{1}_B$ if $B \subset A$.
    - $\mathbb{1}_{A \Delta B} = |\mathbb{1}_A - \mathbb{1}_B|$.
    - $\mathbb{1}_{\bigcup_{n=1}^{\infty} A_n} = \sup_{n \geq 1} \mathbb{1}_{A_n}$.
      - Also equal to $\sum_{i=1}^{\infty} \mathbb{1}_{A_n}$ if A_i is pairwise disjoint.
    - $\mathbb{1}_{\bigcap_{n=1}^{\infty} A_n} = \inf_{n \geq 1} \mathbb{1}_{A_n} = \prod_{n=1}^{\infty} \mathbb{1}_{A_n}$.
    - $\mathbb{1}_{\limsup A_n} = \limsup_{n \to \infty} \mathbb{1}_{A_n}$.
    - $\mathbb{1}_{\liminf A_n} = \liminf_{n \to \infty} \mathbb{1}_{A_n}$.
- **Class**: A collection of sets.
- **Power Set**, or $\mathcal{P}$: The collection of all subsets of U.
  - $\mathcal{P}(U) = \{E \subset U\}$.
  - If U is finite, then $|\mathcal{P}(U)| = 2^{|U|}$.
- **Partition**: A class $\mathcal{C}$ of non-empty sets of U such that $C_i \cap C_j = \varnothing$ for $C_i, C_j \in \mathcal{C}$ and $i \neq j$, and $\bigcup_{C \in \mathcal{C}} C = U$.
- **Equivalent**: $x \sim y$ iff x and y are contained in the same element of the partition.
  - Properties of equivalence:
    - Reflexivity: $x \sim x$.
    - Symmetry: If $x \sim y$, then $y \sim x$.
    - Transitivity: If $x \sim y$ and $y \sim z$, then $x \sim z$.
- **Countable**: Anything in one-to-one correspondence with $\mathbb{N}$.
  - A set is countable if we can create a "queue," or "enumeration" of every element in the set without omitting anything.
  - Examples of countable sets include $\mathbb{N}^k$, $\mathbb{Z}^k$, $\mathbb{Q}^k$.
  - Examples of uncountable sets include $(a, b]$, $\{0, 1\}^{\infty}$.
    - *Proof*: Show that $\{0, 1\}^{\infty}$ is uncountable.

      This proof uses Cantor's diagonal argument. Suppose this set is countable. That means that there exists a queue of the elements in this set. Let $s_i = (s_{i1}, s_{i2}, \dots)$ be the elements of the queue. Define x_i as the complement of the diagonal element of s_i, or $x_i = 1 - s_{ii}$. Join the x_i together to form x. By construction of the queue, every nth digit of x is different than s_n, so x is not in this queue. Hence, by contradiction, $\{0, 1\}^{\infty}$ is not countable.

- If A is one-to-one with B and A is uncountable, then B is uncountable.

- **At Most Countable**: A set that is countable and/or finite.
  - If A_n is at most countable, then $\bigcup_{n=1}^{\infty} A_n$ is at most countable.
  - If A is countable, then $B \subset A$ is at most countable.

- If we have a class of measurable events, then every event we can form using countably many set operations on this class will also be measurable.

- **Semifield**: The class $\mathcal{C} \in \mathcal{P}(\Omega)$ such that:
  1. $\varnothing, \Omega \in \mathcal{C}$.
  2. $C_1, C_2 \in \mathcal{C} \implies C_1 \cap C_2 \in \mathcal{C}$.
     - This means that $\mathcal{C}$ is closed under finite intersection.
  3. $C \in \mathcal{C} \implies C^c$ can be written as a finite union of disjoint sets from $\mathcal{C}$.

  - Examples of semifields include $\{\varnothing, \Omega\}$ and the set of half-open intervals.
  - **Example**: If $\mathcal{S}_1$ and $\mathcal{S}_2$ are two semifields of subsets of Ω, show that the class $\mathcal{S}_1 \cap \mathcal{S}_2 = \{A_1 \cap A_2 : A_1 \in \mathcal{S}_1, A_2 \in \mathcal{S}_2\}$ is a semifield of subsets of Ω.

    Verify that the definition of a semifield is upheld. Let $\mathcal{S} = \mathcal{S}_1 \cap \mathcal{S}_2$:
    1. $\Omega \in \mathcal{S}$: Note that $\mathcal{S}_1$ and $\mathcal{S}_2$ are semifields, so Ω is in both of them. $\Omega \cap \Omega = \Omega \in \mathcal{S}$ ✓
    2. $\varnothing \in \mathcal{S}$: Note that $\mathcal{S}_1$ and $\mathcal{S}_2$ are semifields, so $\varnothing$ is in both of them. $\varnothing \cap \varnothing = \varnothing \in \mathcal{S}$ ✓
    3. Let $S_{1i}, S_{2i} \in \mathcal{S}_i$. $\underbrace{(S_{11} \cap S_{21})}_{\in \mathcal{S}_1} \cap \underbrace{(S_{12} \cap S_{22})}_{\in \mathcal{S}_2} \in \mathcal{S}$ ✓
    4. Suppose $S_i \in \mathcal{S}_i$. Once again, since $\mathcal{S}_1$ and $\mathcal{S}_2$ are semifields, we can represent S_i^c as $\bigsqcup_{j=1}^{n} S_{ji}$.

    $$(S_1 \cap S_2)^c = S_1^c \cup S_2^c = \left(\bigsqcup_{j=1}^{n} S_{j1}\right) \cup \left(\bigsqcup_{j=1}^{n} S_{j2}\right) = \bigsqcup_{j=1}^{n} \underbrace{(S_{j1} \cap S_{j2})}_{\in \mathcal{S}}.$$

    Thus, we can represent $(S_1 \cap S_2)^c$ as a finite disjoint union of elements in $\mathcal{S}$ ✓

    Thus, $\mathcal{S}_1 \cap \mathcal{S}_2$ is a semifield by definition. ■

- **Field**: The class $\mathcal{F} \subset \mathcal{P}(\Omega)$ such that:
  1. $\varnothing, \Omega \in \mathcal{F}$.
  2. $A, B \in \mathcal{F} \implies A \cap B \in \mathcal{F}$.
  3. $A \in \mathcal{F} \implies A^c \in \mathcal{F}$.

  - Any field is a semifield.
  - Examples of fields include $\mathcal{P}(\Omega)$ and finite disjoint unions of half-open intervals in $\mathbb{R}$.
  - If $\{\mathcal{F}_\alpha, \alpha \in \Lambda\}$ are a collection of fields, then the class $\mathcal{F} = \bigcap_{\alpha \in \Lambda} \mathcal{F}_\alpha = \{F : F \in \mathcal{F}_\alpha \forall \alpha \in \Lambda\}$ is a field.
  - Let $\mathcal{C}$ be any collection of subsets of Ω. Consider $\mathcal{F} = \bigcap\{\mathcal{G} : \mathcal{G} \text{ is a field, and } \mathcal{G} \supset \mathcal{C}\}$. Then, $\mathcal{F}$ is a field, $\mathcal{F} \supset \mathcal{C}$, and if $\mathcal{H}$ is any field containing $\mathcal{C}$, then $\mathcal{F} \supset \mathcal{H}$.
  - Fields are closed under all finite set operations.
  - **Example**: Suppose $\mathcal{A}$ is a class of sets containing Ω that satisfies

    $$A, B \in \mathcal{A} \implies A \setminus B = A \cap B^c \in \mathcal{A}.$$

    Show that $\mathcal{A}$ is a field.

    1. $\Omega \in \mathcal{A}$: stated in problem ✓
    2. $\varnothing \in \mathcal{A}$: $\varnothing = \Omega \setminus \Omega$. Since $\Omega \setminus \Omega \in \mathcal{A}$ by the construction of $\mathcal{A}$, $\varnothing \in \mathcal{A}$ ✓
    3. $A \in \mathcal{A} \to A^c \in \mathcal{A}$: Suppose $A \in \mathcal{A}$. $A^c = \Omega \setminus A \in \mathcal{A}$ by construction of $\mathcal{A}$ ✓
    4. Closed under finite intersection. Suppose $A, B \in \mathcal{A}$. $A \cap B = A \setminus B^c \in \mathcal{A}$ ✓

    All of the conditions for a field are satisfied, so $\mathcal{A}$ must be a field. ■

- **Example**: Suppose $\mathcal{A}$ is a class of subsets of Ω such that $\Omega \in \mathcal{A}$, $A \in \mathcal{A} \implies A^c \in \mathcal{A}$, and $\mathcal{A}$ is closed under finite disjoint unions. Show that $\mathcal{A}$ is not guaranteed to be a field.
  Proceed with an example. Let $\Omega = \{1,2,3,4\}$, and let $\mathcal{A}$ be the field generated by two point subsets of Ω, so $\mathcal{A} = \{\varnothing, \{1,2\}, \{1,3\}, \ldots, \{3,4\}, \Omega\}$. $\mathcal{A}$ satisfies the three conditions as described in the problem, and we now show that this is not a field. Consider $\{1,2\}$, and $\{1,3\} \in \mathcal{A}$. The intersection of this is $\{1\}$, which is not in $\mathcal{A}$, since there is only one element. Therefore, $\mathcal{A}$ is not a field. ■
- **Example**: Let $\mathcal{A}$ be a field of subsets of Ω, and define $\bar{\mathcal{A}} = \{A \subset \Omega : \exists A_n \in \mathcal{A} \text{ and } A_n \to A\}$. Show that $\mathcal{A} \subset \bar{\mathcal{A}}$, and that $\bar{\mathcal{A}}$ is a field.

  If $A \in \mathcal{A}$, then we can set $A_n = A \implies \mathcal{A} \subset \bar{\mathcal{A}}$.
  Now, we show that $\bar{\mathcal{A}}$ is a field.
  1. $\Omega \in \bar{\mathcal{A}}$: trivial, since $\Omega \in \mathcal{A}$ ✓
  2. $\varnothing \in \bar{\mathcal{A}}$: trivial, since $\varnothing \in \mathcal{A}$ ✓
  3. Suppose $A \in \mathcal{A}$. Then, $\exists A_n \in \mathcal{A}$, and since $\mathcal{A}$ is a field, $A_n^c \in \mathcal{A}$. $\limsup_{n\to\infty} A_n^c = (\liminf_{n\to\infty} A_n)^c = A^c$, so $A_n^c \to A^c$ ✓
  4. Suppose $A, B \in \bar{\mathcal{A}}$. This means that $A_n, B_n \in \mathcal{A}$. Next, we show that $A_n \cap B_n \to A \cap B$, thus showing that $\bar{\mathcal{A}}$ is closed under finite intersections.

  $$\limsup_{n\to\infty}(A_n \cap B_n) = \bigcap_{k=1}^{\infty} \bigcup_{n\geq k} A_k B_k \subset \limsup_{n\to\infty} A_n = A.$$

  Similarly,

  $$\limsup_{n\to\infty}(A_n \cap B_n) = \bigcap_{k=1}^{\infty} \bigcup_{n\geq k} A_k B_k \subset \limsup_{n\to\infty} B_n = B.$$

  Thus, $\limsup_{n\to\infty}(A_n \cap B_n) \subset AB$.
  In addition,

  $$\liminf_{n\to\infty}(A_n \cap B_n) = \liminf_{n\to\infty} A_n \cap \liminf_{n\to\infty} B_n = A \cap B.$$

  Since $\limsup_{n\to\infty}(A_n \cap B_n) = \liminf_{n\to\infty}(A_n \cap B_n) = A \cap B$, $A_n \cap B_n \to A \cap B$. Therefore, $\bar{\mathcal{A}}$ is closed under finite intersection/union ✓

  Thus, $\bar{\mathcal{A}}$ must be a field. ■

- ***σ*-Field**: The class $\mathcal{A} \subset \mathcal{P}(\Omega)$ such that:
  1. $\varnothing, \Omega \in \mathcal{A}$.
  2. $A_1, A_2, \cdots \in \mathcal{A} \implies \bigcup_{n=1}^{\infty} A_i \in \mathcal{A}$.
     - In other words, $\mathcal{A}$ is closed under countable union.
     - Due to DeMorgan's Law, showing $A_1, A_2, \cdots \in \mathcal{A} \implies \bigcap_{n=1}^{\infty} A_i \in \mathcal{A}$ is equivalent.
  3. $A \in \mathcal{A} \implies A^c \in \mathcal{A}$.
     - In other words, $\mathcal{A}$ is closed under complementation.

  - Any σ-field is a field, so any σ-field is a semifield.
  - Denote $\mathcal{C}\langle\sigma\rangle$ as the smallest σ-field containing class $\mathcal{C}$.
  - Examples of σ-fields include $\{\varnothing, \Omega\}$, $\mathcal{P}(\Omega)$, and the countable-cocountable set.
  - **Example**: Show that the countable-cocountable set (the set of all sets such that each set is countable, or their complement is countable) is a σ-field.

    Denote S as the countable-cocountable field. Verify the definition of a σ-field holds.
    1. $\varnothing \in S$. $\varnothing$ is finite, so it must be countable ✓
    2. $\Omega \in S$: $\Omega^c = \varnothing \in S$ ✓
    3. $A_1, A_2, \cdots \in S \implies \bigcup_{n=1}^{\infty} A_n \in S$: Split into two cases:
       (a) All A_n are countable. Any finite union of countable sets must also be countable ✓
       (b) There exists *finitely* uncountable $A_{k1}, A_{k2}, \ldots, A_{k\ell}$. Since we are using the countable co-countable field, if A_{ki} is uncountable, then A_{ki}^c must be countable. We would then have an intersection of countable sets, which is countable ✓
    4. $A \in S \implies A^c \in S$: If A is uncountable, then A^c must be countable, once again since we are using the countable co-countable field ✓

With all of the conditions satisfied, S must be a σ-field. ■

– **Example**: Is the union of a countable collection of σ-fields $\mathcal{B}_j$, $j \geq 1$ a σ-field?

No. Let $\Omega = \mathbb{N} = \{1, 2, \dots\}$, and $\mathcal{B} = \sigma\langle\{\{1\}, \dots, \{j\}\}\rangle$. Clearly, the $\mathcal{B}_j$'s are nested, with $\mathcal{B}_1 \subset \mathcal{B}_2 \subset \dots$.

Claim that $\bigcup_{j=1}^{\infty} \mathcal{B}_j$ is not a σ-field. Take $c_j = \begin{cases} \{j\} & , j \text{ is odd} \\ \varnothing & , j \text{ is even} \end{cases}$. $\bigcup_{j=1}^{\infty} c_j \notin \bigcup_{j=1}^{\infty} \mathcal{B}_j$, so $\bigcup_{j=1}^{\infty} \mathcal{B}_j$ is not closed under union. Therefore, $\bigcup_{j=1}^{\infty} \mathcal{B}_j$ is not a σ-field. ■

– The intersection of σ-fields is a σ-field.

– **Example**: Let $\mathcal{P}$ be a countable partition of Ω. Show that unions of sets of $\mathcal{P}$ form a σ-field.

Since the partition is countable, we can express $\mathcal{P} = \{c_1, c_2, \dots\}$.

1. $\Omega \in \mathcal{P}$: If we include every element of $\mathcal{P}$ in our union, then we form Ω by the definition of partitions ✓
2. $\varnothing \in \mathcal{P}$: If we don't include any element of $\mathcal{P}$ in our union, then we will form $\varnothing$ ✓
3. Closed under complementation: Suppose $A_n \in \mathcal{P}$. Since $\mathcal{P}$ is a partition of Ω, A_n^c is all the elements of $\mathcal{P}$ not in A_n, so $A_n^c \in \mathcal{P}$ ✓
4. Closed under countable intersection: Suppose $A_n \in \mathcal{P}$. $\bigcap_{n=1}^{\infty} A_n = (\bigcup_{n=1}^{\infty} A_n^c)^c$. We have already shown $\mathcal{P}$ is closed under complementation, so $A_n^c \in \mathcal{P}$. In addition, since we are interested in the unions of sets in $\mathcal{P}$, then by construction $\mathcal{P}$ is closed under union. Therefore, $(\bigcup_{n=1}^{\infty} A_n^c)^c \in \mathcal{P}$ ✓

Thus, the unions of sets in $\mathcal{P}$ form a σ-field. ■

– σ-fields are closed under all countable set operations.

– **Example**: Show that a σ-field $\mathcal{F}$ cannot be countably infinite.

Proceed with contradiction. Suppose $\mathcal{F}$ is countably infinite. This means that we can write out all of the elements using $\mathcal{F} = \{F_1, F_2, \dots\}$.
Consider the partition generated by $\mathcal{F}$ consisting of the form $\bigcap_{i=1}^{\infty} F_i^{\epsilon_i}$, where $\epsilon_i = 1$ means we use F_i^c, F_i otherwise. Discard all empty sets within this partition so the elements in this set are all unique. Denote $\mathcal{P}$ as this collection of non-empty sets. $\mathcal{P} \subset \mathcal{F}$, which means that $\mathcal{P}$ is finite or countable. We can express $\mathcal{P} = \{E_1, E_2, \dots\}$. This also means $\forall F \in \mathbb{F}$, we can express $F = \bigcup_{j=1}^{\infty} E_{i_j}$.

* **Case 1**: $\mathcal{P}$ is finite. Therefore, $F = \bigcup_{j=1}^{\infty} E_{i_j}$ must be finite, hence a contradiction.
* **Case 2**: $\mathcal{P}$ is countably infinite. $F = \bigcup_{j=1}^{\infty} E_{i_j}$ would then be uncountable, hence a contradiction.

Thus, $\mathcal{F}$ cannot be countably infinite. ■

- **Good Sets Principle**: Suppose we have some property P that we want to verify holds for all sets in σ-field $\mathcal{A}$. When $\mathcal{A}$ doesn't have a good description, what we can do is create the collection of good sets $\mathcal{G}$ such that $\mathcal{G} = \{G \in \mathcal{A} : P \text{ holds for } G\} \subset \mathcal{A}$. Then we show that $\mathcal{G} = \mathcal{A}$, and thus the property must hold for $\mathcal{A}$.

– Is a useful proof technique.

– Suppose we can find a class $\mathcal{C}$ such that:

1. $\mathcal{C}$ is a generator for $\mathcal{A}$. For example, $\mathcal{A} = \sigma\langle\mathcal{C}\rangle$
2. $\mathcal{C} \subset \mathcal{G}$. In other words, P holds for all $C \in \mathcal{C}$.
3. $\mathcal{G}$ is a σ-field.

Then, $\mathcal{A} = \sigma\langle\mathcal{C}\rangle \subset \mathcal{G}$, so $\mathcal{G} = \mathcal{A}$.

– **Example**: Suppose $\mathcal{B}$ is a σ-field of subsets of Ω, and suppose $A \notin \mathcal{B}$. Show that $\sigma\langle\mathcal{B} \cup \{A\}\rangle$ is the smallest σ-field containing both $\mathcal{B}$ and A consists of sets of the form $(A \cap B) \cup (A^c \cap B')$ for $B, B' \in \mathcal{B}$.

Apply the good sets principle. Define $\mathcal{G} = \{A \in \mathcal{A} : (A \cap B) \cup (A^c \cap B') \text{ for } B, B' \in \mathcal{B}\}$. Clearly, $\mathcal{G} \subset \mathcal{A}$, so we now show that $\mathcal{G} = \mathcal{A}$. The next step is showing that $\mathcal{G}$ is in fact a σ-field.

1. $\Omega \in \mathcal{G}$: $\Omega \in \mathcal{A}$ and $\mathcal{B}$ (since $\mathcal{B}$ is a σ-field). Therefore, $(\Omega \cap \Omega) \cup (\varnothing \cap \Omega) = \Omega$, so $\Omega \in \mathcal{G}$ ✓
2. $\varnothing \in \mathcal{G}$: $\varnothing \in \mathcal{A}$ and $\mathcal{B}$ (since $\mathcal{B}$ is a σ-field). Therefore, $(\varnothing \cap \varnothing) \cup (\Omega \cap \varnothing) = \varnothing$, so $\varnothing \in \mathcal{G}$ ✓
3. Closed under complementation. Suppose $A \in \mathcal{A}$.

$$\begin{aligned}((A \cap B) \cup (A^c \cap B'))^c &= (A \cap B)^c \cap (A^c \cap B')^c = (A^c \cup B^c) \cap (A \cup (B')^c) \\ &= ((A^c \cap (B')^c) \cup (A \cap B^c)) \cup B^c \cap (B')^c \\ &= \cdots = (A \cap B^c) \cup (A^c \cap (B')^c).\end{aligned}$$

$\mathcal{B}$ is a σ-field, so it is closed under complementation. ✓

4. Closed under countable union. Once again, since $\mathcal{B}$ is a σ-field, then elements in $\mathcal{B}$ are closed under countable union. Suppose $B_n, B'_n \in \mathcal{B}$ for $n \geq 1$.

$$\bigcup_n (A \cap B_n) \cup (A^c B'_n) = A \cap \left(\bigcup_n B_n \right) \cup A^c \cap \left(\bigcup_n B'_n \right) \in \mathcal{G}. \ \checkmark$$

Thus, $\mathcal{G}$ is a σ-field. $\mathcal{G} \subset \sigma\langle \mathcal{B} \cup \{A\} \rangle$, since the right side contains sets of the form $BA + B'A^c$. Lastly, noting that since A and $\mathcal{B} \subset \mathcal{G}$, $\mathcal{G} \supset \sigma\langle \mathcal{B}, A \rangle$. Thus, $\sigma\langle \mathcal{B} \cup \{A\} \rangle$ is the smallest σ-field containing both $\mathcal{B}$ and A. ■

– **Example**: Suppose $\mathcal{C}$ is a class of subsets of Ω and suppose $B \subset \Omega$ satisfies $B \in \sigma\langle \mathcal{C} \rangle$. Show that there exists a countable class $\mathcal{C}_B \subset \mathcal{C}$ such that $B \in \sigma\langle \mathcal{C}_B \rangle$.

Proceed with the good sets principle. Let $\mathcal{A} = \sigma\langle \mathcal{C} \rangle$, and $\mathcal{G} = \{B \in \mathcal{A} : \exists \text{ countable subclass } \mathcal{C}_B \subset \mathcal{C} : B \in \sigma\langle \mathcal{C}_B \rangle\}$. $\mathcal{C} \subset \mathcal{G}$ is straightforward by construction.
Now, suppose $B \in \mathcal{C}$, and let $\mathcal{C}_B = \{B\} \subset \mathcal{C}$. So, every set from generator $\mathcal{C}$ is a good set. Next, we show $\mathcal{G}$ is a σ-field:

1. $\varnothing, \Omega \in \mathcal{G}$: trivial ✓
2. $\bigcup_{n=1}^{\infty} B_n \in \mathcal{G}$: We need to find a generator for $\bigcup B_n$. Since $\mathcal{C}_{B_n} \subset \mathcal{C}$ is countable, we can define $\mathcal{C}_{\bigcup B_n} = \bigcup \mathcal{C}_{B_n} \subset \mathcal{C}$ ✓
3. $B_n^c \in \mathcal{G}$: $\mathcal{C}_{B^c} = \mathcal{C}_B \subset \mathcal{C}$ ✓

This means $\mathcal{G}$ is a σ-field, so $\mathcal{G} = \mathcal{A} = \sigma\langle \mathcal{C} \rangle$. ■

– **Example**: Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be σ-fields of Ω. Show that the σ-field $\mathcal{B}_1 \vee \mathcal{B}_2$ (which is defined to be the smallest σ-field containing both $\mathcal{B}_1$ and $\mathcal{B}_2$) is generated by sets of the form $B_1 \cap B_2$, where $B_1 \in \mathcal{B}_1$, and $B_2 \in \mathcal{B}_2$.

Define $\mathcal{G} = \{B_1 \cap B_2 : B_1 \in \mathcal{B}_1, B_2 \in \mathcal{B}_2\}$. Next, we show that $\mathcal{G}$ is a σ-field.

1. $\Omega \in \mathcal{G}$: Since $\mathcal{B}_1, \mathcal{B}_2$ are σ-fields, $\Omega \in \mathcal{B}_1$ and $\Omega \in \mathcal{B}_2$. Therefore, $\Omega = \Omega \cap \Omega \in \mathcal{G}$ ✓
2. $\varnothing \in \mathcal{G}$: Since $\mathcal{B}_1, \mathcal{B}_2$ are σ-fields, $\varnothing \in \mathcal{B}_1$ and $\varnothing \in \mathcal{B}_2$. Therefore, $\varnothing = \varnothing \cap \varnothing \in \mathcal{G}$ ✓
3. Closed under countable intersection: Suppose $B_{i1}, B_{i2}, \cdots \in \mathcal{B}_i$. Since $\mathcal{B}_i$ is a σ-field, $\mathcal{B}_i$ is closed under countable intersection. Therefore,

$$\underbrace{\left(\bigcap_{j=1}^{\infty} B_{1j} \right)}_{\in \mathcal{B}_1} \cap \underbrace{\left(\bigcap_{j=1}^{\infty} B_{2j} \right)}_{\in \mathcal{B}_2} \in \mathcal{G} \ \checkmark$$

4. Closed under complementation: Once again, since $\mathcal{B}_i$ is a σ-field, $\mathcal{B}_i$ is closed under complementation. Suppose $B_1 \in \mathcal{B}_1$, $B_2 \in \mathcal{B}_2$. $B_1^c \cap B_2^c = \underbrace{(\Omega \setminus B_1)}_{\in \mathcal{B}_1} \cap \underbrace{(\Omega \setminus B_2)}_{\in \mathcal{B}_2} \in \mathcal{G}$ ✓

Thus, $\mathcal{G}$ is a σ-field, and $\mathcal{G} \supset \sigma\langle \mathcal{B}_1 \vee \mathcal{B}_2 \rangle$. Next, we show that $\mathcal{G} \subset \sigma\langle \mathcal{B}_1 \vee \mathcal{B}_2 \rangle$. This is relatively easier, since we know that $\mathcal{B}_i$ and $\sigma\langle \mathcal{B}_1 \vee \mathcal{B}_2 \rangle$ are σ-fields. Therefore, $B_i \in \mathcal{B}_i \implies B_1 \cap B_2 = (B_1^c \cup B_2^c)^c \in \sigma\langle \mathcal{B}_1 \vee \mathcal{B}_2 \rangle$. So, $\mathcal{G} \subset \sigma\langle \mathcal{B}_1 \vee \mathcal{B}_2 \rangle$, and thus $\mathcal{G} = \sigma\langle \mathcal{B}_1 \vee \mathcal{B}_2 \rangle$. ■

– **Example**: Let P be a probability measure on $\mathcal{B}(\mathbb{R})$. For any $B \in \mathcal{B}(\mathbb{R})$ and $\epsilon > 0$, show that there exists a finite union of intervals A such that $P(A \Delta B) < \epsilon$.

Let $\mathcal{G} = \{B \in \mathcal{B}(\mathbb{R}) : \forall \epsilon > 0, \text{there exists a finite union of intervals } A_\epsilon \text{ such that } P(A \Delta B) > \epsilon\}$. This means that $\mathcal{G} \supset \mathcal{B}$, since we can choose $A_\epsilon = B$.
Next, we show that $\mathcal{G}$ is a σ-field.

1. $\varnothing, \Omega \in \mathcal{B}(\mathbb{R}) \subset \mathcal{G}$ ✓
2. Closed under complementation: Suppose $B \in \mathcal{G}$, and let A_ϵ be the relevant finite union of intervals. For B^c, choose A^c. Since $A^c \Delta B^c = A \Delta B$, $B^c \in \mathcal{G}$ ✓
3. Closed under countable union: Suppose $B_n \in \mathcal{G}$. Given $\epsilon > 0$, we can choose an $A_{\epsilon,n}$ such that $P(A_n \Delta B) < \frac{\epsilon}{2^{n+1}}$. Now, choose $A_\epsilon = \bigcup_{n=1}^{N} A_{\epsilon,n}$.

$$P(B \setminus A_n) \leq P\left(\left(\bigcup_{n=1}^{\infty} B_n \right) \setminus \left(\bigcup_{n=1}^{N} B_n \right) \right) + P\left(\left(\bigcup_{n=1}^{N} B_n \right) \setminus \left(\bigcup_{n=1}^{N} A_n \right) \right) \leq \frac{\epsilon}{2} + \sum_{i=1}^{\infty} P(B_i \setminus A_i).$$

In addition, $P(A\backslash B) \leq P\left(\left(\bigcup_{n=1}^{N} A_n\right) \setminus \left(\bigcup_{n=1}^{N} B_n\right)\right) \leq \sum_{i=1}^{N} P(A_n \backslash B_n)$. Adding these together, we get that

$$P(B\Delta A) \leq \frac{\epsilon}{2} + \sum_{i=1}^{n} P(B_n \Delta A_n) < \frac{\epsilon}{2} + \epsilon \sum_{n=1}^{N} \frac{1}{2^{n+1}} < \epsilon. \checkmark$$

Thus, $\mathcal{G}$ is a σ-field, and hence $\mathcal{G} = \mathcal{B}(\mathbb{R})$, so the assertion holds for all $B \in \mathcal{B}(\mathbb{R})$. ■

– **Example**: Let $(\Omega, \mathcal{A}, P)$ be a probability space, and $\mathcal{F}$ be a field on Ω such that $\sigma\langle\mathcal{F}\rangle = A$. Show that for any $\epsilon > 0$ and $A \in \mathcal{A}$, $\exists F \in \mathcal{F}$ that depends on A and ϵ such that $P(A\Delta F) < \epsilon$.

Let $\mathcal{G} = \{A \in \mathcal{A} : \forall \epsilon > 0, \exists F \in \mathcal{F} : P(A\Delta F) < \epsilon\}$. Choosing $A = F \implies \mathcal{F} \subset \mathcal{G}$. Next, we show $\mathcal{G}$ is a σ-field.

1. $\varnothing, \Omega \in \mathcal{G} \subset \mathcal{G}$ ✓
2. Suppose $A_n \in \mathcal{G}$, which has relevant F_n such that $P(A_n \Delta F_n) < \frac{\epsilon}{2^{n+1}}$. This means that $P\left(\bigcup_{n=1}^{N} A_n\right) = P\left(\bigcup_{n=1}^{\infty} A_n\right)$, so $\exists N : P\left(\left(\bigcup_{n=1}^{N} A_n\right) \setminus \left(\bigcup_{n=1}^{\infty} A_n\right)\right) < \frac{\epsilon}{2}$.
Now, let $F = \bigcup_{n=1}^{N} F_n \in \mathcal{F}$. Then,

$$\begin{aligned}
P\left(\left(\bigcup_{n=1}^{N} A_n\right) \setminus F\right) =& P\left(\left(\bigcup_{n=1}^{\infty} A_n\right) \setminus F\right) + P\left(F \setminus \left(\bigcup_{n=1}^{\infty} A_n\right)\right) \\
\leq& P\left(\left(\bigcup_{n=1}^{\infty} A_n\right) \setminus \left(\bigcup_{n=1}^{N} A_n\right)\right) + P\left(\left(\bigcup_{n=1}^{N} A_n\right) \setminus \left(\bigcup_{n=1}^{N} F_n\right)\right) + P\left(\left(\bigcup_{n=1}^{N} F_n\right) \setminus \left(\bigcup_{n=1}^{N} A_n\right)\right) \\
<& \frac{\epsilon}{2} + P\left(\bigcup_{n=1}^{\infty} (A_n \setminus F_n)\right) + P\left(\bigcup_{n=1}^{\infty} (F_n \setminus A_n)\right) \leq \frac{\epsilon}{2} + \sum_{n=1}^{N} P(A_n \setminus F_n) + \sum_{n=1}^{N} P(F_n \setminus A_n) \\
=& \frac{\epsilon}{2} + \sum_{n=1}^{N} P(A_n \Delta F_n) < \frac{\epsilon}{2} + \sum_{n=1}^{N} \frac{\epsilon}{2^{n+1}} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{G} \checkmark
\end{aligned}$$

3. Suppose $A \in \mathcal{G}$ such that $P(A\Delta F)\epsilon$.

$$A^c \Delta F^c = A\Delta F, \text{ so } P(A^c \Delta F^c) < \epsilon \implies A^c \in \mathcal{G} \checkmark$$

Therefore, $\mathcal{G}$ is a σ-field, so $\mathcal{G} = \sigma\langle\mathcal{G}\rangle = \mathcal{A} \implies \mathcal{G} = \mathcal{A}$. ■

• **Field Generator**: $\mathcal{C}$ is a generator for field $\mathcal{F}$ if $\mathcal{C}$ applies of operations to yield $\mathcal{F}$.

– Generators are not unique, but the fields generated are unique.

– Examples include $\mathcal{C} = \{A, B\}$, which yields $\mathcal{F} = \{\varnothing, A, B, A^c, B^c, A \cap B, A \cup B, A \backslash B, B \backslash A, A\Delta B, (A\Delta B)^c, A^c \cap B^c, A^c \cup B^c, A^c \cup B, A \cup B^c, \Omega\}$, and for a given semifield $\mathcal{C}$, the class of all finite disjoint unions.

– **Countably Generated**: A σ-field such that there exists a countable collection of sets $\mathcal{C}$ such that $\mathcal{A} = \sigma\langle\mathcal{C}\rangle$.

∗ The countable co-countable σ-field is not countably generated.

– **Example**: Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, and let $\mathcal{C} = \{\{2, 4\}, \{6\}\}$. What is the field generated by $\mathcal{C}$, and what is the σ-field?
The field generated by $\mathcal{C}$ is (being careful to omit duplicates)

$$\begin{aligned}
\mathcal{F} =& \{\varnothing, \{2, 4\}, \{2, 4\}^c, \{6\}, \{6\}^c, \{2, 4\} \cup \{6\}, \{2, 4\}^c \cap \{6\}^c, \Omega\} \\
=& \{\varnothing, \{2, 4\}, \{1, 3, 5, 6\}, \{6\}, \{1, 2, 3, 4, 5\}, \{2, 4, 6\}, \{1, 3, 5\}, \Omega\}.
\end{aligned}$$

Since this field contains a finite number of elements, it is also a σ-field. ■

– **Example**: Suppose $\mathcal{C}$ is a class of subsets of $\mathbb{R}$ with the property that $A \in \mathcal{C} \implies A^c$ is a countable union of elements of $\mathcal{C}$. Show that $\sigma\langle\mathcal{C}\rangle$ is the smallest class containing $\mathcal{C}$ which is closed under the formation of countable unions and intersections.

We define $\mathcal{A}$ as a c-class, denoted as $\mathcal{D} \equiv c\langle\mathcal{A}\rangle$, such that:

1. $\Omega, \varnothing \in \mathcal{A}$
2. $\mathcal{A}$ is closed under countable union/intersection.

Note that a c-class that is closed under complements is a σ-field. This means that $c\langle\mathcal{C}\rangle \subset \sigma\langle\mathcal{C}\rangle$.
We now need to show that $\mathcal{D}$ is a σ-field. Proceed with the good sets principle. Define $\mathcal{G} := \{A \in \mathcal{D} : A^c \in \mathcal{D}\}$. We now need to show these properties hold:

1. $\mathcal{C} \subset \mathcal{G}$: Take $A \in \mathcal{C}$, and $c_n \in \mathcal{G}$. $A^c = \bigcup_{n=1}^{\infty} c_n \subset \mathcal{D}$ ✓
2. $\mathcal{G}$ is a c-class. Suppose $A_n \in \mathcal{G} \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{G}$. $\left(\bigcup_{n=1}^{\infty} A_n\right)^c = \bigcap_{n=1}^{\infty} \underbrace{A_n^c}_{\in \mathcal{D}} \in \mathcal{D}$ ✓

Therefore, $\mathcal{G} \supset \sigma\langle\mathcal{C}\rangle = \mathcal{D}$ is a σ-field. Since this is the case, this means that $\mathcal{D}$ is a σ-field. Therefore, $\mathcal{D}$ must be the smallest class containing $\mathcal{C}$ that is closed under countable unions/intersections. ■

- **Induced σ-Fields**: Let Ω, Ω' be sets, with $f : \Omega \to \Omega'$. Also let $\mathcal{A}'$ be a σ-field on Ω'. Suppose $\mathcal{A} = \{f^{-1}(A) : A \in \mathcal{A}'\} \subset \mathcal{P}(\Omega)$. Then, $\mathcal{A}$ is the σ-field induced by the map f from $\mathcal{A}'$ on Ω.
  - $\mathcal{A}$ is a σ-field since inverses commute with set operations.
- **Trace σ-Field**: Let $(\Omega, \mathcal{A})$ be a measurable space with $\Omega' \subset \Omega$. Then, $\mathcal{A}^* = \{A \cap \Omega' : A \in \mathcal{A}\}$ is the trace σ-field on Ω'.
  - If $\Omega' \in \mathcal{A}$, then $\mathcal{A}^* = \{A \in \mathcal{A} : A \subset \Omega'\}$.
  - $A_n \cap \Omega' = \iota^{-1}(A)$.
  - Let Ω be a set with a σ-field $\mathcal{A}$ on it, and let $\mathcal{C}$ be a generator for $\mathcal{A}$. Let $A \subset \Omega$. Then, $\mathcal{C}' = \{C \cap A : C \in \mathcal{C}\}$ is a generator for the trace σ-field $\mathcal{A}' = \{B \cap A : B \in \mathcal{A}\}$ of $\mathcal{A}$ induced by A.
- **Borel σ-Field**, or $\boldsymbol{\mathcal{R}}$ or $\boldsymbol{\mathcal{B}}$: The σ-field generated by the class of all open sets.
  - **Borel Sets**: Elements of $\mathcal{R}$.
  - $\mathcal{R}$ is countably generated by the set of all intervals in $\mathbb{R}$ (doesn't matter if half/purely open/closed, although by definition the intervals are open on the left, closed on the right).
  - $\mathcal{R}^k$ is the Borel σ-field on $\mathbb{R}^k$.
  - **Example**: If $A \subset \mathbb{R}$, then the trace of $\mathcal{R}$ on A is given by $\mathcal{A} = \{B \cap A : B \in \mathcal{R}\}$. However, there is another natural σ-field in the sets that are open in A, given by $\mathcal{T} = \{U \cap A : U \text{ is open in } \mathbb{R}\}$. Consider $\mathcal{A}' = \sigma\langle\mathcal{T}\rangle$. Show that $\mathcal{A}' = \mathcal{A}$.
    First off, every open set must be a Borel set, so $\mathcal{T} \in \mathcal{A} \implies \mathcal{A} \supset \mathcal{A}'$.
    We now need to show that $\mathcal{A} \subset \mathcal{A}'$. Apply the good sets principle. Define $\mathcal{G} = \{B \in \mathcal{R} : A \cap B \in \mathcal{A}'\}$. Next, we show that $\mathcal{G} = \mathcal{R}$. $\mathcal{T}$ is a generator for $\mathcal{B}$, which is the first condition. Now, we show that $\mathcal{G}$ is a σ-field.
    1. $\varnothing, \mathcal{R} \in \mathcal{T} \subset \mathcal{G}$ ✓
    2. $A \cap \left(\bigcup_{n=1}^{\infty} B_n\right) = \bigcup_{n=1}^{\infty} \underbrace{(A \cap B_n)}_{\in \mathcal{A}'} \in \mathcal{A}'$ ✓
    3. Suppose $B \in \mathcal{G}$. $A \cap B^c = \underbrace{A}_{\in \mathcal{A}'} \setminus \underbrace{(A \cap B)}_{\in \mathcal{A}'} \in \mathcal{A}'$ ✓

    Thus, by the good sets principle, $\mathcal{G} = \mathcal{R}$. This means that $\mathcal{A} \subset \mathcal{A}'$, and therefore $\mathcal{A} = \mathcal{A}'$. ■
- **Example**: Let P be a given probability measure on $(\mathbb{R}, \mathcal{R})$. Let $\mathcal{F} = \{(B_1 \times C) \cup (B_2 \times C^c) : B_1, B_2 \in \mathcal{R}\}$, where $C \subset \mathbb{R}$ is a non-empty proper subset of $\mathbb{R}$.
  1. Show that $\mathcal{F}$ is a σ-field on $\mathbb{R} \times \mathbb{R}$.
  2. Let $0 < \alpha < 1$ be fixed. Show that $Q(F) = \alpha P(B_1) + (1-\alpha)P(B_2)$ for $F = (B_1 \times C) \cup (B_2 \times C^c)$ that defines a probability measure on $\mathcal{F}$, and satisfies $Q(B \times \mathbb{R}) = P(B)$ for all $B \in \mathcal{R}$, and $Q(\mathbb{R} \times C) = \alpha$.

  1. (a) $\varnothing = (\varnothing \times C) \cup (\varnothing \times C^c) \in \mathcal{F}$. In addition, $\Omega = \mathbb{R}^2 = (\mathbb{R} \times C) \cup (\mathbb{R} \times C^c) \in \mathcal{F}$ ✓
     (b) Let $A_n = (B_{1n} \times C) \cup (B_{2n} \times C^c)$. Then,

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} (B_{1n} \times C) \cup \bigcup_{n=1}^{\infty} (B_{2n} \times C^c) = \left(\left(\bigcup_{n=1}^{\infty} B_{1n}\right) \times C\right) \cup \left(\left(\bigcup_{n=1}^{\infty} B_{2n}\right) \times C^c\right) \in \mathcal{F} \text{ ✓}$$

     (c) Suppose $A_c \in \mathcal{F}$.

$$A_n^c = (B_1 \times C)^c \cap (B_2 \times C^c)^c = ((B_1^c \cup B_2^c) \times C) \cup ((B_1 \cap B_2) \times C^c) \in \mathcal{F} \text{ ✓}$$

2. Let $F_n = (B_{1n} \times C) \cup (B_{2n} \times C^c)$ for $n = 1, 2, \dots$ be disjoint. Then,

$$\begin{aligned} Q\left(\bigcup_{n=1}^{\infty} F_n\right) =& Q\left[\left(\left(\bigcup_{n=1}^{\infty} B_{1n}\right) \times C\right) \cup \left(\left(\bigcup_{n=1}^{\infty} B_{2n}\right) \times C^c\right)\right] = \alpha P\left(\bigcup_{n=1}^{\infty} B_{1n}\right) + (1-\alpha) P\left(\bigcup_{n=1}^{\infty} B_{2n}\right) \\ =& \alpha P\left(\bigcup_{n=1}^{\infty} B_{1n}\right) + (1-\alpha) P\left(\bigcup_{n=1}^{\infty} B_{2n}\right) = \alpha \sum_{i=1}^{\infty} P(B_{1i}) + (1-\alpha) \sum_{i=1}^{\infty} P(B_{2i}) \\ =& \sum_{i=1}^{\infty} [\alpha P(B_{1i}) + (1-\alpha) P(B_{2i})] = \sum_{i=1}^{\infty} Q(F_n). \blacksquare \end{aligned}$$

9.2 Probability Measures

Return to Table of Contents

- **Measure**: A function $\mu : \mathcal{A} \to [0, \infty)$, where $\mathcal{A}$ is a σ-field on Ω, such that $\mu(\varnothing) = 0$, and countable additivity: if $A_1, A_2, \dots \in \mathcal{A}$ are disjoint, then $\mu(\bigsqcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$.
  - A measure is finite if $\mu(\Omega) < \infty$.
  - A measure is a nonnegative, countably additive set function.
  - **Example**: Let Ω be a non-empty set, and $\mathcal{F}_0$ be the countable co-countable collection. Define for $E \in \mathcal{F}_0$ the set function P by $P(E) = \begin{cases} 0, & E \text{ is finite} \\ 1, & E^c \text{ is finite} \end{cases}$. If Ω is countably infinite, show P is finitely additive but not σ-additive.

    We need to show that P is finitely, but not countably, additive. Consider $E_1 \sqcup \dots \sqcup E_k$. If E_i is cofinite, then this union must be cofinite, and we can convert this to be a finite intersection. Therefore, $P(E) = \sum_{i=1}^{k} P(E_i)$, so P is finitely additive.
    Now, consider $\bigsqcup_{i=1}^{\infty} E_i$, and $\Omega = \mathbb{N}_{E_i = \{i\}}$.

$$1 = P(\Omega) \neq P\left(\bigsqcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} 0 = 0.$$

    This means that P is not countably additive. $\blacksquare$
  - **Example**: Let $\mathcal{A}$ be a field of subsets of Ω, and let μ be a finitely additive probability measure on $\mathcal{A}$ (which means that $\mu(\Omega) = 1$). Show that, if $\mathcal{A}_n \in \mathcal{A}$ and $\mathcal{A}_n \downarrow \varnothing$, it may not necessarily be the case that $\mu(A_n) \downarrow 0$.

    Let $\mathcal{A}$ be the finite-cofinite set, and let $A_n = \{n, n+1, \dots\} \in \mathcal{A}$. Note that $A_n \downarrow \varnothing$, and $A_n^c = \{1, \dots, n-1\}$ is finite. Define the measure $P(A) = 0$ if A is finite, 1 if A^c is finite. Clearly, P is finitely additive. However, A_n^c will always be finite, so although $A_n \downarrow \varnothing$, $P(A_n) = 1 \not\to 0$. $\blacksquare$
- **Probability Space**, or $(\Omega, \mathcal{A}, P)$: A triplet where Ω is the sample space corresponding to outcomes of some (potentially hypothetical) experiment, $\mathcal{A}$ is the σ-field of subsets of Ω, and P is a probability measure.
- **Probability Measure**: $(\Omega, \mathcal{A}, P)$, where $P : \mathcal{A} \to [0, 1]$ for σ-field $\mathcal{A}$ on Ω such that $P(\Omega) = 1$ and countable additivity: if $A_1, A_2, \dots \in \mathcal{A}$ are disjoint, then $P(\bigsqcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.
  - A probability measure is a measure such that $\mu(\Omega) = 1$.
  - $P(\varnothing) = 0$.
    - *Proof*: $\varnothing = \bigcup_{n=1}^{\infty} \varnothing$, so $P(\varnothing) = \sum_{n=1}^{\infty} P(\varnothing)$. Since $P(A) \geq 0$ for any $A \in \Omega$, the only way this equality holds is if $P(\varnothing) = 0$. $\blacksquare$
  - $P(A^c) = 1 - P(A)$.
    - *Proof*: $A \sqcup A^c = \Omega$, so $P(A) + P(A^c) = 1$. $\blacksquare$
  - Monotonicity: if $A \subset B$, then $P(A) \leq P(B)$.
    - *Proof*: $B = A \sqcup (B \setminus A)$. Thus, $P(B) = P(A) + P(B \setminus A) \geq P(A) + 0 = P(A)$. $\blacksquare$
  - Sub-additivity: for any $A_1, A_2, \dots \in \mathcal{A}$, $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$.
    - *Proof*: Disjointification: define $B_1 = A_1$, $B_i = A_i \setminus (\bigcup_{n=1}^{i-1} A_i)$. Then, $P(\bigcup_{i=1}^{n} A_i) = P(\bigsqcup_{i=1}^{n} B_i) = \sum_{i=1}^{n} P(B_i) \leq \sum_{i=1}^{n} P(A_i)$. $\blacksquare$

- Continuity: If $A_n \to A^-$, then $P(A_n) \to P(A)^-$, and if $A_n \to A^+$, then $P(A_n) \to P(A)^+$.
  - *Proof*: Disjointification: define $B_1 = A_1$, $B_i = A_i \setminus (\bigcup_{n=1}^{i-1} A_i)$. Then, $P(A) = \sum_{i=1}^n P(B_i) = \lim_{n\to\infty} P(A_n)$.
    For the other direction, define $B_n := A_1 \setminus A_n$. Thus, $B_n \to (A_1 \setminus A)^+$, so $P(B_n) \to P(A_1 \setminus A)^+ = P(A_1) - P(A)$. ■
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
  - *Proof*: Note that $P(A) = P(A \cap B) + P(A \cap B^c)$, similarly for $P(B)$.

$$\begin{aligned} P(A \cup B) =& P\left((A \cap B^c) \cup (B \cap A^c) \cup (A \cap B)\right) \\ =& P(A \cap B^c) + P(B \cap A^c) + P(A \cap B) \\ =& P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B) \\ =& P(A) + P(B) - P(A \cap B). \blacksquare \end{aligned}$$

- *Inclusion-Exclusion Formula*: $P(\bigcup_{j=1}^n A_j) = \sum_{j=1}^n P(A_j) - \sum_{1\le i<j\le n} P(A_iA_j) + \sum_{1\le i<j<k\le n} P(A_iA_jA_k) - \cdots + (-1)^{n+1} P(A_1 \ldots A_n)$.
  - *Proof*: Induction using the previous result for $n = 2$ as the base case.
  - **Bonferroni Inequality**: Omitting remainder terms.
    - For instance, $P(\bigcup_{j=1}^n A_j) \le \sum_{j=1}^n P(A_j)$, and $P(\bigcup_{j=1}^n A_j) \ge \sum_{j=1}^n P(A_j) - \sum_{1\le i<j\le n} P(A_iA_j)$.
    - **Example**: Events $A_1, A_2, \ldots$ are almost disjoint if $P(A_i \cap A_j) = 0$ for all $i \ne j$. Show that, for such events, $P\left(\bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^\infty P(A_i)$.

      Using the Bonferroni inequality, we get that $P\left(\bigcup_{i=1}^\infty A_i\right) \le \sum_{i=1}^\infty P(A_i)$. In addition,

$$P\left(\bigcup_{i=1}^\infty A_i\right) \ge \sum_{i=1}^\infty P(A_i) - \sum_{1\le i<j\le n} P(A_i \cap A_j) = \sum_{i=1}^\infty P(A_i).$$

      Since $\sum_{i=1}^\infty P(A_i) \le P\left(\bigcup_{i=1}^\infty A_i\right) \le \sum_{i=1}^\infty P(A_i)$, it stands that $P\left(\bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^\infty P(A_i)$. ■
- **Example**: Let $(\Omega, \mathcal{B}, P)$ be a probability space. Show that for events $B_i \subset A_i$, $P(\bigcup_i A_i) - P(\bigcup_i B_i) \le \sum_i P(A_i \cap B_i^c) = \sum_i (P(A_i) - P(B_i))$.

  Since $B_i \subset A_i$, then $\bigcup_i B_i \subset \bigcup_i A_i$.

$$\bigcup_i A_i \setminus \bigcup_i B_i = \left(\bigcup_i A_i\right) \cap \left(\bigcup_i B_i\right)^c = \left(\bigcup_i A_i\right) \cap \left(\bigcap_i B_i^c\right) \subset \bigcup_i (A_i \cap B_i^c);$$

$$P\left(\bigcup_i A_i\right) - P\left(\bigcup_i B_i\right) \le P\left(\bigcup_i (A_i \cap B_i^c)\right) \le \sum_i P(A_i \cap B_i^c) = \sum_i \left(P(A_i) - P(B_i)\right). \blacksquare$$

- If the sample space is $\mathbb{R}$ or $(0, 1]$, then for semi-closed intervals, we can use the CDF function $F : \mathbb{R} \to [0, 1]$ using $P((a, b]) = F(b) - F(a)$.
- If the sample space is more abstract, we start with a specification of probability on a generating semifield $\mathcal{C}$.
  - This requires a countably additive P on $\mathcal{C}$, and to extend it to $\sigma\langle\mathcal{C}\rangle$. Then, if $C_1, C_2, \ldots$ forms a partition of $\mathcal{C}$, then $P(\bigsqcup_{n=1}^\infty C_n) = \sum_{i=1}^\infty P(C_n)$.
  - If P is countably additive on $\mathcal{C}$, then we extend P to the field $\mathcal{F}$ that consists of all finite disjoint unions of elements of $\mathcal{C}$, which then if $C_1, C_2, \ldots, C_n$ forms a partition of $\mathcal{C}$, then $P(F) = \sum_{i=1}^n P(C_i)$.
  - The general method is to start with a sample space Ω and a restricted class of subsets S of Ω to which the assignment of probabilities is easier. Then, we extend this probability assignment to $\sigma\langle S\rangle$.
  - **Example**: Let $\Omega = \mathbb{R}$, and suppose S_1 consists of intervals including $\varnothing$. In other words, $S_1 = \{(a, b] : -\infty \le a \le b \le \infty\}$. If $I_1, I_2 \in S_1$, then $I_1 \cap I_2$ is also an interval, and I_1^c is a union of disjoint intervals. ■
- **Example**: Suppose $\mathcal{B}$ is a σ-field of subsets of Ω and suppose $Q : \mathcal{B} \to [0, 1]$ is a set function such that:
  1. Q is finitely additive on $\mathcal{B}$.

2. $0 \le Q(A) \le 1$ for all $A \in \mathcal{B}$ and $Q(\Omega) = 1$.
3. If $A_i \in \mathcal{B}$ are disjoint and $\sum_{i=1}^{\infty} A_i = \Omega$, then $\sum_{i=1}^{\infty} Q(A_i) = 1$.

Add in A_0 such that $\bigsqcup_{i=0}^{\infty} A_i = \Omega$. $\sum_{i=1}^{\infty} Q(A_0) = 1$. $Q(A_0) + Q\left(\bigsqcup_{i=1}^{\infty} A_n\right) = 1 \implies \sum_{n=1}^{\infty} Q(A_n) = Q\left(\bigsqcup_{n=1}^{\infty} A_n\right)$. ■

– **Example**: If $A_1, \dots, A_n$ are events, Define $S_1 = \sum_{i=1}^{n} P(A_i)$, $S_2 = \sum_{1 \le i < j \le n} P(A_i A_j), \dots$ and so on. Show that the probability $(1 \le m \le n)$ $p(m) = P\left[\sum_{i=1}^{n} \mathbb{1}_{A_i} = m\right]$ of exactly m of the events occurring is $p(m) = S_m - \binom{m+1}{m} S_{m+1} + \binom{m+2}{m} S_{m+2} - \cdots \pm \binom{n}{m} S_n$.

We need to prove that $\mathbb{1}\left\{\sum_{i=1}^{n} \mathbb{1}_{A_i} = m\right\} = \sum_{1 \le i_1 < \cdots < i_m \le n} \mathbb{1}_{\bigcup_{j=1}^{m} A_{i_j}} - \sum_{1 \le i_1 < \cdots < i_{m+1} \le n} \mathbb{1}_{\bigcup_{j=1}^{m+1} A_{i_j}} + \cdots + (-1)^{n-m} \mathbb{1}_{\bigcup_{i=1}^{n} A_i}$.
Case 1: Fewer than m events occur. $\mathbb{1}\left\{\sum_{i=1}^{n} \mathbb{1}_{A_i} = m\right\} = 0$, so $0 = 0 - 0 + \cdots \pm 0 = 0$ ✓
Case 2: $k > m$ events occur.

$$\binom{k}{m} - \binom{k}{m+1} + \cdots \pm \binom{k}{k}$$

This will be shown later TODO ■

- **Example**: Show that if F is a distribution function, then F has at most countably many discontinuities.

Define $F(x-) = \lim_{y \uparrow x} F(y)$. Since F is monotone increasing, $F(x) - F(x-) \ge \frac{1}{n}$ for some $n \in \mathbb{N}$.
Define $D_n = \{x : F(x) - F(x-) \ge \frac{1}{n}\} = P(\{x\})$. Singletons are disjoint, and since we are collecting a union of singletons, $1 \ge P(D_n) \ge \frac{1}{n} \cdot \#D_n \implies \#D_n \le n$, which is finite. Therefore, $\bigcup_{n=1}^{\infty} D_n$ is countable. ■

- **Outer Probability Measure**: A (set) function $P^* : \mathcal{A} \to [0, 1]$ for σ-field $\mathcal{A}$ on Ω such that:

1. $P^*(\varnothing) = 0$ and $P^*(\Omega) = 1$.
2. $A \subset B \implies P^*(A) \le P^*(B)$.
3. Countable subadditivity, where $P^*(\bigcup_{n=1}^{\infty} A_n) \le \sum_{i=1}^{n} P^*(A_n)$.

- ***P**-Measurable**: A set A such that for all $B_1 \subset A$ and $B_2 \subset A^c$, $P^*(B_1 \sqcup B_2) = P^*(B_1) + P^*(B_2)$.

– This is more strict than the probability rule that $P^*(A) + P^*(A^c) = 1$.

- **Caratheodory Extension Theorem**: Let $\mathcal{C}$ be a semifield generating a σ-field $\mathcal{A}$, and let P be countably additive on $\mathcal{C}$. Then,:

1. The class of P^*-measurable sets $\mathcal{A}^*$ forms a σ-field containing $\mathcal{A}$.
2. The restriction of P^* to $\mathcal{A}^*$ is countably additive, so is a probability measure on $\mathcal{A}^* \supset \mathcal{A}$.
3. The extension of P from $\mathcal{C}$ to $\mathcal{A}$ is unique.

– This theorem is very important for this topic.
– $\mathcal{A}^*$ becomes a probabilistic completion of $\mathcal{A}$ with all negligible sets added to the collection.
  * This is where the notion of probability is defined!
– *Proof*: Show that $\mathcal{A}^*$ is a σ-field.
  1. $\varnothing, \Omega \in \mathcal{A}^*$: Since both are in $\mathcal{A}$, both must be in $\mathcal{A}^*$ ✓
  2. Suppose $A_n \in \mathcal{A}^*$: $A_n = B_n \cup N_n \subset D_n \in \mathcal{A}$, where $P(D_n) = 0$, and $N \subset \Omega$ is negligible if $\exists D \supset N$ and $P(D) = 0$.

$$\bigcup_{n=1}^{\infty} A_n = \underbrace{\left(\bigcup_{n=1}^{\infty} B_n\right)}_{\in \mathcal{A}} \cup \underbrace{\left(\bigcup_{n=1}^{\infty} D_n\right)}_{\bigcup N_n \subset \bigcup D_n \in \mathcal{A}} \in \mathcal{A} \in \mathcal{A}^* \checkmark$$

  3. Suppose $A \cup N \in \mathcal{A}^*$. Note that $N \subset D \in \mathcal{A} \implies D^c \subset N^c$.

$$(A \cup N)^c = A^c \cap N^c) = \underbrace{(A^c \cap N^c \cap D)}_{\subset D \therefore \text{ negligible}} \cup \underbrace{(A^c \cap N^c \cap D^c)}_{\in \mathcal{A}} = (A^c \cap D^c) \in \mathcal{A} \checkmark$$

- **Almost Surely**: A property P holds almost surely if $\{\omega : P \text{ does not hold at } \omega\}$ is negligible.

– In other words, the above set is contained within some other set D such that $P(D) = 0$.

- **Lebesgue Measure**, or $\boldsymbol{\lambda}$: The unique extension of the concept of the length of intervals on the real line.
  - $\lambda(\{x\}) = 0$, where x is a single point.
    - This means that $\lambda(C) = 0$ for any countable set C.
  - Is σ-finite, meaning it can be represented as a countable union of sets with finite measure.
  - Is translation invariant.
  - Any translation-invariant measure is a constant multiple of the Lebesgue measure. In other words, let μ be such that $\mu(0,1] = c$ for some $c > 0$. Then, $\mu = c\lambda$.
  - If B has an interior point, $\lambda(B) > 0$.
    - The converse is not necessarily true. For instance, $\mathbb{Q}$ has infinite Lebesgue measures, but since $\mathbb{Q}$ is countable, $\lambda(\mathbb{Q}) = 0$.
  - If $A \in \mathcal{R}$ has $\lambda(A) = 0$, then A^c is dense in $\mathbb{R}$.
  - Uncountable Borel sets C may satisfy $\lambda(C) = 0$.
    - **Example**: Cantor set in $[0,1]$. In a divide-and-conquer fashion, remove the middle third of every remaining set. So, $C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$, and $C_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$. $\lambda(C_n) \to 0$, and C will be uncountable. ■
- **Example**: Suppose A_n are events such that $P(A_n) \to 0$, and $\sum_{i=1}^{\infty} P(A_i \cap A_{i+1}^c) < \infty$. Show that $P(\limsup_{n\to\infty} A_n) = 0$.

  Let $E = \limsup_{n\to\infty} A_n$ and $F = \limsup_{n\to\infty} A_n^c$. $P(E) \le P(E \cap F) + P(F^c)$.

  $$P(F^c) = P(\liminf_{n\to\infty} A_n) = \lim_{n\to\infty} P\left(\bigcap_{m\ge n} A_m\right) \le \liminf_{n\to\infty} P(A_n) = 0.$$

  If $\omega \in E \cap F$, then $\omega \in A_n$ for infinitely many n, and $\omega \in A_n^c$ for infinitely many A_n^c. This means that there are infinitely many n such that $\omega \in A_n$ and $\omega \in A_{n+1}^c$. Thus, $(E \cap F) \subset \limsup_{n\to\infty}(A_n \cap A_{n+1}^c)$.
  However, $\sum_{i=1}^{\infty} P(A_i \cap A_{i+1}^c) < \infty$, so by Borel-Cantelli, $P\left(\limsup_{n\to\infty}(A_n \cap A_{n+1}^c)\right) = 0$. Therefore, $P(E) \le P(E \cap F) + P(F^c) = 0 + 0 = 0$. ■
- **Λ-Class**: A class of sets that is closed under countable disjoint unions.
  - If a Λ-class is closed under countable unions, it is a σ-field.
  - Unlike a σ-field, showing closure under countable disjoint union is not equivalent to showing closure under at most countable intersection.
- **Π-Class**: A class of sets that is closed under finite intersection.
  - Π-classes contains what Λ-classes need to become a σ-field.
- **Bootstrapping**: To prove that a property holds for all pairs (A, B) coming from σ-field $\mathcal{A}$, find a generator $\mathcal{C}$ such that the property holds if $A, B \in \mathcal{C}$ and:
  1. For any $A \in \mathcal{C}$, the good sets $\mathcal{G}_A = \{B \in \mathcal{A} : \text{Property holds for } (A, B)\}$ is a σ-field.
  2. For any $B \in \mathcal{A}$, the good sets $\mathcal{G}_A = \{A \in \mathcal{A} : \text{Property holds for } (A, B)\}$ is a σ-field.

  - This is an extension of the good sets principle.
- **Dynkin's $\Pi - \Lambda$ Theorem**: If $\mathcal{P}$ is a Π-class and $\mathcal{L}$ is a Λ-class containing $\mathcal{P}$ then $\mathcal{L} \supset \sigma\langle\mathcal{P}\rangle$.
  - *Proof*: Bootstrapping. Assume $\mathcal{L}$ is the smallest Λ-class containing $\mathcal{P}$, and proceed to show that $\mathcal{L}$ is also a Λ-class. In other words, $\forall A, B \in \mathcal{L}$, $A \cap B \in \mathcal{L}$.
    Fix $A \in \mathcal{P}$, and define $\mathcal{G}_A = \{B \in \mathcal{L} : A \cap B \in \mathcal{L}\}$. We can show that $\mathcal{G}_A$ is a Λ-class containing $\mathcal{P}$.
    Next, fix $B \in \mathcal{L}$, and define $\mathcal{G}_B = \{A \in \mathcal{L} : A \cap B \in \mathcal{L}\}$. We can also show that $\mathcal{G}_A$ is a Λ-class containing $\mathcal{P}$. ■
  - The Λ-class generated by $\mathcal{P}$ is $\sigma\langle\mathcal{P}\rangle$.
  - If a class is both a Π and Λ-class, then it is a σ-field.
  - If two probability measures P and Q agree on a generating Π-class, then they agree on the whole σ-field.

* On $\mathbb{R}$, if P and Q agree on $(-\infty, x]$ for all x, then P and Q are equal on all Borel sets. This means that a CDF uniquely determines a probability measure.

- **Dynkin's Monotone Class Theorem**: A class $\mathcal{M}$ is a monotone class if $A \in \mathcal{M}$ is closed under increasing or decreasing limit. That is, for $A_n \in \mathcal{M}$, $A_n \uparrow A \implies A \in \mathcal{M}$ (increasing), or for $A_n \in \mathcal{M}$, $A_n \downarrow A \implies A \in \mathcal{M}$ (decreasing).
  - Any Λ-class is a monotone class.
  - If $\mathcal{F}$ is a field and monotone class, then $\mathcal{F}$ is a σ-field.
  - If $\{\mathcal{M}_\alpha : \alpha \in \Lambda\}$ is a collection of monotone classes on Ω, then $\mathcal{M} = \bigcap_{\alpha\in\Lambda} \mathcal{M}_\alpha$ is a monotone class.
  - For any class $\mathcal{C}$,
    $$\mathcal{M} = \bigcap\{\mathcal{M}' : \mathcal{M}' \text{ is a monotone class, } \mathcal{M}' \supset \mathcal{C}\}$$
    is the smallest monotone class containing $\mathcal{C}$, or the monotone class generated by $\mathcal{C}$.
  - If $\mathcal{F}$ is a field and $\mathcal{M}$ is a monotone class containing $\mathcal{F}$, then $\mathcal{M} \supset \sigma\langle\mathcal{F}\rangle$. In particular, the monotone class generated by $\mathcal{F}$ is $\sigma\langle\mathcal{F}\rangle$.
    * *Proof*: Bootstrapping. Define $\mathcal{L} = \{A \in \sigma\langle\mathcal{F}\rangle : A \cap B \in \mathcal{M} \text{ for all } B \in \mathcal{F}\}$. We show that $\mathcal{L}$ is a monotone class, and $\mathcal{F} \subset \mathcal{L}$.
      Show $\mathcal{L}$ is a monotone class.
      1. Nonempty set: $\Omega \cap B = B \in \mathcal{F} \subset \mathcal{M}$, so $\Omega \in \mathcal{L}$. Similar for $\varnothing$ ✓
      2. Closed under increasing limits: Let $\{A_n\} \in \mathcal{L}$ be increasing, and define $A = \bigcup_{n=1}^\infty A_n$. For any $B \in \mathcal{F}$, $A \cap B = \bigcup_{n=1}^\infty (A_n \cap B)$. Since $A_n \cap B \in \mathcal{M}$, and $\mathcal{M}$ is a monotone class, $A \cap B \in \mathcal{M} \implies A \in \mathcal{L}$ ✓
      3. Closed under decreasing limits: Let $\{A_n\}$ be a decreasing sequence in $\mathcal{L}$ and $A = \bigcap_{n=1}^\infty A_n$. For any $B \in \mathcal{F}$, $A \cap B = \bigcap_{n=1}^\infty (A_n \cap B)$. Since $A_n \cap B \in \mathcal{M}$ and $\mathcal{M}$ is closed under decreasing intersections, $A \cap B \in \mathcal{M} \implies A \in \mathcal{L}$.

      Thus, $\mathcal{L}$ is a monotone class. More importantly, $\mathcal{L}$ is a Λ-class.
      Next, we show $\mathcal{F} \subset \mathcal{L}$. Take any $A \in \mathcal{F}$. For any $B \in \mathcal{F}$, $A \cap B \in \mathcal{F}$. Since $\mathcal{F} \subset \mathcal{M}$ by assumption, $A \cap B \in \mathcal{M}$ for every $B \in \mathcal{F} \implies A \in \mathcal{L}$.
      Now, since $\mathcal{F}$ is a field that is closed under finite intersections, it is a Π-class. Since $\mathcal{F} \subset \mathcal{L}$ (a Λ-class), then by Dynkin's $\Pi - \Lambda$ theorem, $\sigma\langle\mathcal{G}\rangle \subset \mathcal{L}$. Thus, for every $A \in \sigma\langle\mathcal{F}\rangle$ and every $B \in \mathcal{F}$, $A \cap B \in \mathcal{M}$. Taking $B = \Omega \implies A \in \mathcal{F} \implies \sigma\langle\mathcal{F}\rangle \subset \mathcal{M}$. ■

- **Example**: Let Ω be a sample space with σ-field $\mathcal{A}$, and P and Q are probability measures on $(\Omega, \mathcal{A})$.
  1. Show that $\mathcal{D} = \{A \in \mathcal{A} : P(A) = Q(A)\}$ is a Λ-class.
  2. If P and Q on the real line satisfy $P((-\infty, x]) = P((-\infty, y])$ for all $x \in \mathbb{Q}$, show that $P(A) = Q(A)$ for all $A \in \mathcal{R}$.

  1. Show that the definition of a Λ-class is met.
     (a) $P(\varnothing) = 0 = Q(\varnothing)$, and $P(\Omega) = 1 = Q(\Omega)$, so $\Omega, \varnothing \in \mathcal{D}$ ✓
     (b) Closure under disjoint union: suppose $A_n \in \mathcal{D}$ are disjoint, and $P(A_n) = Q(A_n)$ for all n.
     $$P\left(\bigsqcup_n A_n\right) = \sum_n P(A_n) = \sum_n Q(A_n) = Q\left(\bigsqcup_n A_n\right) \checkmark$$
     (c) Closure under complementation: Suppose $A \in \mathcal{D}$. $P(A^c) = 1 - P(A) = 1 - Q(A) = Q(A^c)$ ✓
     Therefore, $\mathcal{D}$ is a Λ-class.
  2. Let $\mathcal{C} = \{(-\infty, x] : x \in \mathbb{Q}\}$ be the Π-class generator for $\mathcal{R}$. $\forall A \in \mathcal{C}$, $P(A) = Q(A)$, since $\mathcal{G} = \{A \in \mathcal{R} : P(A) = Q(A)\} \supset \mathcal{C}$.
     Since $\mathcal{G}$ is a Λ-class, $\mathcal{G} \supset \sigma\langle\mathcal{C}\rangle = \mathcal{R}$ by Dynkin's $\Pi - \Lambda$ theorem. This means that $\forall A \in \mathcal{R}$, $P(A) = Q(A)$.
     ■

9.3 Random Variables and Independence

Return to Table of Contents

- **Random Variable**: A function $X : \Omega \to \mathbb{R}$.
  - We also indicate $(\Omega, \mathcal{A}, P)$, where Ω is the sample space, $\mathcal{A}$ is a σ-field, and P is a probability measure.
  - We allow X to map to $\pm\infty$ for more flexibility.
  - $X^{-1}(B) = \{\omega : X(\omega) \in B\}$ needs to be an event. In other words, $X^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{R}$. This means we need to have a restriction that $P(X \in B) \in \mathcal{A}$.
  - **Example**: Let $(\Omega, \mathcal{B}, P)$ be $([0,1], \mathcal{B}([0,1]), \lambda)$, where λ is the Lebesgue measure on $[0,1]$. Define $\{X_t, 0 \le t \le 1\}$ by $X_t(\omega) = \mathbb{I}(t = \omega)$. Show that each X_t is a random variable. What is the σ-field generated by $\{X_t, 0 \le t \le 1\}$?

    We need to show that $X^{-1} \in \mathcal{R}$. $X_t^{-1}(\{1\}) = \{\omega : X_t(\omega) = 1\} = \{t\} \in \mathcal{R}$. The only other outcome is $X_t^{-1}(\{0\}) = X_t^{-1}(\{1\})^c \in \mathcal{R}$. Since every $B \in \Omega' = \{0,1\}$ satisfies $X_t^{-1}(B) \in \mathcal{R}$, X is a random variable.
    The smallest σ-field that contains all of the singletons is the countable-cocountable σ-field. ■
  - **Example**: Suppose $X : \Omega \to \mathbb{R}$ has a countable range $\mathcal{R}$. Show $X \in \mathcal{B}/\mathcal{B}(\mathbb{R})$ (that is, $X : (\Omega, \mathcal{B}) \to (\Omega', \mathcal{B}(\mathbb{R}))$ iff $X^{-1}(\{x\}) \in \mathcal{B}$ for all $x \in \mathcal{R}$.

    First, assume $X \in \mathcal{B}/\mathcal{B}(\mathbb{R})$. Thus, since $\{x\} \in \mathcal{B}(\mathbb{R})$, $X^{-1}(\{x\}) \in \mathcal{B}$ for all $x \in \mathcal{B}(\mathbb{R})$.
    Now, suppose $X^{-1}(\{x\}) \in \mathcal{B}$ for all $x \in \mathcal{R}$. Thus, using the definition of an inverse map, for any $B \in \mathcal{B}(\mathbb{R})$,
    $$\begin{aligned} X^{-1}(B) =& \{\omega : X(\omega) \in B\} \overset{B \subseteq \mathcal{R}}{=} \{\omega : X(\omega) \in B \cap \mathcal{R}\} \\ =& \bigcup_{r \in \mathcal{R}, r \in B} \{\omega : X(\omega) = r\} = \bigcup_{r \in \mathcal{R}, r \in B} X^{-1}(\{r\}) \in \mathcal{B}. \end{aligned}$$
    Thus, $X \in \mathcal{B}/\mathcal{B}(\mathbb{R})$. ■
  - **Example**: If X is a random variable satisfying $P(|X| < \infty) = 1$, then show that for any $\epsilon > 0$, there exists a bounded RV Y such that $P(X \neq Y) < \epsilon$.

    If $P(|X| < \infty) = 1$, then $\lim_{M \to \infty} P(|X| > M) \to 0$. Denote $\epsilon > 0$ such that $P(|X| > M) < \epsilon$.
    Define $Y(\omega) = |X(\omega)| \cdot \mathbb{1}(|X(\omega) \le M)$. Clearly, $|Y|$ is bounded by M for all ω. In addition, $P(X \neq Y) = P(|X| > M) < \epsilon$ by construction. Therefore, we have found a bounded random variable Y that satisfies this property. ■
  - **Example**: Show that if X is an RV, then $\sigma\langle X \rangle$ is countably generated. Conversely, show that if $\mathcal{B}$ is any countably generated σ-field, show that $\mathcal{B} = \sigma\langle X \rangle$ for some RV X.

    Since X is a random variable, then $\sigma\langle X \rangle = \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$.
    Since the Borel σ-field is countably generated, we can express a generator for $\mathcal{B}(\mathbb{R})$ as $\{B_1, B_2, \dots\}$. Consider $\mathcal{A} = \{X^{-1}(B_i) : i \in \mathbb{N}\}$. Since the inverse map preserves countable operations, $\sigma\langle \mathcal{A} \rangle$ contains all inverse elements of $\mathcal{B}(\mathbb{R})$. Thus, $\sigma\langle \mathcal{A} \rangle = \sigma\langle X \rangle$. Finally, since $\sigma\langle \mathcal{A} \rangle$ is trivially countably generated, so must $\sigma\langle X \rangle$.
    Now, we show the converse. Since $\mathcal{B}$ is countably generated, we can represent the generating set as $\mathcal{A} = \{B_1, B_2, \dots\} \subset \mathcal{B}$.
    Apply the good sets principle. We already know that $\mathcal{B}$ is a σ-field, so we don't need to show it here. Express $X(\omega) = \sum_{i=1}^{\infty} \frac{\mathbb{1}_{B_n}(\omega)}{2^n}$, where the ith digit means ω exists in B_i. This maps $\omega \to [0,1]$, since we have the sum of indicators that can be at most one. Clearly, $B_i \in \mathcal{B}$, and since X is a pointwise limit, $\sigma\langle X \rangle \subset \mathcal{B}$.
    Now, define $A_n = \{x \in [0,1] :$ the nth binary digit of x is $1\}$. Since A_n is a countable intersection of intervals, it must be Borel-measurable. Thus, $X^{-1}(A_n) = \{\omega \in \Omega : \mathbb{1}_{B_n}(\omega) = 1\} = B_n$, which means that $B_n \in \sigma\langle X \rangle$ for all n.
    Since $\mathcal{A}$ generates $\mathcal{B}$, $\mathcal{B} = \sigma\langle \mathcal{A} \rangle \subset \sigma\langle X \rangle$. Combining with above, we get that $\mathcal{B} = \sigma\langle \mathcal{A} \rangle = \sigma\langle X \rangle$. ■
  - **Example**: Suppose $-\infty < a \le b < \infty$. Show that the indicator function $\mathbb{1}_{(a,b]}(x)$ can be approximated by bounded and continuous functions. That is, show that there exists a sequence of continuous functions $0 \le f_n \le 1$ such that $f_n \to \mathbb{1}_{(a,b]}$ pointwise.

    Define $f(x) = \mathbb{1}_{(a,b]}$ and $f_n(x) = \begin{cases} 1 & , a + \frac{1}{n} < x \le b \\ 0 & , x \le a \text{ or } x \ge b + \frac{1}{n} \end{cases}$, which are all positive, bounded and

continuous.
Next, we show that we can approximate $f(x)$ with $f_n(x)$. If $x \leq a$ or $x > b$, then $f_n(x) = 0 = f(x)$. If $x \in (a, b]$, then $f(x) = 1$. For large n such that $\frac{1}{n} \approx 0$, if $x \in (a, b]$, then $f_n(x) = 1$. Similarly, for large n, if $x > b$, then $f_n(x) = 0$. Therefore, we can approximate $f(x)$ with $f_n(x)$. ■

- **Measurable Function**: A function $f : \Omega \to \Omega'$ such that $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ are measurable spaces where $f^{-1}(A') \in \mathcal{A}$ for all $A' \in \mathcal{A}'$.
  - $f(A) \in \mathcal{A}'$ doesn't necessarily need to be true.
  - An RV is a measurable real-valued function on a sample space.
  - f is measurable iff $f^{-1}(A') \in \mathcal{A}$ for all $A' \in \mathcal{C}'$, where $\mathcal{C}'$ is any generator for $\mathcal{A}'$.
    - *Proof*: Show $f^{-1}(\mathcal{A}') \subset \mathcal{A}$. Apply the good sets principle. Let $\mathcal{G} = \{A' \in \mathcal{A}' : f^{-1}(A') \in \mathcal{A}\}$. $\mathcal{C}' \subset \mathcal{G}$ by the given condition. In addition, by the commutativity of inverse functions with set operations, a routine verification gives that $\mathcal{G}$ is a σ-field. Thus, $\mathcal{G} \supset \sigma\langle\mathcal{C}'\rangle = \mathcal{A}'$. ■
  - **Example**: Suppose that $\{B_n, n \geq 1\}$ is a countable partition of Ω and define $\mathcal{B} := \sigma\langle B_n, n \geq 1\rangle$. Show that a function $X : \Omega \to (-\infty, \infty]$ is $\mathcal{B}$-measurable iff X is of the form $X = \sum_{i=1}^{\infty} c_i \mathbb{1}_{B_i}$.

    First, suppose X is $\mathcal{B}$-measurable. This means that $X^{-1}(C) \in \mathcal{B}$, which can be represented as a countable union $\bigcup_{n\geq 1} B_n$. Since X is a step function represented by indicators, X is constant on each B_n. This means that $X(\omega) = c_n$, and hence $X = \sum_{n=1}^{\infty} c_n \mathbb{1}_{B_n}$.
    Now, suppose X is of the form $X = \sum_{i=1}^{\infty} c_i \mathbb{1}_{B_i}$. For any Borel set $A \subset \mathbb{R}$, $X^{-1}(A) = \bigcup\{B_n : c_n \in A\}$. Since B_n is countable and disjoint, we can express $X^{-1}(A)$ as a countable union of disjoint elements. Therefore, X is $\mathcal{B}$-measurable. ■
  - **Example**: A real function f on the real line is upper semi-continuous (usc) at x if for every ϵ, $\exists\delta$ such that $|x - y| < \delta \implies f(y) < f(x) + \epsilon$. Check that if f is everywhere usc, then it is measurable.

    The set $\mathcal{A} = \{x : f(x) < t\}$ corresponds to the interval $(-\infty, t)$, which means that if f is usc, then $f^{-1}((-\infty, t))$ is open in $\mathbb{R}$.
    This means that $\sigma\langle\mathcal{A}\rangle = \mathcal{B}(\mathbb{R})$, so $f^{-1}(\mathcal{A}) \subset \mathcal{B}(\mathbb{R})$. Therefore, $f \in \mathcal{B}(\mathbb{R})/\mathcal{B}(\mathbb{R})$, which means that if f is usc everywhere, then it must be measurable. ■
  - **Example**: Let λ be the Lebesgue measure on the interval $[0, 1]$. Let A be a Borel subset of $[0, 1]$, and $F(x) = \lambda(A \cap [0, x])$ for $x \in [0, 1]$.
    1. Show that $|F(x) - F(y)| \leq |x - y|$ for all $x, y \in [0, 1]$.
    2. Show that if $\lambda(A) > 0$, then for a given $0 < t < \lambda(A)$, there exists a Borel set $B \subset A$ such that $\lambda(B) = t$.

    1. WLOG, suppose $y \geq x \implies F(y) \geq F(x)$.

    $$F(y) - F(x) = \lambda(A \cap [0, y]) - \lambda(A \cap [0, x]) \leq \lambda(A \cap (x, y]) \leq \lambda((x, y]) = y - x.$$

    2. Note that F is continuous, with $F(0) = 0$, and $F(1) = \lambda(A)$. Therefore, $\exists x^* : F(x^*) = t$. Then, for $B = A \cap [0, x^*]$, $B \subset A$, and $\lambda(B) = t$. ■
- **Borel Measurable**: f such that $(\Omega, \mathcal{A}) = (\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{R})$ and $f^{-1}(B) \in \mathcal{R}$ for every $B \in \mathcal{R}$.
  - **Example**: If $B \in \mathcal{R}$, show that $\mathbb{1}_B : \mathbb{R} \to \mathbb{R}$ is Borel measurable.

    $\mathbb{1}_B^{-1}(\{0\}) = B^c \in \mathcal{R}$, $\mathbb{1}_B^{-1}(\{1\}) = B \in \mathcal{R}$, $\mathbb{1}_B^{-1}(\{0, 1\}) = \mathbb{R} \in \mathcal{R}$, $\mathbb{1}_B^{-1}(\{\varnothing\}) = \varnothing \in \mathcal{R}$. ■
- **Extended Real Line**, or $\bar{\mathbb{R}}$: $\{-\infty\} \cup \mathbb{R} \cup \{\infty\}$.
  - $\bar{\mathbb{R}}$ is compact.
  - If $x \to \infty$, then x converges in $\bar{\mathbb{R}}$.
- **Extended Borel σ-Field**, or $\bar{\mathcal{R}}$: $\sigma\langle\{-\infty\}, \mathcal{R}, \{\infty\}\rangle = \{B, B \cup \{-\infty\}, B \cup \{\infty\}, B \cup \{-\infty, \infty\} : B \in \mathcal{R}\}$.
- If $f : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ and $g : (\Omega', \mathcal{A}') \to (\Omega'', \mathcal{A}'')$ are measurable, then $h := g \circ f : (\Omega, \mathcal{A}) \to (\Omega'', \mathcal{A}'')$ is also measurable.
  - *Proof*: Take $A'' \in \mathcal{A}''$. Then, $h^{-1}(A'') = f^{-1}(g^{-1}(A'')) \in \mathcal{A}$. ■
- Let f, g, f_n be RVs, and $a \in \mathbb{R}$, then the following expressions are measurable functions:

- $a \cdot f$.
  - *Proof*: $\{x : a \cdot f(x) \in B\} = f^{-1}(B')$, where $B' = \{y : a \cdot y \in B\} \in \bar{\mathcal{R}}$. ■
- $f + g$ (if well defined).
  - *Proof*: Let $h : \Omega \to \bar{\mathbb{R}}^2$ be defined as $h(\omega) = (f(\omega), g(\omega))$, and $\mathcal{C} = \{(-\infty, x] \times (-\infty, y]\}$. $\{\omega : f(\omega) \in (-\infty, x]\} \cap \{\omega : g(\omega) \in (-\infty, y]\} \in \mathcal{A}$, so h is measurable.
    Now, consider $\pi : \bar{\mathbb{R}}^2 \to \bar{\mathbb{R}}$ defined as $\pi(x, y) = x + y$. π is continuous (if well defined), so π must be Borel measurable.
    Since π and h are measurable, then by the previous theorem, $\pi \circ h = f(\omega) + g(\omega)$ is measurable. ■
- fg.
  - *Proof*: Same as above, but using $\pi(x, y) = x \cdot y$. ■
- f/g (if well defined).
  - *Proof*: Same as above, but using $\pi(x, y) = \frac{x}{y}$. ■
- $\sup_n f_n$, $\inf_n f_n$.
  - *Proof*:

$$\{\omega : \sup_n f_n(\omega) \in (-\infty, x]\} = \{\omega : f_n(\omega) \leq x \ \forall n\} = \bigcap_{n=1}^{\infty} \underbrace{\{\omega : f_n(\omega) \in (-\infty, x]\}}_{\in \bar{\mathcal{R}}} \in \bar{\mathcal{R}};$$

$$\{\omega : \inf_n f_n(\omega) \in (-\infty, x]\} = \{\omega : f_n(\omega) \geq x \ \forall n\} = \bigcap_{n=1}^{\infty} \underbrace{\{\omega : f_n(\omega) \in [x, \infty)\}}_{\in \bar{\mathcal{R}}} \in \bar{\mathcal{R}}. \ ■$$

- $\limsup_{n\to\infty} f_n$, $\liminf_{n\to\infty} f_n$, $\lim_{n\to\infty} f_n$ (if exists).
  - *Proof*: $\limsup_{n\to\infty} f_n = \inf_{n\geq 1}\{\sup_{m\geq n} f_m\}$, which is measurable by the previous statement. Similarly, $\liminf_{n\to\infty} f_n = \sup_{n\geq 1}\{\inf_{m\geq n} f_m\}$, which is also measurable as before.
    If the limit exists, then $\lim_{n\to\infty} f_n = \limsup_{n\to\infty} f_n$, which was just proven to be measurable. ■

- **Simple Function**: A measurable function that takes only finitely many different values. That is, $f(\omega) = \sum_{i=1}^{n} c_i \mathbb{1}_{A_i}(\omega)$ for $n \in \mathbb{N}$, $A_1, \ldots, A_n \in \mathcal{A}$, and $c_1, \ldots, c_n \in \mathbb{R}$.
- **Simple Random Variable**: A random variable that takes only finitely many different values. That is, $f(\omega) = \sum_{i=1}^{n} c_i \mathbb{1}_{A_i}(\omega)$ for $n \in \mathbb{N}$, $A_1, \ldots, A_n \in \mathcal{A}$, and $c_1, \ldots, c_n \in \mathbb{R}$.
  - We can approximate random variables with simple RVs. For measurable function f and $n \in \mathbb{N}$, define

$$f_n(\omega) = \begin{cases} k2^{-n} & , k2^{-n} \leq f(\omega) < (k+1)2^{-n} \text{ and } k = -2^{2n}, \ldots, 2^{2n} \\ -2^n & , f(\omega) < -2^n \\ 2^n & , f(\omega) \geq 2^n \end{cases}.$$

    - $f_n(\omega) \leq f(\omega)$ unless $f(\omega) < -2^n$.
    - $|f_n(\omega) - f(\omega)| < 2^{-n}$ unless $|f(\omega)| > 2^n$.
    - $\forall \omega, f_n(\omega) \to f(\omega)$.
    - If $f \geq 0$, then $0 \leq f_n \leq f$, and $f_n(\omega) \uparrow f(\omega)$ for all ω.
    - If $|f|$ is bounded above, then for large n, $\sup_\omega |f_n(\omega) - f(\omega)| \leq 2^{-n} \to 0$.
- **Induced σ-Field**: $\sigma\langle f \rangle = \{f^{-1}(B) : B \in \mathcal{R}\}$ for measurable f on $(\Omega, \mathcal{A})$.
  - **Example**: Suppose $\Omega = [0, 1]$, and $\mathcal{A}$ is the σ-field of Borel subsets of $[0, 1]$. The Lebesgue measure on $[0, 1]$ is an induced σ-field. This is actually the uniform distribution! ■
- **Example**: Let $\{X_n\} \overset{\text{iid}}{\sim} \text{Exp}(1)$: that is, $P(X_n \leq x) = 1 - e^{-x}$ for all n. Show that $\liminf_{n\to\infty} nX_n = 0$ almost surely.

  Let $\{A_n\} = \{nX_n < \epsilon\}$, which are independent events.

$$P(A_n) = P(nX_n < \epsilon) = P\left(X_n < \frac{\epsilon}{n}\right) = 1 - e^{-\epsilon/n}.$$

  It can be shown that $\lim_{x\to 0} \frac{1-e^{-x}}{x} = 1$, so $1 - e^{-\epsilon/n} \approx \frac{\epsilon}{n}$. Therefore, $\sum_{i=1}^{\infty} P(A_i) = \epsilon \sum_{i=1}^{\infty} \frac{1}{n} = \infty$. By the Borel-Cantelli lemma, $P(\limsup_{n\to\infty} A_n) = 1$ a.s.
  This means that there are infinitely many n such that $nX_n < \epsilon$. This means that $\liminf_{n\to\infty} nX_n \leq \epsilon$ a.s. Since $\epsilon > 0$ was arbitrary, $\liminf_{n\to\infty} nX_n = 0$ a.s. ■

- **Independence (Events)**: Let $(\Omega, \mathcal{A}, P)$ be a probability space. Two events $A, B \in \mathcal{A}$ are independent if $P(A \cap B) = P(A)P(B)$.
  - If A and B are independent, then so are (A, B^c), (A^c, B), and (A^c, B^c).
  - $\varnothing$ and Ω are independent of any event.
  - If $P(A) \in \{0, 1\}$ iff A is independent of any event, including itself.
  - Let $A \in \mathcal{A}$ and $\mathcal{D} = \{B \in \mathcal{A} : P(A \cap B) = P(A)P(B)\}$. Then, $\mathcal{D}$ is a Λ-class.
    - *Proof*: Need to show $\mathcal{D}$ is a Λ-class.
      1. $\varnothing, \Omega \in \mathcal{D}$: Straightforward, since $\varnothing$ and Ω are independent of any event ✓
      2. $\mathcal{D}$ is closed under proper difference: Suppose $B_1, B_2 \in \mathcal{D}$ such that $B_1 \subset B_2$.

         $$P(A \cap (B_2 \setminus B_1)) = P(A \cap B_2) - P(A \cap B_1) = P(A)[P(B_2) - P(B_1)] = P(A)P(B_2 \setminus B_1) \in \mathcal{D}$$ ✓

      3. $\mathcal{D}$ is closed under countable disjoint unions: Let $B_n \in \mathcal{D}$. Then,

         $$P(A \cap (\sqcup_n B_n)) = P(\sqcup_n (A \cap B_n)) = \sum_n P(A)P(B_n) = P(A) \sum_n P(B_n) = P(A)P(\sqcup_n B_n) \in \mathcal{D}$$ ✓

      Therefore, by definition, $\mathcal{D}$ is a Λ-class ■
    - If $\mathcal{D}$ contains a Π-class $\mathcal{C}$, then $\mathcal{D}$ contains $\sigma\langle\mathcal{C}\rangle$.
  - **Example**: Let A, V, C be disjoint events in a probability space with $P(A) = 0.6$, $P(B) = 0.3$, and $P(C) = 0.1$. Calculate the probabilities of every event in $\sigma\langle A, B, C\rangle$.

    First off, $P(\Omega) = 1$, $P(\varnothing) = 0$, $P(A) = 0.6$, $P(B) = 0.3$, and $P(C) = 0.1$. Since A, B, and C are all disjoint,

    $$P(A \cup B \cup C) = P(\Omega) = 1, \text{ and } P(A \cap B) = P(B \cap C) = P(A \cap C) = P(A \cap B \cap C) = P(\varnothing) = 0.$$

    Similarly, $P(A \cup B) = P(C^c) = 0.6 + 0.3 = 0.9$, $P(B \cup C) = P(A^c) = 0.3 + 0.1 = 0.4$, and $P(A \cup C) = P(B^c) = 0.6 + 0.1 = 0.7$. Lastly, $P(A \cap B^c) = P(A) = 0.6$, and $P(A \cup B^c) = P(B^c) = 0.7$ due to disjointedness. Similar results hold for $P(A \cap C^c)$, $P(B \cup C^c)$, etc. ■
  - **Example**: Let $(\Omega, \mathcal{A}, P)$ be a probability space. If $A \in \mathcal{A}$ is independent of any Π-class $\mathcal{P} \subset \mathcal{A}$ such that $A = \sigma\langle\mathcal{P}\rangle$, then show that $P(A) \in \{0, 1\}$.

    Apply the good sets principle. Let $\mathcal{G} = \{B \in \mathcal{A} : P(A \cap B) = P(A)P(B)\} \supset \mathcal{P}$.
    Next, we show that $\mathcal{G}$ is a Λ-class.
    1. $\Omega, \varnothing \in \mathcal{G}$: $\Omega, \varnothing \in \mathcal{P}$, since $\mathcal{P}$ is a Λ-class ✓
    2. Closed under countable disjoint union: Suppose $B_n \in \mathcal{G}$.

       $$A \cap \left(\bigcup_{n=1}^{\infty} B_n \right) = \bigcup_{n=1}^{\infty} \underbrace{A \cap B_n}_{\in \mathcal{G}} \in \mathcal{G}$$ ✓

    3. Closed under complementation: Suppose $B \in \mathcal{G}$. $A \cap B^c = A \cap (\Omega \setminus B) \in \mathcal{G}$ ✓

    Thus, $\mathcal{G}$ is a Λ-class. Therefore, by Dynkin's theorem, $\mathcal{G} \supset \sigma\langle\mathcal{P}\rangle$. Since $A \in \sigma\langle\mathcal{P}\rangle$, A is independent of itself. Therefore, $P(A) \in \{0, 1\}$. ■

- **Independence (Classes)**: Let $(\Omega, \mathcal{A}, P)$ be a probability space. Two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ are independent if any $C_1 \in \mathcal{C}_1$ and $C_2 \in \mathcal{C}_2$ are independent.
  - If $\mathcal{C}_1$ and $\mathcal{C}_2$ are independent, then the corresponding Λ-classes generated by them $\mathcal{D}_1 = D\langle\mathcal{C}_1\rangle$ and $\mathcal{D}_2 = D\langle\mathcal{C}_2\rangle$ are independent. Furthermore, if $\mathcal{C}_1$ and $\mathcal{C}_2$ are independent Π-classes, then $\sigma\langle\mathcal{C}_1\rangle$ and $\sigma\langle\mathcal{C}_2\rangle$ are independent.
    - *Proof*: Bootstrapping. Let $\mathcal{G}_1$ denote the class of sets independent of $\mathcal{C}_2$. Then, by assumption, $\mathcal{C}_1 \subset \mathcal{D}_1 \subset \mathcal{G}_1$. Thus, $\mathcal{D}_1$ is independent of $\mathcal{C}_2$. Now, let $\mathcal{G}_2$ denote the class of sets independent of $\mathcal{C}_1$. Then, by assumption, $\mathcal{C}_2 \subset \mathcal{D}_2 \subset \mathcal{G}_2$. Thus, $\mathcal{D}_2$ is independent of $\mathcal{C}_1$. Therefore, $\mathcal{D}_1$ is independent of $\mathcal{D}_2$. ■

- **Mutually Independent**: Arbitrary, non-empty collections of events $\mathcal{C}_1, \ldots, \mathcal{C}_n$ are mutually independent if $P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P\left(A_{i_j}\right)$ for every $k \leq n$ and $1 \leq i_1 < \cdots < i_k \leq n$, where $A_{i_j} \in \mathcal{C}_{i_j}$.

- The collection of classes $\{\mathcal{C}_t : t \in T\}$ is mutually independent if every finite subcollection is mutually independent.
- If $\{\mathcal{C}_t : t \in T\}$ are mutually independent Λ-classes, then $\{\sigma\langle\mathcal{C}_t\rangle : t \in T\}$ are independent.
  - *Proof*: WLOG, assume T is finite. For $T = \{1, 2\}$, $D\langle\mathcal{C}_1\rangle = \sigma\langle\mathcal{C}_1\rangle$ and $D\langle\mathcal{C}_2\rangle = \sigma\langle\mathcal{C}_2\rangle$ are independent, as shown earlier. The general case is followed by repeated bootstrap, where we keep the $n-1$ fixed, and we look at $D\langle\mathcal{C}_{n-1}\rangle = \sigma\langle\mathcal{C}_{n-1}\rangle$ and $D\langle\mathcal{C}_n\rangle = \sigma\langle\mathcal{C}_n\rangle$. ■
- Consider independent sub-σ-fields $\{\mathcal{A}_t : t \in T\}$. If $T_j \subset T$ are pairwise disjoint for $j \in J$, and $\mathcal{A}_{T_j} = \sigma\langle\mathcal{A}_t : t \in T_j\rangle$, then $\{\mathcal{A}_{T_j} : j \in J\}$ are mutually independent. In other words, different portions of disjoint T_j are independent.

- **Mutually Independent (Random Variables)**: Let $X_t, t \in T$ be a collection of RVs on a probability space $(\Omega, \mathcal{A}, P)$, and put $\mathcal{A}_t = \sigma\langle X_t\rangle, \sigma\langle X_t, t \in T\rangle = \sigma\langle\mathcal{A}_t, t \in T\rangle$. X_t are mutually independent if $\mathcal{A}_t$ are independent σ-fields. In other words, X_t are mutually independent if for every collection $\{t_1, \ldots, t_n\} \subset T$,

$$P(X_{t_1} \in B_1, \ldots, X_{t_n} \in B_n) = \prod_{i=1}^{n} P(X_{t_i} \in B_i)$$

  for all choice of Borel sets $B_1, \ldots, B_n \in \mathcal{R}$.
  - To verify independence, it is enough to verify the equality holds for all sets from a generating Λ-class $\mathcal{C}$ of $\mathcal{R}$, as opposed to all Borel sets.
  - Given $\mathcal{C} = \{(-\infty, x] : x \in \mathbb{R}\}$, independence is equivalent to showing that $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 \le x_1, \ldots, X_n \le x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$.
    - For discrete RVs, this is equivalent to $P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i)$.
  - Let $X_t, t \in T$ be independent, and $T_j \subset T$ are pairwise disjoint for $j \in J$, then $\{\psi_j(X_t : t \in T) : j \in J\}$ are independent for arbitrary Borel measurable functions ψ_j, $j \in J$.
    - *Proof*: Define $\mathcal{A}_t = \sigma\langle X_t\rangle$. Then, apply the independence of disjoint groupings for σ-fields. ■
  - We can construct independent RVs given probability space $(\Omega = [0, 1], \mathcal{A} = \mathcal{R}_{[0,1]}, P = \lambda)$ by using the binary expansion of ω. Define $X_i(\omega) = x_i$ for $i \ge 1$, which are iid RVs taking values 0 and 1.
    - If $\omega \ne \frac{m}{2^k}$, this expansion is unique. Otherwise, we choose the terminating expansion.

- **Borel-Cantelli Lemmas**: Suppose A_n are arbitrary events. If $\sum_{i=1}^{\infty} P(A_i) < \infty$, then $P(\limsup_{n\to\infty} A_n) = 0$. If $\sum_{i=1}^{\infty} P(A_i) = \infty$ for independent A_i, then $P(\limsup_{n\to\infty} A_n) = 1$.
  - $P(\limsup_{n\to\infty} A_n) = 0$ means only finitely many A_n's can occur.
  - *Proof*: Start by showing when $P(\limsup_{n\to\infty} A_n) = 0$. Note that $\sum_{i=1}^{\infty} P(A_m)$ is a convergent series.

$$P(\limsup_{n\to\infty} A_n) = P\left(\bigcap_{n=1}^{\infty} \bigcup_{m\ge n} A_m\right) = \lim_{n\to\infty} P\left(\bigcup_{m\ge n} A_m\right) \le \lim_{n\to\infty} \sum_{m\ge n} P(A_m) = 0.$$

    Next, we show the other claim. By independence,

$$P((\limsup_{n\to\infty} A_n)^c) = \lim_{n\to\infty} P\left(\bigcap_{m\ge n} A_m^c\right) = \lim_{n\to\infty} \prod_{m\ge n} P(A_m^c) = \lim_{n\to\infty} \prod_{m\ge n} [1 - P(A_m)];$$

    Using the fact that $1 - x \le e^{-x}$ for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \prod_{m\ge n} [1 - P(A_m)] \le \lim_{n\to\infty} \prod_{m\ge n} \exp\{-P(A_m)\} = \lim_{n\to\infty} \exp\left\{-\sum_{m\ge n} P(A_m)\right\} = 0;$$

    Hence, $P(\limsup_{n\to\infty} A_n) = 1 - P((\limsup_{n\to\infty} A_n)^c) = 1 - 0 = 1$. ■
  - **Example**: Suppose $X_n \overset{\text{iid}}{\sim} \text{Exp}(1)$, and $A_n = \{X_n \ge c \log n\}$. If $c > 1$, then $P(A_n \text{ i.o.}) = 0$, but if $c \le 1$, then $P(A_n \text{ i.o.}) = 1$. ■
  - **Record**: X_n if $X_n > \bigvee_{i=1}^{n-1} X_i$.
  - **Renyi's Theorem**: Suppose $\{X_n, n \ge 1\}$ are iid with common CDF $F(x)$. Then, the sequence of RVs $\{R_n, n \ge 1\}$ is independent with $P(R_n = k) = \frac{1}{n}$ for $k \in \{1, \ldots, n\}$ (where $R_n = \sum_{i=1}^{n} \mathbb{1}\{X_i \ge X_n\}$), and the sequence of events $\{A_n, n \ge 1\}$ is independent and $P(A_n) = \frac{1}{n}$, where $A_n = \{X_n \text{ is a record}\}$.

– **Example**: Prove that for $\{X_n, n \geq 1\}$ is iid with continuous dist., $P\left(\left\{X_n = \bigvee_{i=1}^{n-1} Xi\right\} \text{ i.o.}\right) = 1$.

We need to show that $P\left(\limsup_{n\to\infty} X_n = \bigvee_{i=1}^{n} X_n\right) = 1$. Since $\{X_n\}$ are iid, then by the first Borel-Cantelli Lemma, we need to show that $\sum_{i=1}^{\infty} P(R_i) = \infty$, where R_n is the independent sequence of record times $R_n = \{X_n > \max\{X_1, \ldots, X_{n-1}\}\}$.
Applying Renyi's Theorem, since $\{X_n\}$ are once again independent, $P(R_n) = \frac{1}{n} \implies \sum_{i=1}^{\infty} P(R_i) = \sum_{i=1}^{\infty} \frac{1}{i} = \infty$. Therefore, $P\left(\limsup_{n\to\infty} X_n = \bigvee_{i=1}^{n} X_n\right) = 1$ by Borel-Cantelli. ■

- **Example**: Suppose $\{E_n\}$ is a sequents of events such that

$$\lim_{n\to\infty} P(E_n) = 0, \text{ and } \sum_n P(E_n E_{n+1}^c) < \infty.$$

Prove that $P(E_n \text{ i.o.}) = 0$.

Proceed with disjointification. Note that $E_n \setminus \bigcup_{i=k}^{n} E_i$ is disjoint from all $E_j \setminus \bigcup_{i=k}^{j} E_i$. Also note that $E_j \setminus \bigcup_{k=n}^{m-1} E_k \subset E_j \cap E_{j-1}^c$, since if an outcome is in $E_j \setminus \bigcup_{k=n}^{m-1} E_k$, then it cannot be in E_{j-1}. Therefore,

$$\lim_{n\to\infty} P\left(\bigcup_{j=m}^{n} E_j\right) = \lim_{n\to\infty} P(E_n) + \lim_{n\to\infty} \sum_{j=m+1}^{n} P\left(E_j \setminus \bigcup_{k=m}^{n-1} E_k\right) \leq (0) + \lim_{n\to\infty} \sum_{j=m+1}^{n} P\left(E_j \cap E_{j-1}^c\right) \to 0,$$

since $\sum_n P(E_n \cap E_{n-1}^c) < \infty \implies \lim_{n\to\infty} \sum_{j=n+1}^{\infty} P(E_n \cap E_{n-1}^c) \to 0$. Since both terms converge to zero, then $\forall \epsilon > 0$, $\exists n_0$ such that $\forall n \geq n_0$,

$$P(E_n) < \frac{\epsilon}{2} \text{ and } \sum_{k=n}^{\infty} P(E_k \cap E_{k-1}^c) < \frac{\epsilon}{2} \implies P\left(\bigcup_{j=n}^{\infty} E_j\right) < \epsilon.$$

Lastly, since ϵ is arbitrary, using the definition of lim sup,

$$P(E_n \text{ i.o.}) = P(\limsup_{n\to\infty} E_n) = \lim_{n\to\infty} P\left(\bigcup_{m\geq n} E_m\right) = 0. \blacksquare$$

- P**-Trivial**: A σ-field $\mathcal{A}$ such that $P(A) \in \{0, 1\}$ for all $A \in \mathcal{A}$.
  - Let X be an RV that is measurable wrt $\mathcal{A}$. Then, $P(X = c) = 1$ for some constant c.
    * $c = \inf\{x : P(X \leq x) = 1\}$.
- **Tail σ-Field**: $\mathcal{T}_\infty = \bigcap_{n=1}^{\infty} \mathcal{T}_n$, where $\mathcal{T}_n = \sigma\langle X_{n+1}, X_{n+2}\rangle$ for sequence of RVs X_n.
  - **Tail Event**: An event if its occurrence (or nonoccurrence) is unaffected by the values of any given finite number of variables.
  - **Tail-Measurable**: An RV such that its value is unaffected by any given finitely many X_n's.
    * **Example**: $A = \{\bar{X}_n \text{ converges}\}$ and $B = \{\bar{X}_n \to c\}$ are tail events. While changing the first few X_i's may matter for finite i, it won't matter for an infinite number of i's. ■
    * $A = \{\sum_{i=1}^{\infty} X_i \text{ converges}\}$ is a tail event. However, $A = \{\sum_{i=1}^{\infty} X_i = c\}$ is not a tail event, since changing the first few X_i's can greatly impact what value $\sum X_i$ will equal to. ■
- **Kolmogorov's Zero-One Law**: If $X_1, X_2, \ldots$ are independent, then any tail event is P-trivial.
  - *Proof*: Define $\mathcal{F}_n = \sigma\langle X_1, \ldots, X_n\rangle$ and $\mathcal{F}_\infty = \sigma\langle \cup_n \mathcal{F}_n\rangle = \sigma\langle X_1, X_2, \ldots\rangle$. Let $A \in \mathcal{T}_\infty$ and $B \in \mathcal{F}_\infty$. It is good enough to prove that the generator $\bigcup_{n=1}^{\infty} \mathcal{F}_n$ and $\mathcal{T}_\infty$ are independent.
    Suppose $B \in \bigcup_{n=1}^{\infty} \mathcal{F}_n$. $B \in \mathcal{F}_n$ for some n. $A \in \mathcal{T}_n$ by the definition of $\mathcal{T}_\infty$. Since $\mathcal{F}_n$ and $\mathcal{T}_n$ are independent by definition, A must be independent of B.
    However, $\mathcal{T}_\infty \subset \mathcal{F}_\infty$, so by choosing $B = A$ we have that $P(A) = P(A \cap A) = P(A)P(A) = P^2(A)$, which is only possible if $P(A) \in \{0, 1\}$. ■
  - **Example**: Suppose X_n's are independent. So, $P\left(\sum_{i=1}^{\infty} X_i < \infty\right) \in \{0, 1\}$, so under a given distribution, $\sum_{i=1}^{\infty} X_i$ either always converges, or it never will. ■

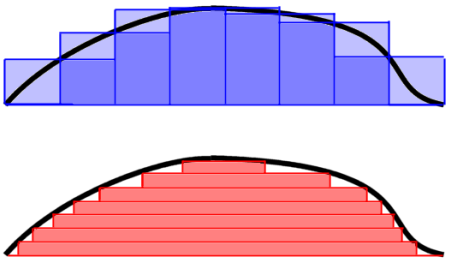
- **Example**: Show that for arbitrary events $\{A_n\}$, for all k, $\sum_{n\geq k} P\left(A_n | \cap_{i=1}^{n-1} A_i^c\right) = \infty$, then $P(\limsup_{n\to\infty} A_n) = 1$.

$$\begin{aligned} P\left((\limsup_{n\to\infty} A_n)^c\right) =& P(\liminf_{n\to\infty} A_n^c) = P\left(\bigcup_{n=1}^{\infty}\bigcap_{m\geq n} A_m^c\right) = \lim_{k\to\infty} P\left(\bigcap_{n\geq k} A_n^c\right) = \lim_{k\to\infty}\prod_{n\geq k} P\left(A_n^c \middle| \bigcap_{j=k}^{n-1} A_j^c\right) \\ &= \lim_{k\to\infty}\prod_{n\geq k}\left[1 - P\left(A_n \middle| \bigcap_{j=k}^{n-1} A_j^c\right)\right] \leq \lim_{k\to\infty}\prod_{n\geq k}\exp\left\{-P\left(A_n \middle| \bigcap_{j=k}^{n-1} A_j^c\right)\right\} \\ &= \lim_{n\to\infty}\exp\left\{-\sum_{n\geq k} P\left(A_n \middle| \bigcap_{j=k}^{n-1} A_j^c\right)\right\} \to \exp(-\infty) = 0. \blacksquare \end{aligned}$$

9.4 Integration and Expectations

Return to Table of Contents

- **Lebesgue Integration**: Splits the range space, rather than domain space used in Riemann integration. In other words, look at all x where $c \leq f(x) \leq d$ for finite c, d.



  - Above: Riemann integration (top) vs. Lebesgue integration (bottom).
  - Lebesgue integration allows the domain to be an abstract measurable space, as opposed to Riemann integration being restricted to the Lebesgue measure.
    - $\mathbb{1}(\mathbb{Q})$ is not Riemann-integrable, but is Lebesgue-integrable.
  - If f is Riemann-integrable on a bounded interval $[a, b]$, then f is also Lebesgue integrable with the Lebesgue integral $\int_{[a,b]} f d\lambda = \int_a^b f(x)dx$.
    - This is because the upper Riemann sum and the lower Riemann sum are both integrals of simple step functions that approximate f from above and below.
  - If f is continuous, then the Lebesgue integral can be evaluated by the second FTOC: $\int_a^b f(x)dx = F(b) - F(a)$.
    - **Example**: Integrate $f(x) = (1 + x^2)^{-1}$.

$$\int \frac{1}{1+x^2} d\lambda(x) \overset{MCT}{=} \lim_{n\to\infty}\int_{-n}^{n} \frac{1}{1+x^2} d\lambda(x) = \lim_{n\to\infty}\int_{-n}^{n} \frac{1}{1+x^2} dx = 2\lim_{n\to\infty}\tan^{-1}(n) = \pi < \infty. \blacksquare$$

  - If a term is Riemann-integrable on a compact set, then it is Lebesgue-integrable.

- **Expectation**, or $\mathbb{E}(\boldsymbol{X})$: For simple RV $X = \sum_{i=1}^k c_i \mathbb{1}_{A_i}$, where $A_1, \ldots, A_k \in \mathcal{A}$ are disjoint,

$$\mathbb{E}(X) = \int X dP = \int X(\omega) dP(\omega) = \int X(\omega) P(d\omega) = \sum_{i=1}^{k} c_i P(A_i).$$

  - $\int \mathbb{1}_A dP = P(A)$.
  - The expectation is well-defined. That is, if $X = \sum_{i=1}^k c_i \mathbb{1}_{A_i} = \sum_{i=1}^m d_i \mathbb{1}_{B_j}$, and if $A_i \cap B_j = \varnothing$, then $c_i = d_j$, and $\sum_{i=1}^k c_i P(A_i) \sum_{i=1}^m d_i P(B_i)$.

- Denote $X^+ = \max(X, 0)$ and $X^- = \max(-X, 0) = -\min(X, 0)$, where X is some general random variable.
  - $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$, which is defined except for when $\mathbb{E}(X^+) = \mathbb{E}(X^-) = \infty$.

- If $\mathbb{E}(X^+) = \infty$ and $\mathbb{E}(X^-) < \infty$, then $\mathbb{E}(X) = \infty$. Similarly, if $\mathbb{E}(X^-) = \infty$ and $\mathbb{E}(X^+) < \infty$, then $\mathbb{E}(X) = -\infty$.
- If $\mathbb{E}(X^+) < \infty$ and $\mathbb{E}(X^-) < \infty$, then $|\mathbb{E}(X)| < \infty$, and X is integrable.
- $\mathbb{E}(X)^+$ and $\mathbb{E}(X)^-$ are nonnegative! A lot of the proofs that follow show that a property holds for nonnegative X, and uses expansion $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$ to extend the property to the general X.
- $(X+Y)^+ \leq X^+ + Y^+$ and $(X+Y)^- \leq X^- + Y^-$.

- Properties of $\mathbb{E}(X)$ for general RVs X:
  - $\mathbb{E}(cX) = c\mathbb{E}(X)$.
    - *Proof*: This trivially holds for simple random variables. Suppose $X \geq 0$ and $0 \leq S \leq X$ for simple random variable S, such that $\mathbb{E}(X) \leq \mathbb{E}(X) + \epsilon$ for some $\epsilon > 0$. Let $c \geq 0$, so that $0 \leq cS \leq cX$. $\mathbb{E}(cS) \leq \mathbb{E}(cX)$, so $c\mathbb{E}(X) - c\epsilon \leq c\mathbb{E}(X) \leq \mathbb{E}(cX)$. Since $\epsilon > 0$ is arbitrary, $\mathbb{E}(cX) \geq c\mathbb{E}(X)$.
      To show the converse, note that $\mathbb{E}(X) = \mathbb{E}(c^{-1}cX) \geq \frac{1}{c}\mathbb{E}(cX)$. If $c < 0$, then
      $$\mathbb{E}(cX) = \mathbb{E}(cX)^+ - \mathbb{E}(cX)^- = 0 - \mathbb{E}\left[(-c)X\right] = c\mathbb{E}(X).$$
      Now, suppose X is general. $\mathbb{E}(cX) = \mathbb{E}(cX)^+ - \mathbb{E}(cX)^- = c[E(X)^+ - \mathbb{E}(X)^-] = c\mathbb{E}(X)$. ■
  - $X \geq Y \implies \mathbb{E}(X) \geq \mathbb{E}(Y)$.
    - *Proof*: Suppose X and Y are nonnegative. Since $X \geq Y$, if $0 \leq S \leq Y$, then $0 \leq S \leq X$. So,
      $$\mathbb{E}(Y) = \sup\{\mathbb{E}(S) : 0 \leq S \leq Y,\ S \text{ simple}\} \leq \sup\{\mathbb{E}(S) : 0 \leq S \leq X,\ S \text{ simple}\} = \mathbb{E}(X).$$
      Now, suppose we have general X and Y. $X \geq Y \implies X^+ \geq Y^+$, and $X^- \leq Y^-$. So,
      $$\mathbb{E}(X) = \mathbb{E}(X)^+ - \mathbb{E}(X)^- \geq \mathbb{E}(Y)^+ - \mathbb{E}(Y)^- = \mathbb{E}(Y).$$ ■
  - $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$.
    - *Proof*:
      $$X \leq |X| \implies \mathbb{E}(X) \leq \mathbb{E}(|X|);\ -X \leq |X| \implies -\mathbb{E}(X) \leq \mathbb{E}(|X|) \therefore |\mathbb{E}(X)| \leq \mathbb{E}(|X|).$$ ■
  - If $X \geq 0$, then $\int_A X dP = \sup\left\{\int_A S dP : 0 \leq S \leq X,\ S \text{ is simple}\right\}$.
    - *Proof*: First, suppose X is nonnegative, and let $0 \leq S \leq X \implies 0 \leq S\mathbb{1}_A \leq X\mathbb{1}_A$. Thus,
      $$\int_A X dP = \mathbb{E}(X\mathbb{1}_A) \geq \mathbb{E}(X\mathbb{1}_A) = \int_A S dP.$$
      So, $\sup\left\{\int_A S dP : 0 \leq S \leq X,\ S \text{ is simple}\right\} \leq \int_A X dP$.
      Now, show the other direction. Let $0 \leq X \leq X\mathbb{1}_A$, so $S = S\mathbb{1}_A$. Thus,
      $$\int_A X dP = \sup\left\{\int_A S dP : 0 \leq S \leq X\mathbb{1}_A,\ S \text{ simple}\right\} \leq \sup\left\{\int_A S dP : 0 \leq S \leq X,\ S \text{ simple}\right\}.$$ ■
  - If $\mathbb{E}(X)$ exists, then $\int_A X dP$ also exists. If $|\mathbb{E}(X)| < \infty$, then $\left|\int_A X dP\right| < \infty$.
    - *Proof*: $(X\mathbb{1}_A)^+ = X^+\mathbb{1}_A \leq X^+$, and $(X\mathbb{1}_A)^- = X^-\mathbb{1}_A \leq X^-$. So, if X is finite, then so is at least one of $\mathbb{E}(X\mathbb{1}_A)^+$ and $\mathbb{E}(X\mathbb{1}_A)^-$. ■
  - If $\mathbb{E}(X)$, $\mathbb{E}(Y)$, and $\mathbb{E}(X) + \mathbb{E}(Y)$ are well defined, then $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.
    - *Proof*: Let $X, Y \geq 0$. Create simple X_n, Y_n such that $0 \leq X_n \uparrow X$ and $0 \leq Y_n \uparrow Y$, and $(X_n + Y_n) \uparrow (X+Y)$. By the Monotone Convergence Theorem (defined later), $\mathbb{E}(X_n) \uparrow \mathbb{E}(X)$ and $\mathbb{E}(Y_n) \uparrow \mathbb{E}(Y) \implies \mathbb{E}(X_n + Y_n) \uparrow \mathbb{E}(X_n) + \mathbb{E}(Y_n)$.
      In addition, $\mathbb{E}(X_n + Y_n) = \mathbb{E}(X_n + Y_n)$, so $\mathbb{E}(X_n) + \mathbb{E}(Y_n) \to \mathbb{E}(X) + \mathbb{E}(Y)$. So, $\mathbb{E}(X_n + Y_n) \to \mathbb{E}(X+Y)$ and $\mathbb{E}(X_n + Y_n) \to \mathbb{E}(X) + \mathbb{E}(Y)$, so $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$. ■

- **Example**: Suppose $X \in L_1$ (the set of integrable RVs) and A, A_n are events.
  1. Show that $\int_{[|X_n| > n]} X dP \to 0$.
  2. Show that if $P(A_n) \to 0$, then $\int_{A_n} X dP \to 0$.
  3. Suppose that $(\Omega, \mathcal{B}, P)$ is a probability space and $A_i \in \mathcal{B}$ for $i \in \{1, 2\}$. Define the distance $d : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$ by $d(A_1, A_2) = P(A_1 \Delta A_2)$. Show that if $A_n, A \in \mathcal{B}$ and $d(A_n, A) \to 0$, then $\int_{A_n} X dP \to \int_A X dP$.

1. Note that $X\mathbb{1}\{|X_n| > n\} \to 0$ pointwise, and that since $X \in L_1$, it has finite expectation. Since $X\mathbb{1}\{|X_n| > n\} \le |X|$, where $\mathbb{E}|X| < \infty$, we can apply DCT to get that $\int_{[|X_n|>n]} X dP \to 0$. ■
2. Since $P(A_n) \to 0$, then for some large M such that $\int_{|X|>M} X dP < \epsilon$,

$$\left|\int_{A_n} X dP\right| \le \int_{A_n} |X| dP = \int_{A\cap\{|X|\le M\}} |X| dP + \int_{A\cap\{|X|>M\}} |X| dP \le M\int_{A_n} dP + \int_{|X|>M} X dP;$$

This has a $\limsup < \epsilon$. Since ϵ is arbitrary, the limit must be 0.
3. This is equivalent to showing that $\left|\int_{A_n} X dP - \int_A X dP\right| \to 0$.

$$\left|\int_{A_n} X dP - \int_A X dP\right| = |\mathbb{E}[X\mathbb{1}_{A_n}] - \mathbb{E}[X\mathbb{1}_A]| = |\mathbb{E}\left[X(\mathbb{1}_{A_n} - \mathbb{1}_A)\right]|;$$

We know that $|\mathbb{E}\left[X(\mathbb{1}_{A_n} - \mathbb{1}_A)\right]| \le \mathbb{E}\big[|X| \cdot |\underbrace{\mathbb{1}_{A_n} - \mathbb{1}_A}_{\to \varnothing}|\big] \to 0$. ■

- **Example**: Show that for independent RVs X and Y, where $\mathbb{E}(X)$ exists, then for all $B \in \mathcal{B}(\mathbb{R})$, $\int_{[Y\in B]} X dP = \mathbb{E}(X)P(Y \in B)$.

  Since X and Y are independent, so must X and $\mathbb{1}_{Y\in B}$.

$$\int_{\mathbb{1}\{Y\in B\}} X dP = \mathbb{E}\left[X\mathbb{1}_{Y\in B}\right] = \mathbb{E}(X)\mathbb{E}\left[\mathbb{1}_{Y\in B}\right] = \mathbb{E}(X)P(Y \in B).$$ ■

- Suppose $X_n \to X$ pointwise. $\mathbb{E}(X_n)$ doesn't always converge to $\mathbb{E}(X)$!
  - **Example**: $X_n(\omega) = n^2\mathbb{I}\left(\omega \le \frac{1}{n}\right)$, and use the Lebesgue measure on $(0, 1]$. Although $X_n \to X = 0$ pointwise, $\mathbb{E}(X_n) = n \to \infty$. ■
  - **Monotone Convergence Theorem**, or **MCT**: Suppose $X_n \ge 0$ is monotone increasing, with $X_n \to X$. Then, $\mathbb{E}(X_n) \to \mathbb{E}(X)$.
    * *Proof*: We know that $0 \le X_n \le X \implies \mathbb{E}(X_n) \le \mathbb{E}(X)$, and $\mathbb{E}(X_n) \le \mathbb{E}(X_{n+1})$ by monotone increasing, so $\ell := \lim_{n\to\infty} \mathbb{E}(X_n) \le \mathbb{E}(X)$.
      Now, suppose $0 < b < 1$, and $0 \le S \le X$, where S is simple and finite-valued. Put $B_n = \{\omega : X_n(\omega) \ge bS(\omega)\} \uparrow \Omega$ as $n \to \infty$. Thus,

$$\ell \ge \mathbb{E}(X_n) \ge b\int_{B_n} S dP \to b\mathbb{E}(S).$$

      This is because $b\int_{B_n} S dP = b \cdot \sum_{i=1}^k c_i P(A_i \cap B_i)$, and since $B_n \uparrow \Omega$, $A_i \cap B_i \uparrow A_i$. Take the supremum over S to get $b\mathbb{E}(X) \le \ell$. Since $b < 1$ is arbitrary, $\mathbb{E}(X) \le \ell$. ■
    * If X_n is any sequence of nonnegative, simple RVs increasing to X, then $\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(X_n)$.
    * **Example**: Prove that if $\mathcal{F}$ is a σ-field on Ω, then the smallest monotone class containing $\mathcal{F}$ is $\sigma\langle\mathcal{F}\rangle$.

      Let $\mathcal{G}$ be the smallest monotone class containing $\mathcal{F}$. Trivially, $\mathcal{G} \subset \sigma\langle\mathcal{F}\rangle$. Next, we show that $\mathcal{G}$ is itself a σ-field, so then $\sigma\langle\mathcal{F}\rangle \subset \mathcal{G} \equiv \mathcal{G} = \sigma\langle\mathcal{F}\rangle$.
      $\mathcal{G}$ is a monotone class. If $\mathcal{G}$ is also a field, then $\mathcal{G}$ is also a σ-field, since the countable union is a monotone increasing limit of a finite union.
      Now, we show that $\mathcal{G}$ is a field. Applying an earlier result, it suffices to show that $\Omega \in \mathcal{G}$ and $A \cap B^c \in \mathcal{G}$, where $A, B \in \mathcal{G}$. Trivially, $\Omega \in \mathcal{G}$.
      Apply bootstrapping. We prove that if $A \in \mathcal{F}$, and $B \in \mathcal{G}$, then $A \cap B^c \in \mathcal{G}$. Let $\mathcal{G}_A = \{B \in \mathcal{G} : A \cap B \in \mathcal{G}\}$. Clearly, $\mathcal{F} \subset \mathcal{G}_A$, since if $B \in \mathcal{F}$, then $A \cap B^c \in \mathcal{F} \subset \mathcal{G}$.
      Next, we claim that $\mathcal{G}_A$ is monotone. Let $B_n \in \mathcal{G}_A$ such that $B_n \uparrow B$. A consequence of this is that $B_n^c \downarrow B^c$. $B \in \mathcal{G}$, since $\mathcal{G}$ is monotone. Therefore, $(A \cap B_n^c) \downarrow (A \cap B^c)$, and so $(A \cap B^c) \in \mathcal{G}$. Similarly, if $B_n \downarrow B$, then $B \in \mathcal{G}_A$. Thus, $\mathcal{G}_A = \mathcal{G}$. That is, $\forall A \in \mathcal{F}$ and $B \in \mathcal{G}$, $A \cap B^c \in \mathcal{G}$.
      Next, fix $B \in \mathcal{G}$ and let $\mathcal{G}_B = \{A \in \mathcal{G} : A \cap B^c \in \mathcal{G}\}$. We now claim that $\mathcal{G}_B = \mathcal{G}$.
      Clearly, $\mathcal{F} \subset \mathcal{G}_B$ by the last step, so it suffices to show that $\mathcal{G}_B$ is a monotone class. If this is the case, then $\mathcal{G}_B \supset \mathcal{G}$, and thus $\mathcal{G}_B = \mathcal{G}$.
      Let $A_n \in \mathcal{G}_B$, where $A_n \uparrow A$. This means A_n is monotone, and so $A \in \mathcal{G}$. Now, $(A_n \cap B^c) \uparrow (A \cap B^c)$, so $(A_n \cap B^c) \in \mathcal{G}$. Since $A_n \in \mathcal{G}_B$, $(A \cap B^c) \in \mathcal{G}$, so $A \in \mathcal{G}_B$.
      Similarly, if $A_n \in \mathcal{G}_B$, where $A_n \downarrow A$, then $A \in \mathcal{G}_B$. This means that $\mathcal{G}_B$ is a monotone class, and so $\mathcal{G}_B = \mathcal{G}$, since $\mathcal{G}$ is the smallest monotone class that contains $\mathcal{F}$.
      Therefore, given $A, B \in \mathcal{G}$, $A \cap B^c \in \mathcal{G}$. This means that $\mathcal{G}$ is a field. Since $\mathcal{G}$ is monotone class, $\mathcal{G}$ is a σ-field, so $\mathcal{G} = \sigma\langle\mathcal{F}\rangle$. ■

- If $X_n \geq 0$ is measurable, then $\mathbb{E}\left(\sum_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} \mathbb{E}(X_i)$.
  - Generalizes the finite sum we've proven earlier with a countable sum.
    - The limit of the series must exist!
  - *Proof*: Apply MCT on the sequence of partial sums. If $X_n \geq 0$, then $\mathbb{E}(X_n)$ is monotone increasing (so the limit must exist).

$$\mathbb{E}\left(\sum_{i=1}^{\infty} X_i\right) = \mathbb{E}\left(\lim_{n\to\infty} \sum_{i=1}^{n} X_i\right) \overset{\text{MCT}}{=} \lim_{n\to\infty} \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{\infty} \mathbb{E}(X_i). \blacksquare$$

- If $X_n \geq Y$, $\mathbb{E}(Y) > -\infty$, and $X_n \uparrow X$, then $\mathbb{E}(X_n) \uparrow \mathbb{E}(X)$, and if $X_n \leq Y$, $\mathbb{E}(Y) < \infty$, and $X_n \downarrow X$, then $\mathbb{E}(X_n) \downarrow \mathbb{E}(X)$.
  - *Proof*: Apply MCT on $X_n - Y$ and $Y - X_n$.

$$\mathbb{E}(X_n - Y) \uparrow \mathbb{E}(X - Y) \implies [\mathbb{E}(X_n) - \mathbb{E}(Y)] \uparrow [\mathbb{E}(X) - \mathbb{E}(Y)] \implies \mathbb{E}(X_n) \uparrow \mathbb{E}(X).$$

$$\mathbb{E}(Y - X_n) \downarrow \mathbb{E}(Y - X) \implies [\mathbb{E}(Y) - \mathbb{E}(X_n)] \downarrow [\mathbb{E}(Y) - \mathbb{E}(X)] \implies \mathbb{E}(X_n) \downarrow \mathbb{E}(X). \blacksquare$$

- **Expectation (Independent RVs)**: If $X_1, \ldots, X_k$ are independent with finite expectations, then $\mathbb{E}(X_1, \ldots, X_k) = \prod_{i=1}^{k} \mathbb{E}(X_i)$.
  - *Proof*: WLOG, only need to show $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, for two independent events X and Y. This is because we can group several RVs into Y.
    We first show that this holds for indicators. Let $X = \mathbb{1}_A$ and $Y = \mathbb{1}_B$ for independent X and Y. Note that $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$.

$$\mathbb{E}(\mathbb{1}_{A\cap B}) = P(A \cap B) \overset{\perp}{=} P(A)P(B)$$

Now, we generalize this to apply to simple RVs. Let $X = \sum_i a_i \mathbb{1}_{A_i}$ and $Y = \sum_j b_j \mathbb{1}_{B_j}$.

$$E(XY) = \sum_i \sum_j a_i b_j \mathbb{1}_{A_i \cap B_j} = \sum_i \sum_j a_i b_j P(A_i \cap B_j) = \sum_i \sum_j a_i P(A_i) b_j P(B_j) = \underbrace{\left(\sum_i a_i \mathbb{1}_{A_i}\right)}_{\mathbb{E}(X)} \underbrace{\left(\sum_j b_j \mathbb{1}_{B_j}\right)}_{\mathbb{E}(Y)}.$$

Now, we further generalize this to apply to nonnegative X, Y. Let $X_n \uparrow X$ and $Y_n \uparrow Y$ be simple.

$$\lim_{n\to\infty} \mathbb{E}(X_n Y_n) = \lim_{n\to\infty} \mathbb{E}(X_n)\mathbb{E}(Y_n) = \mathbb{E}(X)\mathbb{E}(Y).$$

In addition, using the MCT,

$$0 \leq X_n Y_n \uparrow XY \implies \lim_{n\to\infty} \mathbb{E}(X_n Y_n) = \mathbb{E}(XY) \implies \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Lastly, we prove this for the general X, Y. Note that $(XY)^+ = X^+Y^+ + X^-Y^-$, and $(XY)^- = X^+Y^- + X^-Y^+$. Assume finite expectations of all relevant quantities.

$$\begin{aligned}
\mathbb{E}(XY)^+ &= \mathbb{E}(X^+Y^+) + \mathbb{E}(X^-Y^-) = \mathbb{E}(X^+)\mathbb{E}(Y^+) + \mathbb{E}(X^-)\mathbb{E}(Y^-); \\
\mathbb{E}(XY)^- &= \mathbb{E}(X^+Y^-) + \mathbb{E}(X^-Y^+) = \mathbb{E}(X^+)\mathbb{E}(Y^-) + \mathbb{E}(X^-)\mathbb{E}(Y^+); \\
\mathbb{E}(XY) &= \mathbb{E}(XY)^+ - \mathbb{E}(XY)^- \\
&= \mathbb{E}(X^+)\mathbb{E}(Y^+) + \mathbb{E}(X^-)\mathbb{E}(Y^-) - \left[\mathbb{E}(X^+)\mathbb{E}(Y^-) + \mathbb{E}(X^-)\mathbb{E}(Y^+)\right] \\
&= E(X^+)[\mathbb{E}(Y^+) - \mathbb{E}(Y^-)] - E(X^-)[\mathbb{E}(Y^+) - \mathbb{E}(Y^-)] \\
&= [\mathbb{E}(X^+) - \mathbb{E}(X^-)][\mathbb{E}(Y^+) - \mathbb{E}(Y^-)] = \mathbb{E}(X)\mathbb{E}(Y). \blacksquare
\end{aligned}$$

- **Example**: A sequence of RVs X_n is uniformly integrable from above if $\sup_n \mathbb{E}\left[X_n \mathbb{1}\{X_n > C\}\right] \to 0$ as $C \to \infty$.

1. Show that $\log X_n$ is uniformly integrable from above if X_n are positive RVs with $\sup_n \mathbb{E}(X_n) < \infty$.
2. Show that the converse direction is not necessarily true.

1. Let $Y_n = \log X_n$, and $\epsilon > 0$.

$$\mathbb{E}\left[Y_n \mathbb{1}\{Y_n > \epsilon\}\right] = \mathbb{E}\left[\frac{Y_n}{X_n} \cdot X_n \mathbb{1}\{Y_n > \epsilon\}\right] = \mathbb{E}\left[\frac{Y_n}{e^{Y_n}} \cdot X_n \mathbb{1}\{Y_n > \epsilon\}\right] \leq \frac{\epsilon}{e^{\epsilon}} \cdot \mathbb{E}(X_n).$$

Since $\mathbb{E}(X_n) < \infty$, $\sup_n \mathbb{E}\left[Y_n \mathbb{1}\{Y_n > \epsilon\}\right] \to 0$ as $\epsilon \to \infty$. This means that Y_n is uniformly integrable from above.

2. Consider $X_n = \frac{1}{n}\mathbb{1}_{(0,1/2]} + \mathbb{1}_{(1/2,1]}$. $X_n > 0$, and $\mathbb{E}(X_n) = \frac{1}{2n} + \frac{1}{2} \leq 1 < \infty$ for all n. However, $Y_n = \log X_n = (-\log n)\mathbb{1}_{(0,1/2]}$ has expectation $\frac{1}{2}\log n \to \infty$ as $n \to \infty$. So, Y_n is not uniformly integrable. ■

- **Example**: Let $p_1, \ldots, p_k > 1$ such that $\sum_{j=1}^k p_j^{-1} = 1$. Let X_j, $j \in \{1, \ldots, k\}$ be RVs such that $\|X_j\|_{p_j} := (\mathbb{E}|X_j|^{p_j})^{1/p_j} < \infty$. Show that $\mathbb{E}\left(\prod_{j=1}^k |X_j|\right) \leq \prod_{j=1}^k \|X_j\|_{p_j}$.

  First, a corollary. We show that for positive $x_1, \ldots, x_k$, $\prod_{i=1}^k x_i \leq \prod_{i=1}^k \frac{x_i^{p_i}}{p_i}$, where $\sum_{i=1}^k p_i^{-1} = 1$.

$$\log\left(\prod_{i=1}^k x_i\right) = \sum_{i=1}^k \log(x_i) = \sum_{i=1}^k \frac{1}{p_i}\log(x_i)^{p_i} \leq \log\left(\sum_{i=1}^k \frac{x_i^{p_i}}{p_i}\right).$$

  By concavity of logs, $\prod_{i=1}^k x_i \leq \prod_{i=1}^k \frac{x_i^{p_i}}{p_i}$.
  Now, let $x_i = \frac{|X_i|}{(\mathbb{E}[|X_i|^{p_i}])^{1/p_i}}$.

$$\prod_{i=1}^k \frac{|X_i|}{(\mathbb{E}[|X_i|^{p_i}])^{1/p_i}} \leq \sum_{i=1}^k \frac{|X_i|^{p_i}}{p_i \cdot \mathbb{E}[|X_i|^{p_i}]} \implies \frac{\mathbb{E}\left[\prod_{i=1}^k |X_i|\right]}{\prod_{i=1}^k (\mathbb{E}[|X_i|^{p_i}])^{1/p_i}} \leq \sum_{i=1}^k \frac{\mathbb{E}[|X_i|^{p_i}]}{p_i \cdot \mathbb{E}[|X_i|^{p_i}]}.$$

  Rewrite the expression to get that $\mathbb{E}\left(\prod_{i=1}^k |X_i|\right) \leq \prod_{i=1}^k (\mathbb{E}[|X_i|^{p_i})^{1/p_i}$. ■

- **Example**: Let μ be a measure on a set Ω with σ-field $\mathcal{A}$. Let $f : \Omega \to [0, \infty)$ be a measurable function such that $\int f d\mu = 1$. Define $P : \mathcal{A} \to \mathbb{R}$ by the relation $P(A) = \int_A f d\mu$ for $A \in \mathcal{A}$.

  1. Show that P is a probability measure on $(\Omega, \mathcal{A})$.
  2. If g is another nonnegative measurable function on Ω with $\int g d\mu = 1$, show that $\int |f - g| d\mu = 2\left[P(A_0) - Q(A_0)\right]$, where $A_0 = \{\omega : f(\omega) \geq g(\omega)\}$, and $Q(A) = \int_A g d\mu$ for $A \in \mathcal{A}$.

  1. Show that the definition of a probability measure is satisfied. $P(A) \geq 0$ for any $A \in \mathcal{A}$, since f maps to $[0, \infty)$. In addition, $P(\Omega) = 1$.

$$P\left(\bigsqcup_{n=1}^\infty A_n\right) = \int_{\bigsqcup_{n=1}^\infty A_n} f d\mu = \int f \mathbb{1}_{\bigsqcup_{n=1}^\infty A_n} d\mu = \int f\left(\mathbb{1}_{A_n}\right) d\mu \overset{\text{MCT}}{=} \sum_{i=1}^\infty \int f \mathbb{1}_{A_i} d\mu = \sum_{i=1}^\infty P(A_i);$$

     Thus, P is countably additive, so by definition, P is a probability measure.

  2.

$$\int |f-g| d\mu = \int_{f \geq g} (f-g) d\mu + \int_{g > f} (g-f) d\mu = P(A_0) - Q(A_0) + Q(A_0^c) - P(A_0^c) = 2\left[P(A_0) - Q(A_0)\right]. \blacksquare$$

- **Example**: Let $X_1, X_x, \ldots$ be iid with CDF $F(x) = 1 - \exp\{-x^\alpha\}$ for $x \geq 0$ and $\alpha > 0$ is a parameter.

  1. For $A_n = \{X_n \geq c(\log n)^{1/\alpha}\}$, show that $P(\limsup_{n\to\infty} A_n) = \mathbb{I}(c \leq 1)$.
  2. Show that $P\left(\limsup_{n\to\infty} (\log n)^{-1/\alpha} X_n\right) = 1$.

  1. Note that A_n are independent events. In addition, $P(A_n) = n^{-c^\alpha}$.
     If $c > 1 \equiv c^\alpha > 1$, then $\sum_{i=1}^\infty P(A_i) < \infty$. Conversely, if $c \geq 1$, then $\sum_{i=1}^\infty P(A_i) = \infty$. Applying the Borel-Cantelli lemmas yields the results.
  2. For all $\epsilon > 0$,

$$P\left(\limsup_{n\to\infty} \frac{X_n}{(\log n)^{1/\alpha}} = 1\right) = \begin{cases} P\left(\frac{X_n}{(\log n)^{1/\alpha}} > 1 + \epsilon \text{ i.o.}\right) = 0 & , c = 1 + \epsilon \\ P\left(\frac{X_n}{(\log n)^{1/\alpha}} > 1 - \epsilon \text{ i.o.}\right) = 1 & , c = 1 - \epsilon \end{cases}. \blacksquare$$

- **Example**: Suppose $X_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$, such that $\mathbb{E}(X_n^2) = \mu_n^2 + \sigma_n^2$ for $n \in \{1, 2, \ldots\}$. Assume that $\sum_{i=1}^\infty \mu_i^2$ and $\sum_{i=1}^\infty \sigma_i^2 < \infty$.

  1. Justify the claim that $\sum_{n=1}^\infty X_n^2$ is an RV.
  2. Compute $\mathbb{E}\left(\sum_{n=1}^\infty X_n^2\right)$.

1. Since X_n is an RV, and X_n^2 is continuous (and hence measurable), then $\sum_{n=1}^N X_n^2$ is also an RV. $\sum_{n=1}^\infty X_n^2 = \lim_{N\to\infty}\sum_{n=1}^N X_n^2$, so $\sum_{n=1}^\infty X_n^2$ must be an RV.
2. Clearly, $\sum_{n=1}^N X_n^2 \uparrow \sum_{n=1}^\infty X_n^2$. Therefore,

$$\mathbb{E}\left(\sum_{n=1}^\infty X_n^2\right) \overset{\text{MCT}}{=} \lim_{N\to\infty}\left(\sum_{n=1}^N X_n^2\right) = \lim_{N\to\infty}\sum_{n=1}^N \mathbb{E}(X_n^2) = \lim_{N\to\infty}\sum_{n=1}^N(\mu_n^2+\sigma_n^2) = \sum_{i=1}^\infty \mu_i^2 + \sum_{i=1}^\infty \sigma_i^2. \blacksquare$$

- **Dominated Convergence Theorem**, or **DCT**: If $X_n \to X$ pointwise and $|X_n| \le Y$ for all n, where $\mathbb{E}(Y) < \infty$, then $\mathbb{E}(X_n) \to \mathbb{E}(X)$.
  - Does not require monotonicity.
  - **Fatou's Lemma**: If $X_n \ge Y$, where $\mathbb{E}(Y) > -\infty$, then $\mathbb{E}(\liminf_{n\to\infty} X_n) \le \liminf_{n\to\infty}\mathbb{E}(X_n)$. If $X_n \le Y$, where $\mathbb{E}(Y) < \infty$, then $\mathbb{E}(\limsup_{n\to\infty} X_n) \ge \limsup_{n\to\infty}\mathbb{E}(X_n)$.
    - *Proof*: Let $U_n = \inf_{k\ge n} X_k \ge Y$, where $U_n \uparrow \liminf_{n\to\infty} X_n =: X$. Then, $\mathbb{E}(U_n) \uparrow \mathbb{E}(X)$ by MCT. However, $U_n \le X_n$, so $\mathbb{E}(U_n) \le \mathbb{E}(X_n)$, so $\mathbb{E}(X) = \lim_{n\to\infty}\mathbb{E}(U_n) \le \lim_{n\to\infty}\mathbb{E}(X_n)$.
      Let $V_n = -X_n$, where $V_n \ge -Y$, and $\mathbb{E}(-Y) > -\infty$, and $\liminf_{n\to\infty} V_n = -\limsup_{n\to\infty} X_n$. Then,
      $$\mathbb{E}(\limsup_{n\to\infty} X_n) = -\mathbb{E}(\liminf_{n\to\infty} V_n) \ge -\liminf_{n\to\infty}\mathbb{E}(V_n) = \limsup_{n\to\infty}\mathbb{E}(X_n). \blacksquare$$
    - Fatou only provides a one-sided result, but is used to prove DCT.
  - *Proof*: $|X_n| \le Y \equiv -Y \le X_n \le Y$. $E(Y) < \infty \implies \mathbb{E}(-Y) > -\infty$, and X_n is integrable. Since $X_n \to X$, $\limsup_{n\to\infty} X_n = \liminf_{n\to\infty} X_n = X$.
    By Fatou's lemma,
    $$\mathbb{E}(X) \le \liminf_{n\to\infty}\mathbb{E}(X_n) \le \limsup_{n\to\infty}\mathbb{E}(X_n) \le \mathbb{E}(X) \implies \mathbb{E}(X_n) \to \mathbb{E}(X). \blacksquare$$
  - **Bounded Convergence Theorem**: For finite measures, if $X_n \to X$ pointwise, and $|X_n| \le C$ for all n and constant C, then $\mathbb{E}(X_n) \to \mathbb{E}(X)$.
    - We restrict to finite measures, since constants may not be integrable for non-finite measures.
  - For any integrable RV X, $\mathbb{E}\left[X\mathbb{1}\{|X| \le n\}\right] \to \mathbb{E}(X)$, and $\mathbb{E}\left[X\mathbb{1}\{|X| \ge n\}\right] \to 0$.
    - *Proof*: Apply DCT.
      $$\mathbb{E}(X) = \int_{|X|\le n} X dP + \int_{|X|\ge n} X dP \to \mathbb{E}(X) + 0 = \mathbb{E}(X);$$
      As a result, $\lim_{n\to\infty} X\mathbb{1}\{|X| \le n\} = X$, and $\lim_{n\to\infty} X\mathbb{1}\{|X| \ge n\} = 0$. $\blacksquare$
    - We only need one of $\mathbb{E}(X^+)$ and $\mathbb{E}(X^-)$ to be finite. We apply MCT on the finite one. For instance, if $\mathbb{E}(X^+)$ is finite, then $X^+\mathbb{1}\{|X| \le n\} \to X^+$. If neither is finite, we may end up with an undefined $\infty - \infty$ case.
  - If $X_n \ge 0$, where $X_n \to X$ pointwise and $\mathbb{E}(X_n) \to \mathbb{E}(X)$ for integrable X, then $\mathbb{E}(|X_n - X|) \to 0$.
    - *Proof*: $|X_n - X| = X_n + X - 2\min\{X_n, X\}$. $\min\{X_n, X\} \to X$, and $0 \le \min\{X_n, X\} \le X$, so by DCT, $\mathbb{E}[\min\{X_n, X\}] \to \mathbb{E}(X)$. Thus,
      $$\mathbb{E}(|X_n - X|) = \mathbb{E}(X_n) + \mathbb{E}(X) - 2\mathbb{E}[\min\{X_n, X\}] \to \mathbb{E}(X) + \mathbb{E}(X) - 2\mathbb{E}(X) = 0. \blacksquare$$
  - If f_n is a sequence of probability densities converging to probability density f (wrt measure μ), then $\int |f_n - f| d\mu \to 0$.
    - *Proof*: Similar to the above corollary.
      $$\int |f_n - f| d\mu = \int f_n d\mu + \int f d\mu - 2\int \min\{f_n, f\} d\mu \to 1 + 1 - 2 = 0. \blacksquare$$
  - **Example**: Lebesgue integral on $\mathbb{R}$.
    $$\int \frac{\sin x}{1+x^2} d\lambda(x) \overset{\text{DCT}}{=} \lim_{n\to\infty}\int_{-n}^n \frac{\sin x}{1+x^2} d\lambda(x) = \lim_{n\to\infty}\int_{-n}^n \frac{\sin x}{1+x^2} dx = 2\lim_{n\to\infty} 0 = 0. \blacksquare$$
    - The Lebesgue integral would not be defined for $\lambda([0,\infty))$.

- **Example**: Regularity conditions for score function. Suppose $f(x;\theta)$ is differentiable in θ. In other words, $\sup_{|\theta-\theta_0|<\delta} \left|\frac{\partial}{\partial\theta} f(x;\theta)\right| \le H(x)$, where $\int H(x)d\mu(x) < \infty$. Then,

$$\frac{d}{d\theta}\int f(x;\theta)d\mu(x) = \int \frac{\partial}{\partial\theta} f(x;\theta)d\mu(x).$$

  Therefore, $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log p(x;\theta)\right] = 0$.
- **Example**: Limit of a series. Suppose we have sequence a_{nk}, where $|a_{nk}| \le b_k$, and $\sum_{i=1}^\infty b_i < \infty$, and $a_{nk} \to a_k$ as $n \to \infty$.
  Then, $\lim_{n\to\infty}\sum_{k=1}^\infty a_{nk} = \sum_{k=1}^\infty a_k$ by DCT.
  If instead, $0 \le a_{nk} \uparrow a_k$ as $n \to \infty$ by MCT, the result still holds. ■
- For results regarding expectations, WLOG we can insert "almost surely" wherever needed.
  - Applies to MCT, Fatou's lemma, DCT, and all results.
- **Example**: Show for nonnegative RV X, $n\mathbb{E}[X^{-1}\mathbb{1}\{X > n\}] \to 0$, and $n^{-1}\mathbb{E}[X^{-1}\mathbb{1}\{X > n^{-1}\}] \to 0$.

  First, note that we only need to consider the cases where $X > n$. This is because, when $X \le n$, then the indicator term will be zero. So, when $X > n$, $\frac{n}{X}\cdot\mathbb{1}\{X > n\} \le \mathbb{1}\{X > n\} \to 0$. In addition, we know that $\frac{n}{X}\cdot\mathbb{1}\{X > n\} \le 1$. Since we have bounded $\frac{n}{X}\cdot\mathbb{1}\{X > n\} \le 1$ and shown pointwise convergence, we apply the DCT to get that $n\mathbb{E}[X^{-1}\{X > n\}] \to 0$.
  Similarly, $\frac{1}{nX} \to 0$. Thus, $\frac{1}{nX}\cdot\mathbb{1}\{nX > 1\} \to 0$, and is bounded above by 1. The DCT once again results in $n^{-1}\mathbb{E}[X^{-1}\{X > n^{-1}\}] \to 0$. ■
- **Example**: Let X_n, X, Y be RVs on a probability space $(\Omega, \mathcal{A}, P)$ satisfying $|X_n| \le Y$ for all $n = 1, 2, \dots, X_n \to X$ and $\mathbb{E}(Y^2) < \infty$. Show that $\mathbb{E}[(X_n - X)^2] \to 0$ as $n \to \infty$.

  $(X_n - X)^2 = X_n^2 + X^2 - 2X_nX$. We need to show that $\mathbb{E}(X_n^2) \to \mathbb{E}(X^2)$, and $\mathbb{E}(X_n^2X) \to \mathbb{E}(X^2)$.
  $|X_n| \le Y \implies X_n^2 \le Y^2$, and since $X_n^2 \to X^2$ and $\mathbb{E}(Y^2) < \infty$, then by the DCT, $\mathbb{E}(X_n^2) \to \mathbb{E}(X^2)$.
  We also know that $|X_n| \le |Y| \implies |X| \le |Y|$, so $|X_nX| \le Y \cdot Y = Y^2$. Therefore, by the DCT once again, $\mathbb{E}(X_nX) \to \mathbb{E}(X^2)$.
  Therefore,

$$\mathbb{E}[(X_n - X)^2] = \mathbb{E}(X_n^2) + \mathbb{E}(X^2) - 2\mathbb{E}(X_nX) \to \mathbb{E}(X^2) + \mathbb{E}(X^2) - 2\mathbb{E}(X^2) = 0.\ ■$$

- **Example**: Suppose $\{p_k, k \ge 0\}$ is a PMF on $\{0, 1, \dots\}$, and define $P(s) = \sum_{k=0}^\infty p_k s^k$ for $s \in [0, 1)$. Prove that $\frac{d}{dx}P(s) = \sum_{k=1}^\infty p_k k s^{k-1}$.

  We know that $P(s)$ is convergent for $|s| < 1$. This means that for any $s \in [0, 1)$, $\exists\delta > 0$ such that $(s - \delta, s + \delta) \subset (-1, 1)$.
  Apply the limit definition of a derivative. Suppose $h^{-1}\left(P(s + h) - P(s)\right) = \sum_{k=1}^\infty p_k \frac{(s+h)^k - s^k}{h}$. By the previous claim, $\frac{(s+h)^k - s^k}{h} = k\xi^{k-1}$ for some $\xi \in (s - \delta, s + \delta)$, where k can be absorbed into an exponential with a slightly increased δ. Thus, $p_k\frac{(s+h)^k - s^k}{h}$ can be dominated by $p_k\left(|s + \delta|^k - |s - \delta|^k\right)$, which is summable. Thus, by the DCT, the limit of $\frac{d}{dx}P(s)$ is $\sum_{k=1}^\infty p_k k s^{k-1}$. ■

- Some results about X given $\mathbb{E}(X)$:
  1. $\mathbb{E}(|X|) < \infty \implies |X| < \infty$ a.s.
     - *Proof*:

$$\mathbb{E}(|X|) = \int_{|X|<\infty} |X|dP + \infty \cdot P(|X| = \infty);$$

       $P(|X| = \infty) = 0$ a.s., so the entire term is finite, so $|X| < \infty$ a.s. ■
  2. If $X \ge 0$ and $\mathbb{E}(X) = 0$ a.s., then $X = 0$ a.s.
     - *Proof*: Proceed with contradiction. Suppose $X \ne 0$ a.s, with $\mathbb{E}(X) = 0$. As a result, $P(X > 0) > 0 \implies P(X > c) > 0$ for some $c > 0$. This means $\mathbb{E}(X) \ge cP(X \ge c) > 0$, which is a contradiction. Hence, $X = 0$ a.s. ■
  3. $\int_A X dP \ge 0\ \forall A \implies X \ge 0$ a.s.
     - *Proof*: Proceed with contradiction. Suppose $P(X < 0) > 0$, and so $\exists c < 0 : P(X < c) > 0$. This results in $A = \{X < c\} \implies \int_A X dP \le cP(A) < 0$, which is a contradiction. ■
  4. $\int_A X dP = 0\ \forall A \implies X = 0$ a.s.

- *Proof*: $\int_A X dP \geq 0$ for all A, so $X \geq 0$ by the previous result. In addition, $\int_A (-X) dP \geq 0$ for all A, so $-X \geq 0$ a.s. as well. Therefore, $|X| = 0 \equiv X = 0$ a.s. ■

- **Change of Variable Rule**: Let $(\Omega, \mathcal{A}, P)$ be a probability space. Let $T : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ be a measurable space. Also let P' be the measure induced by T on $(\Omega', \mathcal{A}')$. In other words, $P'(A') = P(T^{-1}(A'))$ for $A' \in \mathcal{A}'$. Then, for any RV X on Ω',

$$\mathbb{E}(X) := \int X(\omega') dP'(\omega') = \int X(T(\omega)) dP(\omega) = \mathbb{E}(X \circ T).$$

  - **Example**: Suppose $X \sim \mathcal{N}(0,1)$. This means that $\Phi(x) = \mathcal{N}(0,1)$ is our measure. Define $Y = X^2$. $\mathbb{E}(Y) = \int x^2 d\Phi(x) = 1$. ■
  - *Proof*: Start by showing it's true for indicators. Let $X = \mathbb{1}_{A'}$, where $A' \in \mathcal{A}'$.

$$\int \mathbb{1}_{A'}(\omega') dP'(\omega') = P'(A') = P(T^{-1}(A'));$$

    In addition,

$$\int \mathbb{1}_{A'}(T(\omega)) dP(\omega) = \int \mathbb{1}_{T^{-1}(A')}(\omega) dP(\omega) = P(T^{-1}(A')).$$

    Now, suppose X is simple. This is true by linearity of expectations.
    Now, suppose $X \geq 0$. Create $X_n \geq 0$ that are simple and measurable, such that $X_n \uparrow X$. Then, $X_n \circ T \uparrow X \circ T$. Thus,

$$\int X dP' = \lim_{n\to\infty} \int X_n dP' = \lim_{n\to\infty} \int (X_n \circ T) = \int (X \circ T) dP.$$

    Now, suppose X is general and measurable. $(X \circ T)^+ = (X^+ \circ T)$, and $(X \circ T)^- = (X^- \circ T)$. Therefore,

$$\int X dP' = \int X^+ dP' - \int X^- dP' = \int (X^+ \circ T) dP - \int (X^- \circ T) dP = \int (X' \circ T) dP. \blacksquare$$

- **Uniform Integrability**, or **U.I.**: $\{X_n\}$ is uniformly integrable (u.i.) wit respect to finite measure P if $\sup_{n\geq 1} \int_{|X_n|>C} |X_n| dP \to 0$ as $C \to \infty$.

  - MCT has a very restrictive setting, and DCT needs very strong conditions. Uniform integrability will provide a weaker condition for convergence of expectations.
  - We require a finite measure.
  - Facts about u.i.:
    - $\{X_n\}$ is u.i. iff $\{|X_n|\}$ is u.i.
    - Any finite collection of integrable functions is u.i.
    - If $\{X_n\}$ and $\{Y_n\}$ are u.i., then so is $\{X_n, Y_n : n \geq 1\}$.
    - If $\{X_n\}$ is u.i., then so is $\{cX_n\}$.
    - If $|X_n| \leq |Y|$ for integrable Y, then $\sup_{n\geq 1} \int_{|X_n|>C} |X_n| dP \leq \int_{|Y|>C} |Y| dP \to 0$.
      - This will let us impose a weaker condition for convergence of expectations!
    - If $|X_n| \leq |Y_n|$ for u.i. $\{Y_n\}$, then so is $\{X_n\}$.
    - If X_n and Y_n are u.i., then so is $X_n + Y_n$.
      - *Proof*: Since X_n and Y_n are u.i., then $\exists \delta_i : P(A) < \delta_1 \implies \int_A |X_n| dP < \frac{\epsilon}{2}$, and $P(A) < \delta_2 \implies \int_A |Y_n| dP < \frac{\epsilon}{2}$.

$$\mathbb{E}|X_n + Y_n| \leq \mathbb{E}|X_n| + \mathbb{E}|Y_n| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \blacksquare$$

  - $\{X_n\}$ is u.i. iff $\sup_n \mathbb{E}|X_n| < \infty$ and for any $\epsilon > 0$, $\exists \delta > 0 : \forall n$, $P(B) < \delta \implies \int_B |X_n| dP < \epsilon$.
    - *Proof*: First, suppose X_n is u.i. $\int |X_n| dP = \int_{|X_n|\leq C} |X_n| dP + \int_{|X_n|>C} |X_n| dP$. Note that P is a finite measure. By the definition of u.i., $\int_{|X_n|>C} |X_n| dP < \epsilon \forall \epsilon > 0$. ϵ is arbitrary, so we choose $\epsilon = 1$. With this in mind,

$$\int_{|X_n|\leq C} |X_n| dP + \int_{|X_n|>C} |X_n| dP \leq CP(|X_n| \leq C) + 1 \leq CP(\Omega) + 1 < \infty.$$

Now, let B be any set.

$$\int_B |X_n| dP = \int_{B\cap\{|X_n|>C\}} |X_n| dP + \int_{B\cap\{|X_n|\leq C\}} |X_n| dP \leq \frac{\epsilon}{2} + CP(B) < \epsilon$$

whenever $P(B) < \frac{\epsilon}{2C} = \delta$.
Now, suppose $\sup_n \mathbb{E}|X_n| < \infty$ and for any $\epsilon > 0$, $\exists \delta > 0 : \forall n$, $P(B) < \delta \implies \int_B |X_n| dP < \epsilon$.
Let $M = \sup_n \int |X_n| dP < \infty$, and $\epsilon > 0$. We need to find a δ to establish u.i.
By Markov's inequality, $P(|X_n| > C) \leq \frac{\mathbb{E}|X|}{C} \leq \frac{M}{C} < \delta$ if $C > \frac{M}{\delta}$.
Now, let $A = \{|X_n| > C\}$. This means that $P(A) < \delta$. Thus, $\int_A |X_n| dP < \epsilon$. Since n is arbitrary, we have satisfied the definition of u.i. ■

- If $X_n \to X$ a.s. and $\{X_n\}$ are uniformly integrable, then $\mathbb{E}(X_n) \to \mathbb{E}(X)$.
  - As a result, $|\mathbb{E}(X_n) - \mathbb{E}(X)| \leq \mathbb{E}|X_n - X| \to 0$.
  - *Proof*: Since X_n is u.i., $\sup_n \mathbb{E}|X_n| < \infty$. So, using Fatou's lemma,

$$\mathbb{E}|X| \leq \liminf_{n\to\infty} \mathbb{E}|X_n| \leq \sup_n \mathbb{E}|X_n| < \infty.$$

    Now, let $Y_n = X$. Since X is integrable, Y_n is u.i. In addition, X_n is u.i., so $X_n - Y_n = X_n - X$ is also u.i. So, for any $\epsilon > 0$,

$$\mathbb{E}|X_n - X| = \int_{|X_n - X|\leq\epsilon} |X_n - X| dP + \int_{|X_n-X|>\epsilon} |X_n - X| dP \leq \epsilon + \int_{|X_n-X|>\epsilon} |X_n - X| dP.$$

    Now, consider $A_n = \{\omega : |X_n(\omega) - X(\omega)| > \epsilon\}$. $P(A_n) \to 0$ by u.i. of $X_n - X$.
    This means that $\int_{|X_n-X|>\epsilon} |X_n - X| dP < \epsilon \implies \mathbb{E}|X_n - X| < 2\epsilon$. Since ϵ is arbitrary, this means that $\mathbb{E}(X_n) \to \mathbb{E}(X)$. ■
  - We don't need the X_n's to be defined on the same sample space to satisfy this condition, whereas DCT does.
- $\sup_n \mathbb{E}|X_n| < \infty$ is a necessary but not sufficient condition for convergence.
  - **Example**: $\Omega = [0,1]$, $P = \lambda$, $X_n(\omega) = n\mathbb{1}_{(1-1/n,1)}(\omega)$.

$$\mathbb{E}(X_n) = 0\left(1 - \frac{1}{n}\right) + n \cdot \frac{1}{n} = 1 \implies \mathbb{E}(X_n) = 1.$$

    However, for any arbitrary $C > 0$,

$$\{\omega : |X_n(\omega)| > C\} = \begin{cases} \varnothing & , n \leq C \\ \left(1 - \frac{1}{n}, 1\right] & , n > C. \end{cases}$$

    This means that $\int_{|X_n|>C} |X_n| dP = 1 \not\to 0$ as $C \to \infty$ for $n > C$, so X_n is not u.i. ■
- Let $\psi : [0,\infty) \to [0,\infty)$ be defined such that $\frac{\psi(x)}{x} \uparrow \infty$ as $x \to \infty$. If $\sup_n \mathbb{E}[\psi(|X_n|)] < \infty$, then X_n is u.i.
  - *Proof*: Note that $\frac{\psi(x)}{x} \uparrow \infty \implies \frac{x}{\psi(x)}$ is decreasing. Let $C > 0$.

$$\int_{|X_n|>C} |X_n| dP = \int_{|X_n|>C} \frac{|X_n|}{\psi(|X_n|)} \cdot \psi(|X_n|) dP \leq \frac{C}{\psi(C)} \mathbb{E}[\psi(|X_n|)] \to 0. \blacksquare$$

  - X_n is u.i. if $\sup_n \mathbb{E}(X_n^2) < \infty$.
    - **Example**: Suppose $X_n = n\mathbb{1}\left(c_n, c_n + \frac{1}{n^2}\right)$, where c_n is dense in $[0,1]$. Note that X_n's are not dominated in this example, so we'll never be able to use DCT
      $\mathbb{E}(X_n^2) = n^2 \cdot \frac{1}{n^2} = 1 < \infty$, which is bounded in n, so X_n is u.i. ■
- **Example**: Suppose $\{X_n\}$ is iid, and integrable. Show that $\{n^{-1}S_n, n \geq 1\}$ is u.i.

  First, we show that X_i are u.i. Since X_i are iid and integrable,

$$\mathbb{E}[|X_i|\mathbb{1}\{X_i > C\}] = \mathbb{E}[|X_1|\mathbb{1}\{X_1 > C\}] < \infty \to 0 \implies \sup_{n\geq 1} \mathbb{E}[|X_i|\mathbb{1}\{X_i > C\}] \to 0.$$

  Thus, X_i are u.i. Since this is the case, $\forall i$, $\sup_n \mathbb{E}|X_n| < \infty$.

$$\sup_n \mathbb{E}|n^{-1}S_n| = \sup_n \frac{1}{n}\mathbb{E}\left|\sum_{i=1}^n X_i\right| \overset{\text{iid}}{\leq} \sup_n |\mathbb{E}(X_1)| < \infty.$$

Thus, $n^{-1}S_n$ is integrable. Suppose $\forall \epsilon > 0$, $\exists \delta > 0 : P(A) < \delta \Rightarrow \sup_i \int |X_i| dP < \epsilon$. For the same δ,

$$\sup_n \int_A \left| \frac{S_n}{n} \right| dP = \sup_n \frac{1}{n} \sum_{i=1}^{n} \int_A |X_i| dP < \sup_n \frac{1}{n} \sum_{i=1}^{n} \epsilon = \frac{\epsilon n}{n} = \epsilon.$$

Since $\epsilon > 0$ is arbitrary, $\sup_n \int_A \left| \frac{S_n}{n} \right| dP \to 0 \implies n^{-1}S_n$ is u.i. ■

- **Example**: Let X_n, Y_n, X, and Y be RVs on probability space $(\Omega, \mathcal{A}, P)$ satisfying $|X_n| \leq Y_n$ for all n, where $X_n \to X$, $Y_n \to Y$, and $\mathbb{E}(Y_n) \to \mathbb{E}(Y) < \infty$. Show that $\mathbb{E}(X_n) \to \mathbb{E}(X)$.

  Since X_n is positive, then we can apply Fatou's lemma to get that $\mathbb{E}(X) = \mathbb{E}(\liminf_{n\to\infty} X_n) \leq \liminf_{n\to\infty} \mathbb{E}(X_n)$.

  Next, $0 \leq X_n \leq Y_n$, then $X_n - Y_n \leq 0$. Since $0 > -\infty$, we can once again apply Fatou's lemma to get that $\mathbb{E}\left(\limsup_{n\to\infty}(X_n - Y_n)\right) \geq \limsup_{n\to\infty} \mathbb{E}(X_n - Y_n)$. Since $Y_n \to Y$ and $X_n \to X$,

$$\mathbb{E}\left(\limsup_{n\to\infty}(X_n - Y_n)\right) = \mathbb{E}(\limsup_{n\to\infty} X_n) - \mathbb{E}(\limsup_{n\to\infty} Y_n) = \mathbb{E}(X) - \mathbb{E}(Y).$$

  Applying the same result to the other side of the inequality, $\limsup_{n\to\infty} \mathbb{E}(X_n - Y_n) = \limsup_{n\to\infty} \mathbb{E}(X_n) - \mathbb{E}(Y)$. Therefore,

$$\mathbb{E}(X) - \mathbb{E}(Y) \geq \limsup_{n\to\infty} \mathbb{E}(X_n) - \mathbb{E}(Y) \implies \limsup_{n\to\infty} \mathbb{E}(X_n) \leq \mathbb{E}(X).$$

  Combining with the original result,

$$\mathbb{E}(X) \leq \liminf_{n\to\infty} \mathbb{E}(X_n) \leq \limsup_{n\to\infty} \mathbb{E}(X_n) \leq \mathbb{E}(X) \implies \mathbb{E}(X_n) \to \mathbb{E}(X). \blacksquare$$

- **Example**:

  1. Show that two RVs X_1 and X_2 are independent iff for every pair f_1, f_2 of non-negative continuous functions, $\mathbb{E}[f_1(X_1)f_2(X_2)] = \mathbb{E}[f_1(X_1)]\mathbb{E}[f_2(X_2)]$.
  2. Suppose that for each n, the pair ξ_n and η_n are independent RVs and that pointwise, $\xi_n \to \xi_\infty$, and $\eta_n \to \eta_\infty$. Show that ξ_∞ and η_∞ are independent.

  1. – **Forward Direction**: Suppose X_1 and X_2 are independent. This must mean that $f_1(X_1)$ and $f_2(X_2)$ are independent, so $\mathbb{E}[f_1(X_1)]\,\mathbb{E}[f_2(X_2)] = \mathbb{E}[f_1(X_1)f_2(X_2)]$.
     – **Reverse Direction**: Suppose $\mathbb{E}[f_1(X_1)]\,\mathbb{E}[f_2(X_2)] = \mathbb{E}[f_1(X_1)f_2(X_2)]$ for all bounded and continuous f_1 and f_2.
     Let the expectation be defined over probability measure P, and let $g_i = \mathbb{1}_{(a_i,b_i]}$. By the previous result, we can construct a sequence of nonnegative, bounded, and continuous $g_{in} \in \left(a, b + \frac{1}{n}\right]$ that results in a bound of $g_i \in [0, g_{in}]$. $a \leq g_{in} \leq b + \frac{1}{n} \implies \mathbb{E}(g_{in}) < \infty$, so we can apply the DCT to get that $\mathbb{E}[g_1(X_1)g_2(X_2)] = \mathbb{E}[g_1(X_1)]\,\mathbb{E}[g_2(X_2)]$, which is equivalent to saying that

$$P(X_1 \in (a_1, b_1], X_2 \in (a_2, b_2]) = P(X_1 \in (a_1, b_1])\, P(X_2 \in (a_2, b_2]),$$

     which is the definition of independence.
  2. We can use the previous result, where $f_{1n}(\eta_n) = \eta_n$, and $f_{2n}(\xi_n) = \xi_n$.

$$\mathbb{E}[\xi_\infty \eta_\infty] = \lim_{n\to\infty} \mathbb{E}[\xi_n \eta_n] = \lim_{n\to\infty} \mathbb{E}[\xi_n]\,\mathbb{E}[\eta_n] = \mathbb{E}[\xi_\infty]\,\mathbb{E}[\eta_\infty].$$

     This means that ξ_∞ and η_∞ are independent. ■

9.5 Inequalities and L_p-Spaces

Return to Table of Contents

- **Cauchy-Schwarz Inequality**: Let $(\Omega, \mathcal{A}, P)$ be a probability space, and $X, Y : \Omega \to \mathbb{R}$ be RVs. If $\mathbb{E}(X^2)$ and $\mathbb{E}(Y^2) < \infty$, then XY is integrable, and

$$|\mathbb{E}(XY)| \leq \mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)},$$

  with equality iff X and Y are proportional a.s.

– *Proof*: First, we show XY is integrable. Using the fact that

$$(X+Y)^2 = X^2 + 2XY + Y^2 \implies |XY| \leq \frac{X^2+Y^2}{2},$$

so given finite second moments, XY must be integrable.

$$(X+tY)^2 \geq 0 \implies \mathbb{E}(X^2) + t^2\mathbb{E}(Y^2) + 2t\mathbb{E}(XY) \geq 0 \ \forall t \in \mathbb{R}.$$

Choose $t = -\frac{\sqrt{\mathbb{E}(X^2)}}{\sqrt{\mathbb{E}(Y^2)}}$. Then,

$$\mathbb{E}(X^2) + \mathbb{E}(X^2) - 2\frac{\sqrt{\mathbb{E}(X^2)}}{\sqrt{\mathbb{E}(Y^2)}}\mathbb{E}(XY) \geq 0 \implies |\mathbb{E}(XY)| \leq \mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)}.$$

Lastly, the equality holds when $X \propto Y$ by applying trivial algebra. ■

– **Example**: Let $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ be two permutations of $\{1, \ldots, n\}$. Show that $\sum_{i=1}^n a_i b_i \leq \frac{n(n+1)(n+2)}{6}$, with equality iff $a_i = b_i \ \forall i$.

Define $X = \begin{pmatrix} a_1 & \ldots & a_n \\ \frac{1}{n} & \ldots & \frac{1}{n} \end{pmatrix}$, similarly for Y.

$$\mathbb{E}(X) = \mathbb{E}(Y) = \sum_{i=1}^n \frac{i}{n} = \frac{n+1}{2}, \text{ and } \mathbb{E}(X^2) = \mathbb{E}(Y^2) = \sum_{i=1}^n \frac{i^2}{n} = \frac{(n+1)(2n+1)}{6}.$$

Applying the Cauchy-Schwarz inequality,

$$\frac{1}{n}\sum_{i=1}^n a_i b_i = \mathbb{E}(XY) \leq \sqrt{\frac{(n+1)(2n+1)}{6}}\sqrt{\frac{(n+1)(2n+1)}{6}},$$

where rearranging yields the result. We know that the Cauchy-Schwarz inequality is an equality iff $X \propto Y$, which happens when $a_i = b_i$ for all i. ■

- **Hölder's Inequality**: For RVs X and Y, where $\mathbb{E}|X|^p$ and $\mathbb{E}|Y|^q < \infty$ for $p, q \geq 1 : \frac{1}{p} + \frac{1}{q} = 1$, then

$$|\mathbb{E}(XY)| \leq \left[\mathbb{E}|X|^p\right]^{1/p} \left[\mathbb{E}|Y|^q\right]^{1/q}, \text{ with equality iff } X \propto Y.$$

– $p = q = 2$ yields the Cauchy-Schwarz inequality.

– *Proof*: First, we prove a lemma that for conjugate indices p and q, and for positive constants a and b, $\frac{a^p}{p} + \frac{b^q}{q} \geq ab$.

$$\log\left\{\frac{a^p}{p} + \frac{b^q}{q}\right\} \geq \frac{\log(a^p)}{p} + \frac{\log(b^q)}{q} = \log a + \log b = \log(ab).$$

Applying the concavity of logs proves the lemma.
Now, let $a = \frac{|X|}{\mathbb{E}(|X|^p)^{1/p}}$ and $b = \frac{|Y|}{\mathbb{E}(|Y|^q)^{1/q}}$.

$$\frac{|X||Y|}{\mathbb{E}(|X|^p)^{1/p}\mathbb{E}(|Y|^q)^{1/q}} \leq \frac{|X|^p}{p\mathbb{E}(|X|^p)} + \frac{|Y|^q}{q\mathbb{E}(|Y|^q)}.$$

Integrating and canceling terms yields

$$\frac{\mathbb{E}|XY|}{\mathbb{E}(|X|^p)^{1/p}\mathbb{E}(|Y|^q)^{1/q}} \leq \frac{1}{p} + \frac{1}{q} = 1 \implies |\mathbb{E}(XY)| \leq \left[\mathbb{E}|X|^p\right]^{1/p} \left[\mathbb{E}|Y|^q\right]^{1/q}.$$

Once again, due to algebra the corollary is an equality iff $a^p = b^q \equiv X \propto Y$. ■

- **Markov/Chebyshev Inequality**: For increasing $\psi : [0, \infty) \to [0, \infty)$,

$$P(|X| \geq t) \leq \frac{1}{\psi(t)}\mathbb{E}\left[\psi(|X|)\right].$$

– This inequality is only useful for finite $\mathbb{E}[\psi(|X|)]$.

– *Proof*: Since ψ is increasing,

$$\mathbb{1}\{x \geq t\} \leq \frac{\psi(x)}{\psi(t)} \implies \mathbb{1}\{|X(\omega)| \geq t\} \leq \frac{\psi(|X(\omega)|)}{\psi(t)} \implies P(|X| \geq t) \leq \frac{1}{\psi(t)}\mathbb{E}\left[\psi(|X|)\right]$$

by taking the expectation on both sides. ■

– We can improve this inequality via the inequality given in the proof.

$$P(|X| \geq t) \leq \frac{1}{\psi(t)} \int_{|X| \geq t} \psi(|X|) dP.$$

∗ This is a tighter bound, since $\int_{|X|\geq t} \psi(|X|)dP \leq \mathbb{E}\left[\psi(|X|)\right]$.

∗ Without the improvement, the tail probability is $O\left(\frac{1}{\psi(t)}\right)$, and with the improvement, the tail probability is $o\left(\frac{1}{\psi(t)}\right)$.

- **Minkowski's Inequality**: If $\mathbb{E}(|X|^p)$ and $\mathbb{E}(|Y|^p) < \infty$, where $p \geq 1$, then

$$\left[\mathbb{E}(|X+Y|^p)\right]^{1/p} \leq \left[\mathbb{E}(|X|^p)\right]^{1/p} + \left[\mathbb{E}(|Y|^p)\right]^{1/p}.$$

– *Proof*: First, we need to prove a lemma that for any $p > 0$, $(a+b)^p \leq \max\{2^{p-1}, 1\}(a^p + b^p)$.
For $p \geq 1$, we need to show that $(a+b)^p \leq 2^{p-1}(a^p+b^p)$. By the convexity of $x \mapsto x^p$ on $[0,\infty)$, $\left(\frac{a+b}{2}\right)^p \leq \frac{1}{2}a^p + \frac{1}{2}b^p$, which can be rewritten to yield the answer.
For $p < 1$, define $m = \frac{1}{p} > 1$, $x = a^p$, and $y = b^p$. Then, we can once again use convexity to get that

$$(x+y)^m \geq x^m + y^m \implies \frac{x^m}{(x+y)^m} + \frac{y^m}{(x+y)^m} \leq \left(\frac{x}{x+y} + \frac{y}{x+y}\right) = 1,$$

which proves the lemma.
Minkowski's inequality is trivial if $p = 1$. Therefore, we now consider the case where $p > 1$. Let $A = \left[\mathbb{E}(|X+Y|^p)\right]^{1/p}$, $B = \left[\mathbb{E}(|X|^p)\right]^{1/p}$, and $C = \left[\mathbb{E}(|Y|^p)\right]^{1/p}$. Let q be the conjugate index of p. Note that if B and $C < \infty$, then so must $A < \infty$.
Now, $|X+Y|^p \leq |X| \cdot |X+Y|^{p-1} + |Y| \cdot |X+Y|^{p-1}$. Integrate and apply Hölder's inequality to obtain, which is

$$\begin{aligned}
\mathbb{E}(|X+Y|^p) \leq& \mathbb{E}\left[|X| \cdot |X+Y|^{p-1}\right] + \mathbb{E}\left[|Y| \cdot |X+Y|^{p-1}\right] \\
\leq& [\mathbb{E}(|X|)^p]^{1/p} \left[\mathbb{E}(|X+Y|^{q(p-1)}\right]^{1/q} + [\mathbb{E}(|X|)^p]^{1/p} \left[\mathbb{E}(|X+Y|^{q(p-1)}\right]^{1/q} \\
\implies A^p \leq& BA^{p/q} + CA^{p/q} \implies A \leq B + C. \blacksquare
\end{aligned}$$

– **Example**: Suppose $p = \frac{1}{2}$, $\Omega = \{0,1\}$, $P = (\frac{1}{2}, \frac{1}{2})$, $X = (1,0)$, $Y = (0,1)$.
$X + Y = 1$, so $\left[\mathbb{E}(|X+Y|^{1/2})\right]^2 = 1^2 = 1$. However, $\left[\mathbb{E}(|X|^{1/2})\right]^2 = \left[\mathbb{E}(|Y|^{1/2})\right]^2 = \frac{1}{4}$. This is an example of why Minkowski does not hold for $p < 1$. ■

- **Moment Inequality**: If X is an RV and $0 < r < s \leq \infty$, then

$$\alpha_r := \mathbb{E}(|X|^r) \leq \left[\mathbb{E}(|X|^s)\right]^{r/s} = \alpha_s^{r/s},$$

where α_r is called the rth absolute moment.

– Define $\|X\|_r := \alpha_r^{1/r}$.

– **Essential Supremum**, or **esssup**: $\|\cdot\|_\infty$.

– *Proof*: Note that $\frac{s}{r}$ and $\frac{s}{s-r}$ are conjugate indices. Applying Hölder's inequality,

$$\alpha_r = \mathbb{E}(|X|^r) \leq \left[\mathbb{E}(|X|^s)\right]^{r/s}(1) = \left\{\left[\mathbb{E}(|X|^r)\right]^{s/r}\right\}^{r/s} \left\{\mathbb{E}(1^{s/(s-r)})\right\} = \alpha_s^{r/s}. \blacksquare$$

- **Jensen's Inequality**: If X is an RV on $(\Omega, \mathcal{A}, P)$, and $\psi : \mathbb{R} \to \mathbb{R}$ is convex, then

$$\psi\left[\mathbb{E}(X)\right] \leq \mathbb{E}\left[\psi(X)\right].$$

– **Convex**: $\psi : \mathbb{R} \to \mathbb{R}$ such that for x and y, and $0 \leq \alpha \leq 1$,

$$\psi\left(\alpha x + (1-\alpha)y\right) \leq \alpha\psi(x) + (1-\alpha)\psi(y).$$

- *Proof*: A convex function envelops the tangent lines

$$\psi(x) = \sup_{a,b}\{a + bx : a + by \leq \psi(y)\ \forall y\}.$$

As a result, $\mathbb{E}[\psi(X)] \geq \mathbb{E}(a + bX) = a + b\mathbb{E}(X)$. Taking the supremum over a and b yields $a + by \leq \psi(y) = \psi(\mathbb{E}(X))$ for all y. ■

- If ψ is concave, the direction of the inequality flips.
- **Example**: Moment inequality. $\alpha_r = \mathbb{E}(|X|^r)$, and $\alpha_s = \mathbb{E}(|X|^s)$.

$$|X|^s = (|X|^r)^{s/r} \implies \mathbb{E}(|X|^s) \geq [\mathbb{E}(|X|^r)]^{s/r}. \blacksquare$$

- **Arithmetic-Geometric-Harmonic Inequality**: If $a_1, \ldots, a_n > 0$, then

$$\frac{1}{n}\sum_{i=1}^{n} a_i \geq \left(\prod_{i=1}^{n} a_i\right)^{1/n} \geq \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{a_i}}.$$

- *Proof*: Use convexity of logs. ■

- $\boldsymbol{\mathcal{L}_p}$: $\{X : \mathbb{E}(|X|^p) < \infty\}$.
  - $\mathcal{L}_p$ is a linear space. That is, $\mathcal{L}_p$ is closed under addition and scalar multiplication.
  - If $X, Y \in \mathcal{L}_p$, then $\mathbb{E}(|X + Y|^p) \leq \max\{2^{p-1}, 1\}[\mathbb{E}(|X|^p) + \mathbb{E}(|Y|^p)]$.
  - Define $\|X\|_p := [\mathbb{E}(|X|^p)]^{1/p} \geq 0$.
    - This is similar to a norm on a linear space. however $\|X\| + p = 0 \equiv X = 0$ only almost surely, not always.
    - $\|cX\|_p = |c| \cdot \|X\|_p$.
    - If $p \geq 1$, then by Minkowski, $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$.

- Call the space of such equivalent classes of functions L_p, or $L_p(P)$, or $L_p(\Omega, \mathcal{A}, P)$.
  - $\|\cdot\|_p$ is a norm on L_p if $p \geq 1$. Then, L_p becomes a normed linear space.
  - $\boldsymbol{L_p}$**-Distance**: $d_p(X, Y) = \|X - Y\|_p$.
  - $\|\cdot\|_p$ is not a norm for $p < 1$, but we can define $d_p(X, Y) = \mathbb{E}(|X - Y|^p)$.
    - **Triangle Inequality**: $d_p(X, Z) \leq d_p(X, Y) + d_p(Y, Z)$.
  - $\|X\|_\infty = \int\{c \geq 0 : |X| \leq c \text{ a.s.}\} := \text{esssup}|X|$.
    - $\mathcal{L}_\infty = \{X : X \text{ measurable}, \|X\|_\infty < \infty\}$ with corresponding equivalence class L_∞.
      - $\|\cdot\|_\infty$ is a norm on L_∞.
  - $L_{p_1} \supset L_{p_2}$ if $p_1 < p_2$ by the moment inequality.

- **Cauchy Sequence**: $x_n \to x \implies |x_m - x_n| \to 0$ as $m, n \to \infty$ for large m, n.
  - Convergent implies Cauchy, but not necessarily the other way.
    - **Example**: Suppose we have $S \in \mathbb{Q}$. $(1, 1.4, 1.414, \ldots) \to \sqrt{2}$, and every element is in $\mathbb{Q}$, but $\sqrt{2} \notin \mathbb{Q}$. ■

- $X_n \in L_p$ converges to $X \in L_p$ if $d_p(X_n, X) \to 0$.

- **Complete**: A metric space (S, d) such that every Cauchy sequence is convergent.
  - $\mathbb{R}$ and all L_p-spaces are complete.
  - All step functions with rational values and intervals with rational endpoints are dense in L_p for $p < \infty$.
  - **Separable**: $\forall s \in S$ and $\epsilon > 0$, $\exists x \in D : d(s, x) < \epsilon$, where D is the set of countable and dense sets.
    - L_p-spaces are separable.

- $\ell_p = \{\{a_n\} : \sum_{i=1}^{\infty} |a_i|^p < \infty\}$ is an analog of L_p for $\mathbb{N}$ with a counting measure on it.
  - $\ell_\infty = \{\{x_n\} : \{x_n\} \text{ is a bounded sequence}\}$.
  - ℓ_p distance between $\{a_n\}$ and $\{b_n\}$ is:
    - $\{\sum_{i=1}^{\infty} |a_n - b_n|^p\}^{1/p}$ if $1 \leq p < \infty$.

    - $\{\sum_{i=1}^{\infty} |a_n - b_n|^p\}$ if $0 < p < 1$.
    - $\sup\{|a_n - b_n| : n \in \mathbb{N}\}$ if $p = \infty$.
  - All ℓ_p spaces are complete, ℓ_p is separable for $p < \infty$.
  - $\ell_{p_1} \subset \ell_{p_2}$ if $p_1 < p_2$.
    - **Example**: $\sum |x_n| < \infty \implies \sum |x_n|^2 < \infty$.

- **Example**: $(\mathbb{R}, \mathcal{R}, \lambda)$. $\mathcal{L}_p = \left\{f : \int |f|^p d\lambda < \infty\right\}$, and $L_p = \|f\|_p = \left(\int |f|^p d\lambda\right)^{1/p}$ for $p \geq 1$. In addition, $\|f\|_\infty = \text{esssup}|f|$.
  We can't compare L_{p_1} to L_{p_2}. For instance, take $L_1 = \left(\int |f| d\lambda < \infty\right)$ and $L_2 = \left(\int |f|^2 d\lambda < \infty\right)$. Now, take $f(x) = \frac{1}{|x|}\mathbb{1}\{|x| \geq 1\}$. $f \notin L_1$, but $f \in L_2$. Now, consider $f(x) = \frac{1}{\sqrt{|x|}}\mathbb{1}(|x| \leq 1)$. $f \in L_1$, but $f \notin L_2$. ■

9.6 Joint Random Variables

Return to Table of Contents

- **Measurable Rectangle**: $A_1 \times A_2$, where given two measurable spaces $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$, and $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$.
  - The class of measurable rectangles forms a semifield.
  - Given respective probability measures P_1 and P_2, the measure on the measurable rectangle is $P_1(A_1)P_2(A_2)$.
    - This is countably additive on the semifield.
  - **Product Measure**: $P_1 \times P_2$.
    - Is unique by Caratheodory.
    - Stands for the joint distribution of the pair of independent RVs.
- **Product σ-Field**: The σ-field $\mathcal{A}$ generated by the collection of measurable rectangles $\mathcal{A}_1 \otimes \mathcal{A}_2$.
- **Transition**: $P_2(\omega, A_2)$ on $(\Omega_2, \mathcal{A}_2)$, where P_1 is a probability measure on $(\Omega_1, \mathcal{A}_1)$ such that:
  1. $\forall A_2 \in \mathcal{A}_2$, $\omega_1 \mapsto P_2(\omega_1, A_2)$ is $\mathcal{A}_1$-measurable.
  2. $\forall \omega_1 \in \Omega_1$, $A_2 \mapsto P_2(\omega_1, A_2)$ is a probability measure.
     - In other words, if ω_1 is fixed, A_2 is a probability measure.
  - Let $P(E) = \int P_2(\omega_1, E_{\omega_1} P(d\omega_1)$. By the good sets principle, if $E \subset \Omega_1 \times \Omega_2$ is measurable, then $E_{\omega_1} = \{\omega_2 \in \Omega_2 : (\omega_1, \omega_2) \in E\}$.
  - P satisfies countable additivity on the product σ-field.
  - If $E_1, E_2, \cdots \in \mathcal{A}$ are disjoint, then

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \int P_2\left(\omega_1, \left(\bigcup_{n=1}^{\infty} E_n\right)_{\omega_1}\right) P_1(d\omega_1) = \int P_2\left(\omega_1, \bigcup_{n=1}^{\infty} (E_n)_{\omega_1}\right) P_1(d\omega_1)$$
$$= \int \sum_{n=1}^{\infty} P_2\left(\omega_1, (E_n)_{\omega_1}\right) P_1(d\omega_1) \overset{\text{MCT}}{=} \sum_{n=1}^{\infty} \int P_2\left(\omega_1, (E_n)_{\omega_1}\right) P_1(d\omega_1) = \sum_{n=1}^{\infty} P(E_n).$$

- **Example**: Show that the product σ-algebra is the smallest σ-algebra making the coordinate mappings π_1, π_2 measurable.

  Proceed with bootstrapping. Let $\mathcal{G}$ be a σ-algebra on the cartesian product of two random variables $X_1 \times X_2$ such that coordinate projections are measurable with respect to $\mathcal{G}$. That is,

$$\pi_1^{-1}(A_1) = A_1 \times X_2 \in \mathcal{G}, \text{ and } \pi_2^{-1}(A_2) = X_1 \times A_2 \in \mathcal{G}.$$

  Next, define $\mathcal{G}_2 = \{A_1 \in \mathcal{A}_1 : A_1 \times S_2 \in \mathcal{G}\}$. We claim that $\mathcal{G}_2$ is a σ-algebra on X_1. First, since $X_1 \times A_2 = \pi_2^{-1}(A_2) \in \mathcal{G}$, $X_1 \in \mathcal{G}_2$. Next, we show that $\mathcal{G}_2$ is closed under complementation. Let $A_1 \in \mathcal{A}_1$, $A_2 \in \mathcal{A}_2$ such that $A_1 \times A_2 \in \mathcal{G}$.

$$A_1^c \times A_2 = \underbrace{(X_1 \times A_2)}_{\in \mathcal{G}} \backslash (A_2 \times A_1) \in \mathcal{G}.$$

  Therefore, $\mathcal{G}_2$ is closed under complementation. Lastly, we show that $\mathcal{G}_2$ is closed under countable union. Let $\{A_{1i}\}_{i=1}^{\infty} \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$ such that $A_{1i} \times A_2 \in \mathcal{G}$.

$$\left(\bigcup_{n=1}^{\infty} A_{1n}\right) \times A_2 = \bigcup_{n=1}^{\infty} \underbrace{(A_{1n} \times A_2)}_{\in \mathcal{G}} \in \mathcal{G}.$$

  Therefore, $\mathcal{G}_2$ is a σ-algebra on X_1. Since $A_1 \times A_2 \in \mathcal{G}$ by the measurability of π_1 and π_2, $A_1 \in \mathcal{A}_1 \implies A_1 \in \mathcal{G}_2$. Therefore, combining all of the previous results in this problem, bootstrapping gives us that $\mathcal{G}_2 = \mathcal{A}_1$.
  As a result, since $A_1 \times A_2$ generates the product σ-algebra $\mathcal{A}_1 \otimes \mathcal{A}_2$, then $\mathcal{A}_1 \otimes \mathcal{A}_2 \subset \mathcal{G}$.
  In conclusion, the product σ-field is the smallest σ-algebra making the coordinate maps measurable, since if a σ-algebra on $X_1 \times X_2$ makes π_1 and π_2 measurable, then it must contain every generating rectangle $A_1 \times A_2$, which means it contains the entire σ-algebra $\mathcal{A}_1 \otimes \mathcal{A}_2$. ■

- **Jointly Measurable**: $X : (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2) \to (\bar{\mathbb{R}}, \bar{\mathcal{R}})$ if $\forall B \in \bar{\mathcal{R}}$,

$$\underbrace{\{(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 : X(\omega_1, \omega_2) \in B\}}_{\text{Inverse image}} \in \mathcal{A}_1 \otimes \mathcal{A}_2.$$

  - $X(\omega_1, \omega_2) = X_1(\omega_1)X_2(\omega_2)$ is jointly measurable if X_1 and X_2 are measurable on Ω_1 and Ω_2, respectively.
  - $\sum_{i=1}^{\infty} c_i X_{1i}(\omega_1)X_{2i}(\omega_2)$ is jointly measurable.
  - For Euclidean spaces and Borel measurability, then joint continuity is sufficient.

- On Euclidean spaces, a sufficient condition for joint measurability is continuity in one component and measurability in the other.

  - *Proof*: Assume $\Omega_1 \subset \mathbb{R}$ is compact (otherwise, the function is a pointwise limit of such functions). Partition Ω_1 into 2^n smaller sub-intervals $J_1, \ldots, J_N$, and choose representative values $a_i \in J_i$. Then,

$$\sup_{x \in J_i} |f(x, y) - f(a_i, y)| \to 0 \ \forall y.$$

    Since $f(a_i, y)\mathbb{1}\{x \in J_i\}$ is jointly measurable, so $\sum_{i=1}^{n} f(a_i, y)\mathbb{1}\{x \in J_i\}$ is also jointly measurable. That means that

$$f(x, y)) = \lim_{N \to \infty} \sum_{i=1}^{n} f(a_i, y)\mathbb{1}\{x \in J_i\} \text{ is also measurable. } \blacksquare$$

- We want to determine when it is possible to evaluate joint expectations iteratively by integrating out random variables one at a time.

- **Fubini's Theorem**: Consider $P = P_1 \times P_2$, where P_1 is a probability measure, and P_2 is a transition. Let $X : (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2) \to (\bar{\mathbb{R}}, \bar{\mathcal{R}})$ be jointly measurable.

  1. If $X \geq 0$ a.s., then $\int_{\Omega_2} X(\omega_1, \omega_2)P_2(\omega_1, d\omega_2)$ exists and is nonnegative a.s., and $\omega_1 \mapsto \int_{\Omega_2} X(\omega_1, \omega_2)P_2(\omega_1, d\omega_2)$ is measurable. In addition,

$$\int_{\Omega} X dP = \int_{\Omega_1} \left[\int_{\Omega_2} X(\omega_1, \omega_2)P_2(\omega_1, d\omega_2) \right] P_1(d\omega_1)$$

  2. If $\int_{\Omega} |X| dP < \infty$, then $\int_{\Omega_2} X(\omega_1, \omega_2)P_2(\omega_1, d\omega_2)$ is finite a.s., and $\omega_1 \mapsto \int_{\Omega_2} X(\omega_1, \omega_2)P_2(\omega_1, d\omega_2)$ is measurable. In addition,

$$\int_{\Omega} X dP = \int_{\Omega_1} \left[\int_{\Omega_2} X(\omega_1, \omega_2)P_2(\omega_1, d\omega_2) \right] P_1(d\omega_1)$$

  - **Fubini (Classical)**: Let P_1 and P_2 be probability meaasures on Ω_1 and Ω_2, respectively, and $P = P_1 \times P_2$. If X is jointly measurable an either nonnegative or jointly integrable, then

$$\mathbb{E}(X) = \mathbb{E}_{P_1}\left[\mathbb{E}_{P_2}(X)\right] = \mathbb{E}_{P_2}\left[\mathbb{E}_{P_1}(X)\right].$$

    * We can often show joint integrability by showing that $\int_{\Omega_1} \int_{\Omega_2} |X(\omega_1, \omega_2)| P_2(\omega_1, d\omega_2)P_1(d\omega_1) < \infty$.
    * This result also holds for integration with respect to σ-finite measures in place of probability measures.
    * Nonnegativity (or joint integrability) in Fubini's theorem cannot be dropped. Individual integrability is not enough.
      · **Example**: Suppose $\Omega_1 = \Omega_2 = \mathbb{N}$, $\mathcal{A}_1 = \mathcal{A}_2 = \mathcal{P}(\mathbb{N})$, $\mu_1 = \mu_2$ is the counting measure, and

$$f(i, j) = \begin{cases} i & , j = i \\ -i & , j = i + 1 \\ 0 & , \text{o.w.} \end{cases}$$

      Although $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f(i, j) = \sum_{i=1}^{\infty} 0 = 0$, $\sum_{j=1}^{\infty} \sum_{i=1}^{\infty} f(i, j) = \sum_{j=1}^{\infty} [j - (j-1)] = \infty$. $\blacksquare$
    * σ-finiteness is also important in Fubini's theorem.
      · **Example**: $\Omega_1 = \Omega_2 = [0, 1]$, $\mathcal{A}_1 = \mathcal{A}_2 = \mathcal{R}$, $\mu_1 = \lambda$, and $\mu_2 = \#$, the counting measure, and let $f(x, y) = \mathbb{1}\{x = y\}$.

$$\int \int f d\mu_1 d\mu_2 = \int 0 d\mu_2 = 0 \neq 1 = \int 1 d\mu_1 = \int \int f d\mu_2 d\mu_1. \ \blacksquare$$

- In a multivariate setting, if a joint expectation exists, then Fubini guarantees that

$$\mathbb{E}X = \int X dP = \int \cdots \int X(\omega_1, \ldots, \omega_k) P(\omega_1, \ldots, \omega_{k-1} d\omega_k) \ldots P_1(d\omega_1).$$

  - $P(A_1 \times \cdots \times A_k)$ extends to unique probability measure on $\mathcal{A} \otimes \cdots \otimes \mathcal{A}_k = \sigma\langle \{A_1 \times \cdots \times A_k : A_i \in \mathcal{A}_i\} \rangle$.

- **Finite-Dimensional Set**: $\mathcal{F} = \{\mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_k \times \Omega_{k+1} \times \Omega_{k+2} \times \ldots\}$.

  - Let $A = A_1 \times \cdots \times A_k \times \Omega_{k+1} \times \Omega_{k+2} \times \ldots$. $A \subset \mathcal{F} \subset \mathcal{A}$.

- **Ionescu-Tulcea Theorem**: Give a hierarchical model $X_1 \sim P_1$, $X_2|X_1 \sim P_2, \ldots, (X_i|X_1, \ldots, X_{i-1}) \sim P_i$, a meaningful joint distribution for $(X_1, X_2, \ldots)$ exists.

  - This lets us define a probability measure on a countably infinite product space.
  - We can construct countably many independent RVs by decomposing the joint measure into the product of the individual measures.

- **Kolmogorov Consistency Theorem**: Suppose that all possible finite-dimensional joint distributions P_{T^*} of $(X_t : t \in T^*)$ are "consistent:" that is, if $T_1^* \subset T_2^*$ are finite, then $P_{T_2^*|T_1^*} = P_{T_1^*}$, where $|_F$ stands for marginalization to an index subset F. Then, there exists a unique probability measure P on the product space such that $P|_{T^*} = P_{T^*}$ for all finite T^*.

  - A resulting corollary is that an arbitrary number of independent random variables with arbitrarily specified distributions exist.

- **Example**: IBP. Suppose F and G are two distribution functions with no common points of discontinuity in an interval $(a, b]$. Show that

$$\int_{(a,b]} G(x)F(dx) = F(b)G(b) - F(a)G(a) - \int_{(a,b]} F(x)G(dx).$$

$$G(x) = \int \mathbb{1}\{y \le x\} G(dy) = \int_{(a,b]} \int \mathbb{1}\{y \le x\} G(dy)F(dx) = \int_{(a,b]} G(x)F(dx).$$

  Indicators are nonnegative and jointly measurable. In addition, x and y are dummy variables that we will interchange in the next statement.

$$\int_{(a,b]} F(x)G(dx) = \int_{(a,b]} \int \mathbb{1}\{y \le x\} F(dy)G(dx) = \int_{(a,b]} \int \mathbb{1}\{x \le y, a < y \le b\} F(dx)G(dy).$$

  Applying Fubini's theorem, we get

$$\int_{(a,b]} \int \mathbb{1}\{x \le y, a < y \le b\} F(dx)G(dy) = \int \int \mathbb{1}\{x \le y, a < y \le b\} G(dy)F(dx).$$

  Summing over the two derived terms,

$$\int_{(a,b]} G(x)F(dx) + \int_{(a,b]} F(x)G(dx) = \int \int \left[\{y \le x, a < x \le b\} + \mathbb{1}\{x \le y, a < y \le b\}\right] G(dy)F(dx).$$

  The second term evaluates to $F(b)G(b) - F(a)G(a)$, and rearranging yields the results. ■

- **Example**: For positive RV X,

  1. Use Fubini's theorem applied to σ-finite measures to prove that $\mathbb{E}(X) = \int_{[0,\infty)} P(X \ge t) dt$.
  2. Check also that for any $\alpha > 0$, $\mathbb{E}(X^\alpha) = \alpha \int_{[0,\infty)} x^{\alpha-1} P(X > x) dx$.
  3. Derive a similar expression for $\mathbb{E}(X)$ for X that is not necessarily non-negative.
  4. Suppose $\psi(X)$ is increasing, non-negative, and differentiable. Derive an expression for $\mathbb{E}[\psi(X)]$.

  1.

$$\mathbb{E}(X) = \int X dP = \int x dP(x) = \int \int \mathbb{1}\{0 \le t \le x\} d\lambda(t) \cdot dP(x)$$
$$\overset{\text{Fub.}}{=} \int \int \mathbb{1}\{0 \le t \le x\} dP(x) d\lambda(t) = \int_{[0,\infty)} P(X \ge t) dt.$$

2. Let $s = t^{1/\alpha} \implies dt = \alpha \cdot s^{\alpha-1} ds$.

$$\mathbb{E}(X^\alpha) = \int_0^\infty P(X^\alpha > t) dt = \int_0^\infty P(X > t^{1/\alpha}) dt = \int_0^\infty P(X > s) \alpha \cdot s^{\alpha-1} ds.$$

3. Assume $\mathbb{E}(X^+)$ and $\mathbb{E}(X^-)$ exist.

$$X = X^+ - X^- \implies \mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-) = \int_{[0,\infty)} \left[P(X^+ > t) - P(X^- > t)\right] dt.$$

Using the definitions of X^+ and X^-, $X^+ > t$ iff $X > t$. Conversely, $X^- > t$ iff $X < -t$. Therefore,

$$\int_{[0,\infty)} \left[P(X^+ > t) - P(X^- > t)\right] dt = \int_{[0,\infty)} \left[P(X > t) - P(X < -t)\right] dt = \int_{[0,\infty)} P(|X| > t) dt.$$

4. Note that we once again assume X is non-negative.

$$\psi(x) = \psi(0) + \int_{[0,x)} \psi'(t) dt = \psi(0) + \int_{[0,\infty)} \psi'(t) \mathbb{1}\{x > t\} dt.$$

Taking expectations on both sides yields

$$\mathbb{E}[\psi(X)] = \psi(0) + \int_{[0,\infty)} \psi'(t) P(X > t) dt. \blacksquare$$

- **Example**: For an RV X with distribution F, define the MGF $\phi(\lambda)$ by $\phi(\lambda) = \mathbb{E}(e^{tX})$. Let $\Lambda = \{\lambda \in R : \phi(\lambda) < \infty\}$, with $\lambda_\infty = \sup \Lambda$.

1. Prove that for λ in the interior of Λ that $\phi(\lambda) > 0$, and that $\phi(\lambda)$ is continuous on the interior of Λ.
2. Give examples of $\lambda_\infty \in \Lambda$ and $\notin \Lambda$.

1. $\phi(\lambda) \geq 0$ trivially. $\phi(\lambda) = 0 \implies \exp\{\lambda X\} = 0$ a.s. This means that $X = \pm\infty$ for $\phi(\lambda) = 0$. Since $X \in \mathbb{R} \not\supset \{-\infty, \infty\}$, this means that $\phi(\lambda) > 0$.
Next, suppose $\lambda \in \Lambda^0$. Choose $\varepsilon > 0$ such that $[\lambda - \varepsilon, \lambda + \varepsilon] \in \Lambda$. Suppose $\lambda_n \to \lambda_0$. Thus, for large n,

$$e^{\lambda_n X} \leq e^{(\lambda-\varepsilon)X} + e^{(\lambda+\varepsilon)X} \in L_1(F).$$

Thus, by DCT, $\phi(\lambda_n) \to \phi(\lambda_0)$.

2. Consider $F(x) = c(1+x)^\alpha e^{-x}$ for $x > 0$. In this case, $\lambda_\infty = 1$.
Case 1: $\alpha < 1$. $\int_0^\infty (1+x)^\alpha < \infty$, so $\lambda_\infty \in \Lambda$.
Case 2: $\alpha \geq 1$. $\int_0^\infty (1+x)^\alpha = \infty$, so $\lambda_\infty \notin \Lambda$.

- **Example**: Suppose $\{X_n, n \geq 1\}$ is a sequence of Bernoulli RVs with $P(X_n = 1) = p_n$. Show that $\sum_{i=1}^\infty p_i < \infty \implies \sum_{i=1}^\infty \mathbb{E}(X_i) < \infty \implies P(X_n \to 0) = 1$.

Let $X = \sum_{i=1}^\infty X_i$, which is a well-defined RV.

$$\mathbb{E}(X) \overset{\text{MCT}}{=} \sum_{i=1}^\infty \mathbb{E}(X_i) = \sum_{i=1}^\infty p_i < \infty.$$

Note that $X \geq 0$ and $X_n \geq 0$ for all n.

$$P(X < \infty) = 1 \implies P(\limsup_{n\to\infty} A_n) = 0 \overset{\text{Bor-Cant}}{\implies} P(X_n \to 0) = 1. \blacksquare$$

- **Example**: Pratt's Lemma. Let X_n, Y_n, X, and Y be RVs on the probability space $(\Omega, \mathcal{B}, P)$, such that $0 \leq X_n \leq Y_n$, $X_n \to X$, $Y_n \to Y$, and $\mathbb{E}(Y_n) \to \mathbb{E}(Y)$, where $\mathbb{E}(Y) < \infty$.

1. Show that $\mathbb{E}(X_n) \to \mathbb{E}(X)$.
2. Show that the DCT follows.

1. Since $X_n \geq 0$ and $X_n \to X$, then Fatou's lemma gives that

$$\mathbb{E}\left[\liminf_{n\to\infty} X_n\right] = \mathbb{E}(X) \leq \liminf_{n\to\infty} \mathbb{E}(X_n).$$

Next, we consider $0 \leq Y_n - X_n \to Y - X$. We can once again apply Fatou to get that $\mathbb{E}(Y - X) \leq \liminf_{n\to\infty} \mathbb{E}(Y_n - X_n)$. Since $0 \leq X_n \leq Y_n$ with $\mathbb{E}(Y) < \infty$, then $\mathbb{E}(X) < \infty$ (this eliminates the $\infty - \infty$ case, which is not defined). Thus,

$$\mathbb{E}(Y-X) = \mathbb{E}(Y)-\mathbb{E}(X) \leq \liminf_{n\to\infty} [\mathbb{E}(Y_n) - \mathbb{E}(X_n)] = \mathbb{E}(Y)-\limsup_{n\to\infty} \mathbb{E}(X_n) \implies \limsup_{n\to\infty} \mathbb{E}(X_n) \leq \mathbb{E}(X).$$

Since $\limsup_{n\to\infty} \mathbb{E}(X_n) \leq \mathbb{E}(X) \leq \liminf_{n\to\infty} \mathbb{E}(X_n)$, $\mathbb{E}(X_n) \to = E(X)$.

2. $0 \leq X_n \leq Y_n$ means that X_n is u.i. When combined with the fact that $X_n \to X$, we have thus satisfied the conditions needed for the DCT. ■

- **Example**: Suppose we are on the space $(\mathbb{R}, \mathcal{R}, \lambda)$. Let $f : \mathbb{R} \to \mathbb{R}$ be individually measurable and integrable. Show that f is jointly measurable $\int\int f(x)f(y-x)\lambda(dy)\lambda(dx) \geq 0$.

$f(x)$ is jointly measurable, since it doesn't depend on y, and f is individually measurable. $f(y-x)$ will also be measurable due to the identical measures imposed on y and x, so $f(x)f(y-x)$ is jointly measurable, being the product of two measurable functions.
Note that λ is translation invariant. Let $z = y - x$.

$$\int\int f(x)f(y-x)\lambda(dy)\lambda(dx) \geq 0 = \int |f(x)| \int |f(z)|\lambda(dz)\lambda(dx) = \int |f(z)|dz \cdot \int\int |f(x)|dx < \infty.$$

In fact, this integral evaluates to c^2 for some $c \in \mathbb{R}$, which is trivially ≥ 0. ■

- **Example**: Show that if X and Y are RVs with distribution functions $F(x)$ and $G(x)$ with no common discontinuities, then $\mathbb{E}[F(Y)] + \mathbb{E}[G(X)] = 1$.

Note that indicators are nonnegative, so the integrals involving indicators will also be nonnegative.

$$\begin{aligned}\mathbb{E}\left[F(Y)\right] + \mathbb{E}\left[G(X)\right] &= \int F(y)dG(y) + \int G(x)dF(x) = \int\int \mathbb{1}\{y \leq Y\}dF(x)dG(y) + \int G(x)dF(x) \\ &\overset{\text{Fubini}}{=} \int\int \mathbb{1}\{Y \geq y\}dG(y)dF(x) + \int G(x)dF(x) \\ &= \int \left[1 - G(x)\right] dF(x) + \int G(x)dF(x) = \int \left[1 - G(x) + G(x)\right] dF(x) = 1 \int dF(x) = 1. \blacksquare\end{aligned}$$

- **Example**: Show that for a distribution function $F(x)$, $\int_{\mathbb{R}} \left(F(x+a) - F(x)\right) dx = a$, where dx is the Lebesgue measure.

$$\int_{\mathbb{R}} \left[F(x+a) - F(x)\right] dx = \int_{\mathbb{R}} \int \left[\mathbb{1}\{y \leq x+a\} - \mathbb{1}\{y \leq x\}\right] d\lambda(y)dx = \int_{\mathbb{R}} \int \left[\mathbb{1}\{y \leq x \leq y+a\}\right] d\lambda(y)dx.$$

Since we have a jointly measurable and nonnegative random variable (with marginal Lebesgue measures and it is an indicator), we can apply Fubini's theorem to get that

$$\int_{\mathbb{R}} \int \left[\mathbb{1}\{y \leq x \leq y+a\}\right] d\lambda(y)dx = \int \int_{\mathbb{R}} \left[\mathbb{1}\{y \leq x \leq y+a\}\right] dxd\lambda(y) = \int (y+a-y)d\lambda(y) = a \int d\lambda(y) = a. \blacksquare$$

- **Example**: Suppose X and $Y \in L_1$.
  1. Show that $\mathbb{E}(Y) - \mathbb{E}(X) = \int_{\mathbb{R}} \left[P(X < x \leq Y) - P(Y < x \leq X)\right] dx$.
  2. Show that the expected length of the random interval $(X, Y]$ is the integral with respect to x of $P(x \in (X, Y])$.

1. First, note that $Y - X = \int_{\mathbb{R}} \left[\mathbb{1}\{X \leq x \leq Y\} - \mathbb{1}\{Y < x < X\}\right] dx$. Since X and $Y \in L_1$, they are integrable by construction. Therefore, we can apply Fubini's theorem to get that

$$\mathbb{E}(Y-X) = \mathbb{E}\left\{\int_{\mathbb{R}} \left[\mathbb{1}\{X \leq x \leq Y\} - \mathbb{1}\{Y < x < X\}\right] dx\right\} = \int_{\mathbb{R}} \left\{\mathbb{E}\left[\mathbb{1}\{X \leq x \leq Y\} - \mathbb{1}\{Y < x < X\}\right]\right\} dx.$$

This simplifies down to

$$\mathbb{E}(Y - X) = \mathbb{E}(Y) - E(X) = \int_{\mathbb{R}} \left\{P\{X \leq x \leq Y\} - P\{Y < x < X\}\right\} dx.$$

2. Note that $Y < x < X$ will never occurs, since in that case $(X, Y]$ would be empty, which would never contain x. Applying the previous result,

$$\int_{\mathbb{R}} P\left[x \in (X, Y]\right] dx = \int_{\mathbb{R}} P(X < x \leq Y) dx = \int_{\mathbb{R}} \mathbb{E}\left[\mathbb{1}\{X < x \leq Y\}\right] = \mathbb{E}(Y - X). \blacksquare$$

9.7 Convergence

Return to Table of Contents

- There are two variants of convergence:
  - **Pointwise Convergence**: $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$.
  - **Uniform Convergence**: $\sup_{\omega \in \Omega} |X_n(\omega) - X(\omega)| \to 0$.
  - By introducing probability, we can exclude near-zero probability events from happening.
    - If $P(|X| > a) \approx 0$, then we may say that $-a \leq X \leq a$ w.h.p. This will show up once we start proving things with convergence!
- **Almost Sure Convergence**: $X_n \overset{\text{a.s.}}{\to} X$ $[P]$ if $P(\{\omega : X_n(\omega) \not\to X(\omega)\}) = 0$.
  - $X_n \overset{\text{a.s.}}{\to} X$ iff $\exists N \in \mathcal{A}, P(N) = 0$ such that $\forall \epsilon > 0$ and $\omega \in N^c$, $\exists m \geq 1 : \forall n \geq m, |X_n(\omega) - X(\omega)| < \epsilon$.
    - ϵ can run over a countable set, such as $\frac{1}{k}$. This lets us take a countable union and apply lim sup's. For instance,

$$X_n \overset{\text{a.s.}}{\to} X \equiv P\left[\bigcup_{\epsilon \in \mathbb{Q}^+} \underbrace{\bigcap_{n=1}^{\infty} \bigcup_{m \geq n} (|X_m - X| > \epsilon)}_{= \limsup_{n \to \infty} (|X_m - X| > \epsilon)}\right] = 0 \equiv \forall \epsilon > 0, P(|X_n - X| > \epsilon \text{ i.o.}) = 0.$$

  - A sufficient condition (using Borel-Cantelli) is that $\forall \epsilon > 0$, $\sum_{i=1}^{\infty} P(|X_i - X| > \epsilon) < \infty$.
    - **Example**: $U_1, \ldots, U_n \overset{\text{iid}}{\sim} Unif(0, 1)$. Let $X_n = \max\{U_1, \ldots, U_n\}$.

      Since $X_n \leq 1$,

$$P(|X_n - 1| > \epsilon \ \forall n) = P(1 - X_n > \epsilon \ \forall n) = \prod_{i=1}^{n} P(X_n < 1 - \epsilon) = (1 - \epsilon)^n \to 1.$$

      Therefore, $X_n \overset{\text{a.s.}}{\to} 1$. $\blacksquare$
- **Convergence in Probability**: $X_n \overset{\text{p}}{\to} X$ if $\forall \epsilon > 0$, $P(|X_n - X| \geq \epsilon) \to 0$ as $n \to \infty$.
  - **Example**: $X_n \sim \text{Exp}(c_n)$, where $c_n \to \infty$. $P(|X_n| > \epsilon) = e^{-c_n \epsilon} \to 0$, so $X_n \overset{\text{p}}{\to} 0$. $\blacksquare$
  - If $X_n \overset{\text{p}}{\to} 0$, then $X_n = o_p(1)$. If $X_n = o_p(c_n)$, then $\frac{X_n}{c_n} \overset{\text{p}}{\to} 0$.
  - A sufficient condition (using Chebyshev) is that $P(X_n - X) \geq \epsilon) \leq \frac{1}{\epsilon^p} \mathbb{E}\left(|X_n - X|^p\right)$ for any $p > 0$.
    - If $d_p(X_n, X) \to 0$, then $X_n \overset{\text{p}}{\to} X$.
  - **Example**: Suppose Z_n are uncorrelated, with mean zero and variance $\frac{1}{n}$. Also let $X_n = \sum_{i=1}^{n} \frac{Z_i}{\log n}$. Then,

$$\mathbb{E}(|X_n|^2) = Var(X_n) = \frac{1}{(\log n)^2} \sum_{i=1}^{n} Var(Z_i) = \sum_{i=1}^{n} \frac{1/i}{(\log n)^2} \sim \frac{\log n}{(\log n)^2} \to 0.$$

    Therefore, $X_n \overset{\text{p}}{\to} 0$. $\blacksquare$
  - $\overset{\text{a.s.}}{\to} \implies \overset{\text{p}}{\to}$.
    - *Proof*: Let $A_n = \{\omega : |X_n(\omega) - X(\omega)| \geq \epsilon\}$. $\overset{\text{a.s.}}{\to}$ happens when $P\left(\bigcup_{n=N}^{\infty} A_n\right) \to 0$, and $\overset{\text{p}}{\to}$ happens when $P(A_n) \to 0$. Since $A_n \subset \bigcup_{n=N}^{\infty} A_n$, the result immediately follows $\blacksquare$
    - **Example**: $\Omega = (0, 1]$, $\mathcal{A} = \mathcal{R}_{(0,1]}$, $P = \lambda_{(0,1]} = Unif(0, 1)$. Let $X_1 = 1$, $X_2 = \mathbb{1}(0, \frac{1}{2}]$, $X_3 = \mathbb{1}(\frac{1}{2}, 1]$, $X_4 = \mathbb{1}(0, \frac{1}{4}]$, $X_5 = \mathbb{1}(\frac{1}{4}, \frac{1}{2}]$, $X_2 = \mathbb{1}(\frac{1}{2}, \frac{3}{4}]$, $X_2 = \mathbb{1}(\frac{3}{4}, 1]$, etc.

      Since $P(X_n) = \frac{1}{n} \to 0$, $X_n \overset{\text{p}}{\to} 0$, but $X_n \overset{\text{a.s.}}{\not\to} X$. $\blacksquare$

- If $X_n \xrightarrow{\text{p}} X$ and $Y_n \xrightarrow{\text{p}} Y$, then:
  - $aX_n + bY_n \to aX + bY$.
    - As a result, $o_p(1) + o_p(1) = o_p(1)$.
  - $X_n Y_n \to XY$.
  - $\frac{X_n}{Y_n} \to \frac{X}{Y}$, if $P(Y \neq 0) = 1$.
  - If $g : \mathbb{R}^2 \to \mathbb{R}$ is continuous, $g(X_n, Y_n) \to g(X, Y)$.
    - *Proof*: nts $P\left(|g(X_n, Y_n) - g(X, Y)| > \epsilon\right) \to 0$ w.h.p.

      For all $\epsilon > 0$, $\exists M > 0 : P(|X| > M) < \epsilon$ for any RV X. If $X_n \xrightarrow{\text{p}} X$, then for any $\delta > 0$,

      $$P(|X_n| > M + \delta) \leq P(|X| > M) + P(|X_n - X| > \delta) \implies g \text{ is uniformly continuous.}$$

      This means that $|X_n - X| + |Y_n - Y| < \delta \implies |g(X_n, Y_n) - g(X, Y)| < \epsilon$. ■
- **Rapidly Varying**: A distribution tail $1 - F(x)$ such that $\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = \infty$ if $x \in (0, 1)$ (0 o.w.).
- **Example**: Suppose $\{X_n, n \geq 1\}$ is iid and non-negative, and define $M_n = \bigvee_{i=1}^n X_i$.
  1. Show that $P(M_n > x) \leq nP(X_1 > x)$.
  2. Show that if $\mathbb{E}(X_1^p) < \infty$, then $\frac{M_n}{n^{1/p}} \xrightarrow{\text{p}} 0$.
  3. Show that $\frac{M_n}{n} \xrightarrow{\text{p}} 0$ iff $nP(X_1 > n) \to 0$ as $n \to \infty$.
  4. Show that there exists $b(n) \to \infty$ such that $\frac{M_n}{b(n)} \xrightarrow{\text{p}} 1$ iff $1 - F(x) = P(X_1 > x)$ is rapidly varying at ∞. We may take $b(n) = \left(\frac{1}{1-F}\right)^{\leftarrow}(n)$ to be the $1 - \frac{1}{n}$ quantile of F.
  5. Now, suppose $\{X_n\}$ is an arbitrary sequence of non-negative RVs. Show that $\mathbb{E}\left[M_n \mathbb{1}\{M_n \geq \delta\}\right] \leq \sum_{i=1}^n \mathbb{E}\left[X_i \mathbb{1}\{X_i \geq \delta\}\right]$. Also show that if $\{X_n\}$ is u.i., then $\frac{\mathbb{E}(M_n)}{n} \to 0$.

  1. Note that $\mathbb{1}\{M_n > x\} = \bigcup_{i=1}^n \mathbb{1}\{X_i > x\}$. Therefore,

     $$P(M_n > x) \leq \sum_{i=1}^n P(X_i > x) \overset{\text{iid}}{=} nP(X_1 > x).$$

  2. We need to show that $\lim_{n\to\infty} P\left(\left|\frac{M_n}{n^{1/p}} - 0\right| > \epsilon\right) = 0$ for all $\epsilon > 0$. Note that since $\{X_i\}$ is non-negative, so must M_n be.

     $$P\left(\left|\frac{M_n}{n^{1/p}}\right| > \epsilon\right) = P\left(M_n > n^{1/p}\epsilon\right) \leq n \cdot P(X_1 > n^{1/p}\epsilon) = n \cdot P(X_1^p > n\epsilon^p).$$

     Since $\mathbb{E}(X_1^p) < \infty$, that means that $a \cdot P(X_1^p > a) \to 0$ as $a \to \infty$. This means that $\forall \delta > 0$, $\exists a_0 : \forall a \geq a_0$, $a \cdot P(X_1^p > a) < \delta$. In this problem, $a = n\epsilon^p$. Therefore,

     $$n\epsilon^p \cdot P\left(X_1 > n\epsilon^p\right) < \delta \implies P\left(X_1 > n\epsilon^p\right) < \frac{\delta}{n\epsilon^p}.$$

     Since δ is arbitrary, $P\left(\left|\frac{M_n}{n^{1/p}} - 0\right| > \epsilon\right) \to 0$.
  3. First, suppose $n \cdot P(X_1 > n) \to 0$. Suppose $\epsilon \in (0, 1)$.

     $$P\left(\left|\frac{M_n}{n} - 0\right| > \epsilon\right) = P(M_n > \epsilon n) \leq n \cdot P(X_1 > \epsilon n) \leq \epsilon n \cdot P(X_1 > \epsilon n);$$

     We can let $m = \epsilon n$, so as $m \to \infty$, $m \cdot P(X_1 > m) \to 0$, and the claim holds. Now, if $\epsilon \geq 1$, $P\left(\left|\frac{M_n}{n} - 0\right| > \epsilon\right) \leq P\left(\frac{M_n}{n} > 1\right) = P(M_n < n) \leq n \cdot P(X_1 > n) \to 0$.
     Now, suppose $\frac{M_n}{n} \xrightarrow{\text{p}} 0$. Since $P\left(\left|\frac{M_n}{n} - 0\right| > \epsilon\right)$ holds for all $\epsilon > 0$, we choose $\epsilon = 1$.

     $$P(M_n > n) = 1 - P(M_n < n) = 1 - [1 - P(X_1 > n)]^n \approx n \cdot P(X_1 > n)$$

     when $n \cdot P(X_1 > n)$ is small. Therefore $P(M_n > n) \to 0 \implies n \cdot P(X_1 > n) \to 0$.
  4. First, suppose that $1-F(x)$ is rapidly varying. This means that $\lim_{t\to\infty} \frac{P(X_1>xt)}{P(X_1>t)} = 0$ for $x \in (0, 1)$. Using the definition of $b(n)$,

     $$P\left(M_n \leq \frac{1}{\epsilon}b(n)\right) = \left[1 - P\left(X_1 > \frac{1}{\epsilon}b(n)\right)\right]^n \to 0.$$

     Choosing $\epsilon' = \frac{1}{\epsilon}$ yields $P(M_n > \epsilon' b(n)) = P\left(M_n \leq \frac{1}{\epsilon}b(n)\right) \to 0 \implies \frac{M_n}{n} \xrightarrow{\text{p}} 1$.
     Now, suppose $\exists b(n) \to \infty$ such that $\frac{M_n}{b(n)} \xrightarrow{\text{p}} 1$. We can reverse the above proof to get that $1 - F(x)$ is rapidly varying.

5. Since $M_n = \max\{X_1, \ldots, X_n\}$, $\mathbb{1}\{M_n > a\} \subset \mathbb{1}\{X_i < a\}$, and $M_n \leq \sum_{i=1}^n X_i$. Therefore,

$$M_n \mathbb{1}\{M_n \geq \delta\} \leq \sum_{i=1}^n \left[X_i \mathbb{1}\{X_i \geq \delta\}\right] \implies \mathbb{E}\left[M_n \mathbb{1}\{M_n \geq \delta\}\right] \leq \sum_{i=1}^n \mathbb{E}\left[X_i \mathbb{1}\{X_i \geq \delta\}\right].$$

Next, since X_n is u.i., $\forall \epsilon > 0\ \exists \delta > 0$ such that $\sup_i \mathbb{E}\left[X_i \mathbb{1}\{X_i \geq \delta\}\right] < \epsilon$.

$$\mathbb{E}(M_n) = \mathbb{E}\left[M_n \mathbb{1}\{M_n < \delta\}\right] + \mathbb{E}\left[M_n \mathbb{1}\{M_n \geq \delta\}\right] \leq \delta + \sum_{i=1}^n \mathbb{E}\left[X_i \mathbb{1}\{X_i \geq \delta\}\right] < \delta + n\epsilon$$

Divide by n to get that $\frac{\mathbb{E}(M_n)}{n} \leq \frac{\delta}{n} + \epsilon \to 0 + 0 = 0$. ■

– **Example**: Let $\{X_n\}$ be a sequence of RVs.

1. Show that if $X_n \xrightarrow{\text{p}} 0$, then for any $p > 0$, $\frac{|X_n|^p}{1+|X_n|^p} \xrightarrow{\text{p}} 0$.
2. Show that $X_n \xrightarrow{\text{p}} 0$ iff $\mathbb{E}\left(\frac{|X_n|^p}{1+|X_n|^p}\right) \to 0$.
3. Show that if $\frac{|X_n|^p}{1+|X_n|^p} \xrightarrow{\text{p}} 0$ for some $p > 0$, then $X_n \xrightarrow{\text{p}} 0$.

1. Since we know $X_n \xrightarrow{\text{p}} 0$,

$$P\left(\left|\frac{|X_n|^p}{1+|X_n|^p} - 0\right| > \epsilon\right) = \cdots = P\left(|X_n| > \left(\frac{\epsilon}{1-\epsilon}\right)^p\right) \to 0.$$

This is because we can treat $\left(\frac{\epsilon}{1-\epsilon}\right)^p$ as the ϵ for $X_n \xrightarrow{\text{p}} 0$.

2. First, suppose $X_n \xrightarrow{\text{p}} 0$. We can bound $\frac{|X_n|^p}{1+|X_n|^p}$ above by 1 and below by zero. Therefore, applying the bounded convergence theorem, $\mathbb{E}\left(\frac{|X_n|^p}{1+|X_n|^p}\right) \to \frac{0}{1+0} = 0$.
Now, suppose $\mathbb{E}\left(\frac{|X_n|^p}{1+|X_n|^p}\right) \xrightarrow{\text{p}} 0$. Applying Markov's inequality,

$$P\left(\frac{|X_n|^p}{1+|X_n|^p} \geq \epsilon\right) \leq \frac{1}{\epsilon}\mathbb{E}\left(\frac{|X_n|^p}{1+|X_n|^p}\right) \to 0.$$

$\frac{|x|^p}{1+|x|^p}$ is increasing wrt x, so that means that $\forall \epsilon > 0$, $\exists \delta > 0$ such that $|x| \geq \delta \implies \frac{|x|^p}{1+|x|^p} \geq \epsilon$. This means that

$$P(|X_n| \geq \delta) \leq P\left(\frac{|X_n|^p}{1+|X_n|^p} \geq \epsilon\right) \to 0.$$

3. This will reverse the first proof. Let $\delta = \left(\frac{\epsilon}{1-\epsilon}\right)^p$. Since we know $\frac{|X_n|^p}{1+|X_n|^p} \xrightarrow{\text{p}} 0$,

$$P\left(|X_n| > \delta\right) = \cdots = P\left(\left|\frac{|X_n|^p}{1+|X_n|^p}\right| > \epsilon\right) \to 0.$$ ■

– **Example**: Suppose $\{X_n, n \geq 1\}$ are independent, non-negative RVs where $X_n \sim (\mu_n, \sigma_n^2)$. Suppose $\sum_{i=1}^\infty \mu_i < \infty$ and $\sigma_n^2 \leq c\mu_n$ for some $c > 0$ and all n. Show that $\frac{S_n}{\mathbb{E}(S_n)} \xrightarrow{\text{p}} 1$.

$$P\left(\left|\frac{S_n}{\mathbb{E}(S_n)} - 1\right| > \epsilon\right) = P\left(S_n - \mathbb{E}(S_n) \geq \epsilon \cdot \mathbb{E}(S_n)\right) \leq \frac{Var(S_n)}{\left[\epsilon \cdot \mathbb{E}(S_n)\right]^2} \leq \frac{c}{\mathbb{E}(S_n)}.$$

Since X_n is nonnegative, $S_n \to \infty$ as $n \to \infty$, so $\frac{c}{\mathbb{E}(S_n)}$, and thus $\frac{S_n}{\mathbb{E}(S_n)} \xrightarrow{\text{p}} 1$. ■

- **Example**: Show that in a discrete probability space, $\xrightarrow{\text{p}} \equiv \xrightarrow{\text{a.s.}}$.
$\xrightarrow{\text{a.s.}} \implies \xrightarrow{\text{p}}$ is trivial, so we show the other direction.
Suppose $X_n \xrightarrow{\text{p}} X$ on a discrete probability space. In such a probability space, singletons $\{\omega\}$ are measurable and can have positive probabilities. $X_n \xrightarrow{\text{p}} X$ means that $\forall \epsilon > 0$, $A_n^\epsilon = \{\omega : |X_n(\omega) - X(\omega)| \geq \epsilon\}$ decreases as $n \to \infty$.
Proceed with contradiction. Suppose $\xrightarrow{\text{p}} \not\Rightarrow \xrightarrow{\text{a.s.}}$. This means that there exists a set $A = \{\omega : X_n(\omega) \not\xrightarrow{\text{p}} X(\omega)\}$ with nonzero probability. Since the probability space is discrete, $\exists \omega_0 : P(\{\omega_0\}) > 0$.
Since $X_n(\omega_0) \not\to X(\omega)$, $\exists \epsilon > 0 : |X_n(\omega) - X(\omega)| \geq \epsilon$ for infinitely many n. This means that $P(|X_n - X| \geq \epsilon) \geq P(\{\omega_0\}) > 0$, which is a contradiction, since $X_n \xrightarrow{\text{p}} X$. Therefore, under a discrete probability space, $\xrightarrow{\text{p}} \equiv \xrightarrow{\text{a.s.}}$. ■

- **Convergence in Distribution**: $X_n \xrightarrow{\text{d}} X$ if $F_n(x) \to F(x)$ for all x where F is continuous.
  - Note that the convergence is now wrt the distribution of X, not X by itself.
  - **Example**: F is discontinuous at x. Suppose $X_n \equiv \frac{1}{n}$.

    $F_n(x) = \mathbb{1}[\frac{1}{n}, \infty)(x)$. Therefore, $F_n(0) = 0$, but $\lim_{n\to\infty} F_n(0) = 1$, so $X_n \not\xrightarrow{\text{d}} 0$.
  - **Example**: $U_n \overset{\text{iid}}{\sim} Unif(0,1)$, and $X_n = n \cdot \min\{U_1, \ldots, U_n\}$.

$$P(X_n > x) = P\left(\min\{U_1, \ldots, U_n\} > \frac{x}{n}\right) = \prod_{i=1}^{n}\left[1 - P\left(U_i > \frac{x}{n}\right)\right] = \left(1 - \frac{x}{n}\right)^n \to e^{-x}.$$

    Therefore, $X_n \xrightarrow{\text{d}} \text{Exp}(1)$. ■
  - If $X_n \xrightarrow{\text{d}} X$ iff $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$ for bounded and continuous g.
    - *Proof*: $\mathbb{E}[g(X_n)] = \int g(x) dP_n(x)$. We ignore areas of X_n with low probability, in order to form a compact set in the domain. Since that is the case, we can apply a Riemann integral.
      We need to choose boundary points a and b that are not discontinuous. A previous lemma shows that we can always find such a set of points. Applying the definition of $\xrightarrow{\text{d}}$, $P(a \le X_n \le b) \to P(a \le X \le b)$.

      For the converse direction, construct two sequences

$$0 \le \mathbb{1}\left\{x \in \left(-\infty, a - \frac{1}{k}\right)\right\} \le g_k(x) \le \mathbb{1}\{x \in (-\infty, a]\} \le f_k(x) \le \mathbb{1}\left\{x \in \left(-\infty, a + \frac{1}{k}\right]\right\} \le 1.$$

      For any $k \ge 1$,

$$P\left(X \le a - \frac{1}{k}\right) \le \mathbb{E}[g_k(X)] = \lim_{n\to\infty} \mathbb{E}[g_k(X_n)] \le \liminf_{n\to\infty} P(X_n \le a) \le \limsup_{n\to\infty} P(X_n \le a).$$

      This term is

$$\le \lim_{n\to\infty} \mathbb{E}[g_k(X_n)] = \mathbb{E}[g_k(X)] \le P\left(X \le a + \frac{1}{k}\right).$$

      Taking the limit as $k \to \infty$,

$$P(X < a) \le \liminf_{n\to\infty} P(X_n \le a) \le \limsup_{n\to\infty} P(X_n \le a) \le P(X \le a) \implies P(X_n \le a) \to P(X \le a)$$

      if a is a continuity point of the distribution of X. ■
  - Note that $X_n + Y_n \xrightarrow{\text{d}} X + Y$ is not guaranteed.
    - **Example**: $X \sim \mathcal{N}(0,1)$, $X_n = X$, $Y_n = (-1)^n X$. $X_n + Y_n = 2X$ if n is even, and 0 if odd, which means $X_n + Y_n \not\xrightarrow{\text{d}} \mathcal{N}$. ■
    - This statement is true if we have joint distributional convergence. That is, $(X_n, Y_n) \xrightarrow{\text{d}} (X, Y)$ is good.
  - If $X_n \to X$ and g is continuous, then $g(X_n) \to g(X)$.
    - Note that there is no boundedness required.
    - This allows us to talk about RVs with abstract values.
    - *Proof*: Verify that for any bounded and continuous function f on $\mathbb{R}$, $\mathbb{E}[f(g(X_n))] \to \mathbb{E}[f(g(X))]$. Let $Y_n = g(X_n)$. Since $f \circ g$ is a bounded and continuous function and $X_n \xrightarrow{\text{d}} X$, we can apply the previous theorem to get that

$$\mathbb{E}[f(Y_n)] = \mathbb{E}[f(g(X_n))] \to \mathbb{E}[f(g(X))] = \mathbb{E}[f(Y)].$$ ■

- **Continuous Mapping Theorem**, or **CMT**: Suppose $X_n \xrightarrow{\text{d}} X$ and g is a real-valued, measurable (not necessarily continuous) function with points of discontinuity D. If $P(X \in D) = 0$, then $g(X_n) \xrightarrow{\text{d}} g(X)$.
  - *Proof*: Similar to the previous corollary. Only difference is in accounting for D. ■
- $\xrightarrow{\text{p}} \implies \xrightarrow{\text{d}}$. $\xrightarrow{\text{d}} \implies \xrightarrow{\text{p}}$ iff $X \equiv c$.

- *Proof*: Apply the Bounded Convergence Theorem. For any bounded continuous function, $g(X_n) \xrightarrow{p} g(X) \implies \mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)] \implies X_n \xrightarrow{d} X$.
  For the converse,

$$P(X_n) \leq c + \epsilon) \to F(c + \epsilon) = 1 \text{ and } P(X_n \leq c - \epsilon) \to F(c - \epsilon) = 0.$$

  This implies that

$$P(|X_n - c| > \epsilon) \leq 1 - P(c - \epsilon \leq X_n \leq c + \epsilon) =\to 1 - 1 = 0. \blacksquare$$

- If X_n has density p_n, and X has density p wrt measure μ, and $p_n \to p$, then $X_n \xrightarrow{d} X$.
  - *Proof*: For any A,

$$\left| \int_A (p_n(x) - p(x)) d\mu(x) \right| \overset{\text{Scheffe}}{\leq} \int |p_n(x) - p(x)| \, d\mu(x) \to 0.$$

    $A \subset \infty$, which is why we drop A from the integral. $\blacksquare$
  - **Example**: $N(\mu_n, \sigma_n^2) \xrightarrow{d} \mathcal{N}(\mu, \sigma^2)$ if $\mu_n \to \mu$ and $\sigma_n \to \sigma > 0$. $\blacksquare$
  - **Example**: If $n \to \infty$, $\theta_n \to 0$, $n\theta_n \to \lambda > 0$, then $\text{Bin}(n, \theta_n) \xrightarrow{d} \text{Pois}(\lambda)$. $\blacksquare$

- **Slutsky's Theorem**: Suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$. Then, $X_n + Y_n \xrightarrow{d} X + c$, $X_n Y_n \xrightarrow{d} cX$, and if $c \neq 0$ then $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$.
  - *Proof*: Suffices to show that $(X_n, Y_n) \xrightarrow{d} (X, c)$, since we could then apply CMT.
    If x is a continuity point of the distribution of X, then

$$P(X_n \leq x, Y_n \leq y) \to P(X \leq x) \text{ for } y > c, \text{ and } P(X_n \leq x, Y_n \leq y) \to 0 \text{ for } y < c. \blacksquare$$

- **Stochastically Bounded**, or **Tight**: A sequence $X_n = O_p(1)$ if for any $\epsilon > 0$, $\exists M > 0 : P(|X_n| > M) < \epsilon$ for all n.
  - $X_n = O_p(c_n)$ if $\frac{X_n}{c_n} = O_p(1)$.
  - Fixed sequences $X_n = X$ are stochastically bounded.
  - $\xrightarrow{d}$ implies stochastic boundedness.
    - Suppose $X_n \xrightarrow{d} X$. Take $\epsilon > 0$ and $M > 0$ such that $\forall n \geq n_0$, $|P(|X_n| \leq M) - P(|X| \leq M)| < \frac{\epsilon}{2}$. Therefore, $P(|X_n| > M) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$.
  - A common sufficient condition for $X_n = O_p(1)$ is $\sup_n \mathbb{E}(|X_n|^\delta) < \infty$ for some $\delta > 0$.
    - Applying Markov,

$$\sup_n P(|X_n| \geq M) \leq \sup_n \frac{\mathbb{E}(|X_n|^\delta)}{M^\delta} < \epsilon.$$

  - **Prokhorov's Theorem**: If $X_n = O_p(1)$, then given any subsequence n_k, $\exists n_{k_m}$ such that $X_{n_{k_m}} \xrightarrow{d} X$ for some X.
    - This is valid in abstract spaces.

- **Weak Law of Large Numbers (iid Sequences)**: If X_n's are iid with finite mean μ, then $\bar{X}_n \xrightarrow{p} \mu$.
  - *Proof*: First, assume $Var(X_i) = \sigma^2 < \infty$.

$$P\left(\left| \frac{1}{n} \sum_{i=1}^n [X_i - \mathbb{E}(X)] \right| > \epsilon \right) \leq \frac{Var\left(\sum_{i=1}^n X_i\right)}{n^2 \epsilon^2} = \frac{\sum_{i=1}^n Var(X_i)}{n^2 \epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0.$$

    Now, relax the finite variance assumption. Fix a large $C > 0$ (to be chosen later), and let $Y_i = X_i \mathbb{1}\{|X_i| \leq C\}$, and $Y_i = X_i \mathbb{1}\{|X_i| > C\}$ (note that Z_i may not have finite variance), so $X_i = Y_i + Z_i$. Y_i is iid and has finite variance, so $\bar{Y}_n \to \mathbb{E}(Y) = \mathbb{E}(X \mathbb{1}\{|X| \leq C\})$. Now,

$$\mathbb{E}\left| \frac{1}{n} \sum_{i=1}^n Z_i \right| \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}|Z_i| \overset{\text{iid}}{=} \mathbb{E}(|X| \mathbb{1}\{|X| > C\}).$$

    Let $C \to \infty$. Then, $\mathbb{E}(Y) \to \mathbb{E}(X)$ by DCT, and $\mathbb{E}(|X| \mathbb{1}\{|X| > C\}) \to 0$, also by DCT. $\blacksquare$

- **WLLN, Uncorrelated Sequences**: Let X_n be uncorrelated RVs with bounded variances. If $\mathbb{E}(X_i) = 0$ for all i, then $\bar{X}_n \xrightarrow{\text{P}} 0$, and if the means are non-zero, then $\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i) \xrightarrow{\text{P}} 0$.
  - *Proof*: Let C be the upper bound of the variances.
$$P\left(\left|\frac{1}{n}\sum_{i=1}^n [X_i - \mathbb{E}(X)]\right| > \epsilon\right) \overset{\text{Uncor.}}{\leq} \frac{\sum_{i=1}^n Var(X_i)}{n^2\epsilon^2} \leq \frac{C}{n\epsilon^2} \to 0.$$
  - A sufficient condition is $\frac{1}{n^2}\sum_{i=1}^n \sigma_i^2 \to 0$.
- WLLN holds under independence plus uniform integrability.
  - *Proof*: Just like above, use truncation. $\sup_i \mathbb{E}(|X_i|\mathbb{1}\{|X_i| > C\}) \to 0$ as $C \to \infty$ by u.i. ■

- **Strong Law of Large Numbers (iid Sequences)**: If X_n are iid with finite mean μ and finite fourth moment, then $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$.
  - SLLN is stronger than WLLN, but WLLN is easier to prove, and is often sufficient for statistical applications.
  - *Proof*: WLOG, assume $\mathbb{E}(X) = 0$. By the Borel-Cantelli lemma, it suffices to prove that $\forall \epsilon > 0$, $\sum_{i=1}^n P(|\bar{X}_n| > \epsilon) < \infty$.
$$P(|\bar{X}_n| > \epsilon) \overset{\text{Cheb.}}{\leq} \frac{1}{\epsilon^2}\mathbb{E}(\bar{X}_n^2) = \frac{1}{n\epsilon^2}\mathbb{E}(X^2).$$
    However, this sequence is not enough, since summing over n would get rid of the n in the denominator. Therefore, we need to increase the power of n^{-1}. We want an even moment so we can drop the absolute value from the Chebyshev inequality, so the next choice uses the finite fourth moment assumption, and Chebyshev.
$$P(|\bar{X}_n| > \epsilon) \leq \frac{\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^4\right]}{n^4\epsilon^4}.$$
    By assumed independence,
$$\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^4\right] = \sum_i \sum_j \sum_k \sum_\ell \mathbb{E}(X_i X_j X_k X_\ell) = 0 \text{ for all different } i, j, k, \ell.$$
    Therefore, we only consider cases where an index is equal. There is $n(n-1)$ instances of equal indexes (dropping a constant), so $\mathbb{E}(\bar{X}_n^4) \leq \frac{C}{n^2}$, which is summable. ■
  - **Kronecker's Lemma**: If a_n is a real sequence and b_n is a positive sequence increasing to infinity such that $\sum_{i=1}^n \frac{a_i}{b_i}$ converges to a finite number, then $\frac{1}{b_n^{-1}}\sum_{i=1}^n a_i \to 0$.
  - **SLLN (Independent Sequences)**: Let X_n be independent RVs such that $\sum_{i=1}^n \frac{\mathbb{E}(X_n^2)}{n^2} < \infty$. If $\mathbb{E}(X_n) = 0$ for all n, then $\bar{X}_n \xrightarrow{\text{a.s.}} 0$. Otherwise, $\frac{1}{n}\sum_{i=1}^n [X_i - \mathbb{E}(X_i)] \xrightarrow{\text{a.s.}} 0$.
    - *Proof*: Since $\sum_{i=1}^\infty \frac{\mathbb{E}(X_i^2)}{n^2} < \infty$, $\frac{1}{n}\sum_{i=1}^\infty X_i$ must also converge almost surely. Apply Kronecker's lemma with $a_n = X_n$ and $b_n = n$ to complete the proof. ■
  - If $X_1, X_2, \ldots$ are independent, mean-zero RVs such that $\sum_{i=1}^n \mathbb{E}(X_n^2) < \infty$, then $\sum_{i=1}^n X_n$ converges almost surely.
    - *Proof*: Omitted. However, it involves establishing Kolmogorov's maximal inequality, and then showing that the sequence of partial sums is a Cauchy sequence. ■
  - Let X_n be iid. If $\bar{X}_n \to a$, where a is finite, then $E(X) = a$, and $\mathbb{E}|X| < \infty$.
    - *Proof*:
$$\frac{X_n}{n} = \frac{S_n}{n} - \frac{n-1}{n}\frac{S_{n-1}}{n-1} \to a - a = 0.$$
      This means that $\exists n_0 : \forall n \geq n_0$, $\frac{|X_n|}{n} \leq 1$. Since X_n's are independent, then by the Borel-Cantelli lemma $\sum P\left(\frac{|X_n|}{n} > 1\right) < \infty$. This means that
$$\sum_{i=1}^\infty P\left(\frac{|X_i|}{n} > 1\right) \overset{\text{iid}}{=} \sum_{i=1}^\infty P(|X_1| > n).$$
      The second term is finite iff $\mathbb{E}|X| < \infty$. When this is the case, $\mathbb{E}(X)$ exists, and by the SLLN, $\mathbb{E}(X) = a$. ■

- **Convolution**: $P * Q(A) = \int P(A-y)dQ(y)$, where P and Q are probability measure on $\mathbb{R}$, and $A - y = \{x - y : x \in A\}$.
  - If $X \sim P$ and $Y \sim Q$, then $X + Y \sim P * Q$.
  - Can show $P(A-y)$ is measurable in y by good sets principle.
  - $P * Q$ is a probability measure.
- While we can identify the distribution of a sum of RVs, we can use MGFs. However, this requires identifying a finite MGF, which might be too difficult.
- **Characteristic Function**: $\phi(t) = \phi_X(t) = \mathbb{E}[e^{itX}] = \int e^{itx}dP(x) = \int \cos(tx)dP(x) + i\int \sin(tx)dP(x)$.
  - Uses the fact that $|e^{it}| = \sqrt{\cos^2 t + \sin^2 t} = 1$.
  - e^{itx} is always bounded, since $e^{itx} = \cos(tx) + i\sin(tx)$.
  - For discrete distributions, CF is the same as the MGF, but sub in it wherever t happens.
  - **Example**: $X \sim \text{Unif}(-a, a)$.

$$\phi(t) = \int_{-a}^{a} \cos(tx)\frac{dx}{2a} + i\underbrace{\int_{-a}^{a} \sin(tx)\frac{dx}{2a}}_{=0,\text{ odd function}} = \frac{1}{2a}\cdot\left[\frac{\sin(tx)}{t}\right]_{-a}^{a} = \frac{\sin(at)}{at},\ t \neq 0. \blacksquare$$

  - Properties of CF:
    1. $|\phi(t)| \leq 1$ and always exists.
       - *Proof*; $|\mathbb{E}(e^{itX})| \leq \mathbb{E}|e^{itX}| = 1$. ■
    2. $\phi(0) = 1$.
       - *Proof*: $e^{i0x} = 1$. ■
    3. $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.
       - *Proof*: $\mathbb{E}[e^{it(X+Y)}] = \mathbb{E}[e^{itX}e^{itY}] = \mathbb{E}[e^{itX}]\mathbb{E}[e^{itY}]$. ■
    4. $\phi(-t) = \overline{\phi(t)}$.
       - *Proof*: $e^{-itX} = \overline{e^{itX}}$. ■
    5. $\phi_{-X}(t) = \mathbb{E}[e^{-itX}] = \overline{\phi(t)}$.
       - *Proof*: Direct. ■
    6. $\phi(t)$ is continuous in t.
       - *Proof*: $s \to t \implies \cos(sX) \to \cos(tX)$. Since $|\cos(sX)| \leq 1$, then by DCT, $\mathbb{E}[\cos(sX)] \to \mathbb{E}[\cos(tX)]$. Can apply similar logic to $\sin(tX)$. ■
    7. $\phi(t)$ is uniformly continuous.
       - *Proof*:

$$|\phi(t+h) - \phi(t)| = \left|\mathbb{E}[e^{i(t+h)X}] - \mathbb{E}[e^{itX}]\right| = \left|\mathbb{E}[(e^{itX})(e^{ihX}-1)]\right| \leq \mathbb{E}[|e^{itX}|\cdot|e^{ihX}-1|] \leq \left|\mathbb{E}[e^{ihX}-1]\right|.$$

       By DCT, $\left|\mathbb{E}[e^{ihX}-1]\right| \to 0$ as $h \to 0$. ■
    8. $X_n \xrightarrow{\text{d}} X \implies \phi_{X_n}(t) \to \phi_X(t)$.
       - *Proof*: $x \mapsto e^{itx}$ is bounded and continuous. The result immediately follows ■
  - $\mathbb{E}|X|^k < \infty$ iff $\phi(t)$ is k times differentiable everywhere.
    - *Proof*: Differentiating under the integral sign once, we observe that $|ie^{itx}| \leq |x|$, which is integrable by assumption. We can repeat this k times. ■
- If $a < b$ are continuity points of F of P, then $P(a,b) = \frac{1}{2\pi}\lim_{T\to\infty}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\phi(t)dt$. The limit is $F^*(b) - F^*(a)$, where $F^*(x) = \frac{F(x)-F(x-)}{2}$.
  - Proof is omitted.
  - A characteristic function uniquely determines a distribution.
- If $\phi(t)$ is absolutely integrable, then P has a density given by $f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx}\phi(t)dt$.
  - **Example**: Cauchy CF evaluation. $\int_{\mathbb{R}} e^{itX}\frac{1}{2}e^{-|x|}dx = \frac{1}{1+t^2}$. Applying inversion theorem with $\phi(t) = e^{-|t|}$ and $f(x) = \frac{1}{2}e^{-|x|}$,

$$\frac{1}{2}e^{-|x|} = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx}\frac{1}{1+t^2}dt \implies e^{-|t|} = \int_{\mathbb{R}} e^{-ixt}\frac{1}{\pi(1+x^2)}dt. \blacksquare$$

- **Levy-Cramer Continuity Theorem**: Let F_n and F be distributions with respective CFs ϕ_n and ϕ. $F_n \xrightarrow{\text{d}} F$ iff $\phi_n(t) \to \phi(t)$ for all t.
  - If the rth absolute moment is finite, then as $t \to 0$,
$$\left|\phi(t) - \sum_{k=0}^{r} \frac{(it)^k}{k!}\mathbb{E}(X^k)\right| \leq \frac{|t|^4}{r!}\mathbb{E}\left[\min\left\{2|X|^r, |t|\frac{|X|^{r+1}}{r+1}\right\}\right] = o(|t|^r).$$
    - *Proof*: Using the Taylor expansion, $e^{it} = \sum_{k=0}^{n} \frac{(it)^k}{k!} + R$, where $R = \frac{(it)^{n+1}}{(n+1)!}e^{is}$, where $0 \leq s \leq t$. Thus,
$$\left|e^{it} - \sum_{k=0}^{n} \frac{(it)^k}{k!}\right| \leq \left|\frac{(it)^{n+1}}{(n+1)!}e^{is}\right| \leq \frac{|t|^{n+1}}{(n+1)!}.$$
    In addition,
$$\left|e^{it} - \sum_{k=0}^{n} \frac{(it)^k}{k!}\right| \leq \left|e^{it} - \sum_{k=0}^{n-1} \frac{(it)^k}{k!}\right| + \frac{|t|^n}{n!} \leq \frac{2|t|^n}{n!}.$$
    Thus,
$$\left|e^{itX} - 1 - \sum_{j=1}^{k} \frac{i^j t^j X^j}{j!}\right| \leq \min\left\{\frac{|t|^{k+1}|X|^{k+1}}{(k+1)!}, \frac{2|t|^k|X|^k}{k!}\right\}.$$
    Integrating both sides creates an upper bound for $\left|\phi(t) - 1 - \sum_{j=1}^{k} \frac{i^j t^j \mathbb{E}(X^j)}{j!}\right|$. Applying DCT, we get that the quantity converges to zero. ■
- If all moments exist and $\frac{t^r \mathbb{E}|X|^r}{r!} \to 0$ as $r \to 0$ for some t, then $\phi(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!}\mathbb{E}(X^k)$.
- If $\mathbb{E}|X| < \infty$ and $\mathbb{E}(X) = \mu$, then $\phi(t) = 1 + i\mu t + o(t)$.
  - If $Var(X) = \sigma^2 < \infty$, then $\phi(t) = 1 + i\mu t - \frac{t^2(\mu^2+\sigma^2)}{2} + o(t^2)$.
- **Levy's Central Limit Theorem**: If $X_1, X_2, \ldots \overset{\text{iid}}{\sim} (\mu, \sigma^2)$, then $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2)$.
  - *Proof*: WLOG, suppose $\mu = 0$.
$$\phi_{\sqrt{n}\bar{X}_n(t)} = \mathbb{E}\left(e^{it\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i}\right) \overset{\text{iid}}{=} \prod_{i=1}^{n} \mathbb{E}\left(e^{i(t/\sqrt{n})X_i}\right) = \left[\phi(t/\sqrt{n})\right]^n = \left[1 - \frac{t^2\sigma^2}{2n} + o(n^{-1})\right]^n \to e^{-t^2\sigma^2/2}. \blacksquare$$
  - $\frac{n(\bar{X}_n - \mu)}{\sigma^2} \xrightarrow{\text{d}} \chi^2_1$.
    - *Proof*: $x \mapsto x^2$ is continuous. $Z^2 \sim \chi^2_1$ for $Z \sim \mathcal{N}(0,1)$. ■
  - $\frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \xrightarrow{\text{d}} \mathcal{N}(0,1)$, where $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.
    - *Proof*:
$$s_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right) \xrightarrow{\text{P}} \sigma^2.$$
    Applying Slutsky to the following ratio,
$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} = \frac{\sqrt{n}(\bar{X}_n - \mu)/\sigma}{s_n/\sigma} \overset{\text{Slutsky}}{\xrightarrow{\text{d}}} \frac{\mathcal{N}(0,1)}{1} = \mathcal{N}(0,1). \blacksquare$$
- **Multivariate CLT**: If $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \overset{\text{iid}}{\sim} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are vectors, then $\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{\text{d}} \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{\Sigma})$.
  - **Cramer-Wold Device**: $(X_{1n}, \ldots, X_{kn}) \xrightarrow{\text{d}} (X_1, \ldots, X_k)$ iff $\sum_{i=1}^{k} a_i X_{in} \xrightarrow{\text{d}} \sum_{i=1}^{k} a_i X_i$.
- **Lindeberg's Central Limit Theorem**: Let X_{ni} for $i = 1, \ldots, k_n$ be independent, mean-zero RVs with $\mathbb{E}(X_{ni}^2) = \tau_{ni}^2 < \infty$. Let $\sigma_n^2 = \sum_{i=1}^{k_n} \tau_{ni}^2$, and assume $\forall \epsilon > 0$,
$$\sigma_n^{-2}\sum_{i=1}^{k_n} \mathbb{E}\left[X_{ni}^2 \mathbb{1}\{|X_{ni}| > \epsilon\sigma_n\}\right] \to 0.$$
  Then, $\frac{S_n}{\sigma_n} \xrightarrow{\text{d}} \mathcal{N}(0,1)$.

- Levy's CLT is a special case of Lindeberg's CLT.
  - Can use DCT with $\sigma_n^2 = n\tau^2$.
- **Lyapunov's Condition**: $\sigma_n^{-(2+\delta)} \sum_{i=1}^{k_n} \mathbb{E}|X_{ni}|^{2+\delta} \to 0$, where $\delta > 0$ is arbitrarily small.
  - Useful when incomplete integrals are difficult to work with.
- **Uniform Asymptotic Negligibility**, or **UAN**: $\max\left\{\frac{\tau_{ni}^2}{\sigma_n^2}\right\} \to 0$.
  - If UAN fails, Lindeberg's condition fails.
  - If UAN holds, Lindeberg's condition is necessary for CLT.

- **Asymptotic Normality**: If μ_n and $\sigma_n > 0$ are sequences such that $\frac{X_n - \mu_n}{\sigma_n} \xrightarrow{\text{d}} \mathcal{N}(0,1)$, then $X_n \sim AN(\mu_n, \sigma_n^2)$.
  - This does not mean that $X_n \sim (\mu_n, \sigma_n)$. X_n may not actually have any moments.
- **Delta Method**: Suppose $X_n \sim AN(\mu)n, c_n^2\sigma^2)$ and $g'(\mu) \neq 0$ for some g. Then, $g(X_n) \sim AN\left(g(\mu), c_n^2\sigma(g'(\mu))^2\right)$.
  - *Proof*: Expand $g(x)$ around μ.

$$g(x) = g(\mu) + (x-\mu)g'(\mu) + o(|x-\mu|).$$

Since $|X_n - \mu| = O_p(c_n) \implies o(|X_n - \mu|) = o_p(c_n)$. Thus,

$$g(X_n) = g(\mu) + (X_n - \mu_n)g'(\mu) + o_p(c_n) \implies \frac{g'(\mu)(X_n - \mu)}{c_n} + o_p(1) \xrightarrow{\text{d}} \mathcal{N}\left(0, \sigma^2((g'(\mu)^2)\right). \blacksquare$$

  - **Example**: $U_i \overset{\text{iid}}{\sim} \text{Unif}(0,1)$. Find the asymptotic distribution of $X_n = \left(\prod_{i=1}^n U_i\right)^{1/n}$.

$$\mathbb{E}[\log U] = \int_0^1 \log u du = \cdots = -1; \mathbb{E}[(\log U)^2] = \int_0^1 (\log u)^2 du = \cdots = 2.$$

Let $Z_n = \log X_n = \frac{1}{n}\sum_{i=1}^n \log(U_i) \implies X_n = e^{Z_n}$. $Z_n \sim AN(-1, \frac{1}{n})$. Let $g(x) = e^x \implies g'(x) = e^x$, which is nonzero at $\mu = -1$. Therefore,

$$X_n \sim AN\left(e^{-1}, \frac{(e^{-1})^2}{n}\right) = AN\left(e^{-1}, \frac{e^{-2}}{n}\right). \blacksquare$$

  - **Multivariate Delta Method**: $g'(X_n) \sim AN_k\left(g(\mu), c_n^2(g'(\mu))\Sigma(g'(\mu))^T\right)$, where $g'(\mu) \neq 0$, and $X_n \sim AN_k(\mu, c_n^2\Sigma)$.
    - **Example**: Sample variance, $s_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2$.

WLOG, suppose $\mu = 0$.

$$\begin{pmatrix} \frac{1}{n}\sum_{i=1}^n X_i \\ \frac{1}{n}\sum_{i=1}^n X_i^2 \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix}\right).$$

Thus, $s_n^2 = g(x,y) = y - x^2$. Therefore, $g'(x,y)^T = (-2x, 1)$, and $g'(0, \sigma^2) = (0,1)$. Multiply matrices to get

$$s_n^2 \sim AN\left(\sigma^2, \frac{\mu_4 - \sigma^4}{n}\right). \blacksquare$$

9.8 Conditional Expectations

Return to Table of Contents

- **Absolutely Continuous**: Measure ν is absolutely continuous wrt measure μ, denoted as $\nu \ll \mu$, if $\forall A \in \mathcal{A}$, $\mu(A) = 0 \implies \nu(A) = 0$.
  - Evety μ-null set is also a ν-null set.
  - **Mutually Absolutely Continuous**: $\nu \ll \mu$ and $\mu \ll \nu$.
  - ν defined by $\nu(A) = \int_A f d\mu$ is a measure and $\nu \ll \mu$.
  - **Example**: $\mu = \lambda$, and $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ on $\mathbb{R}$. $\nu(A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ is the Standard Normal probability measure. ■
  - If $f(x) > 0$ almost everywhere (a.e.) wrt μ and $\nu \ll \mu$, then $\mu \ll \nu$.
    - This is because $\mu(A) > 0 \implies \int_A f d\mu > 0$.
  - For any ν and μ, $\nu \ll \nu + \mu$, and $\mu \ll \nu + \mu$.
- Let $(\Omega, \mathcal{A}, \mu)$ be a measurable space, f a probability density function wrt μ, where $P(A) = \int_A f d\mu$. Then, for any RV X, $\mathbb{E}_P(X) = \int X f d\mu$.
  - This lets us integrate with an older measure.
  - **Example**: $X \sim \mathcal{N}(0,1)$. $\mathbb{E}(X^4) = \int x^4 d\Phi$, but we often use $\int_{-\infty}^{\infty} x^4 \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$. ■
  - *Proof*: This trivially holds for indicators. Then, suppose X is simple. $X = \sum_{i=1}^k c_i \mathbb{1}_{A_i}$.

$$\mathbb{E}_P(X) = \sum_{i=1}^k c_i P(A_i) = \sum_{i=1}^k c_i \int_{A_i} f d\mu = \int \left[\sum_{i=1}^k c_i \mathbb{1}_{A_i} \right] f d\mu = \int X f d\mu.$$

    Now, let $X \geq 0$. Take $X_n \uparrow X$, $X_n \geq 0$ is simple. $X_n f \uparrow X f$, so by MCT, $\mathbb{E}_P(X) = \lim_{n\to\infty} \mathbb{E}_P(X_n) = \int X f d\mu$.
    Lastly, Suppose X is measurable. Since $f \geq 0$ a.e. $[\mu]$, $(Xf)^+ = X^+ f$, and $(Xf)^- = X^- f$.

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X-) = \int X^+ f d\mu - \int X^- f d\mu = \int X f d\mu. \blacksquare$$

- **Radon-Nikodym Theorem**: Let $(\Omega, \mathcal{A})$ be measurable, μ a σ-finite measure and ν a measure. If $\nu \ll \mu$, then there exists a Borel-measurable function $f : \Omega \to \mathbb{R}$ such that $\nu(A) = \int_A f d\mu$ for all $A \in \mathcal{A}$. Furthermore f is unique a.e. $[\mu]$.
  - *Proof*: Mostly omitted. For uniqueness, if f and g are two Radon-Nikodym derivatives, then $\int_A f d\mu - \int_A g d\mu \implies f = g$. ■
  - We cannot drop σ-finiteness of μ.
- **Conditional Expectation**: Let $(\Omega, \mathcal{A}, P)$ be a probability space, X an integrable RV. Let $\mathcal{E}$ be a sub σ-field of $\mathcal{A}$. Let X^* be an $\mathcal{E}$-measurable integrable RV such that $\int_E X^* dP = \int_E X dP$ for all $E \in \mathcal{E}$. X^* is the conditional expectation of X given $\mathcal{E}$, denoted as $X^* = \mathbb{E}(X|\mathcal{E})$.
  - The conditional expectation exists and is unique up to null sets and/or P.
  - Conditional expectations are RVs, not deterministic quantities in general.
- Properties of conditional expectations. Note that all of these are a.s. $[P]$:
  - $\mathbb{E}(X|\mathcal{A}) = X$ and $\mathbb{E}(X|\{\varnothing, \Omega\}) = \mathbb{E}(X)$.
    - *Proof*: $\int_A X dP = \int_A X dP \implies \mathbb{E}(X|\mathcal{A}) = X$. ■
  - $X \geq Y \implies \mathbb{E}(X|\mathcal{E}) \geq \mathbb{E}(Y|\mathcal{E})$.
    - *Proof*: First, assume $X \geq 0$. $\nu(A) := \int_A X dP$ is a measure, where $\nu \ll P$, so by Radon-Nikodym, $\mathbb{E}(X|\mathcal{E}) \geq 0$. Apply to $Y - X$. ■
  - $\mathbb{E}[\mathbb{E}(X|\mathcal{E})] = \mathbb{E}(X)$. In particular, $\mathbb{E}[\mathbb{E}(X|\mathcal{E}_2)|\mathcal{E}_1] = \mathbb{E}(X|\mathcal{E}_1)$ for $\mathcal{E}_1 \subset \mathcal{E}_2$.
    - *Proof*: The first part follows from the definition of conditional expectation with $\Omega = E$. Now, $\mathbb{E}(X|\mathcal{E}_1)$ is $\mathcal{E}_1$-measurable, so for any $E \in \mathcal{E}_1 \subset \mathcal{E}_2$,

$$\int_E \mathbb{E}[\mathbb{E}(X|\mathcal{E}_2)|\mathcal{E}_1] dP = \int_E \mathbb{E}(X|\mathcal{E}_2) dP = \int_E X^* dP = \int_E X dP \implies \mathbb{E}[\mathbb{E}(X|\mathcal{E}_2)|\mathcal{E}_1] = \mathbb{E}(X|\mathcal{E}_1). \blacksquare$$

- $\mathbb{E}(X|\mathcal{E}) = X$ iff X is $\mathcal{E}$-measurable.
  - *Proof*: If X is $\mathcal{E}$-measurable, then X itself satisfies the requirement in the definition of conditional expectation. Now, if $X = \mathbb{E}(X|\mathcal{E})$ a.s., then $\mathbb{E}(X|\mathcal{E})$ is $\mathcal{E}$-measurable by definition, so X is a.s. equal to an $\mathcal{E}$-measurable function.
- $\mathbb{E}(a_1X_1 + a_2X_2|\mathcal{E}) = a_1\mathbb{E}(X_1|\mathcal{E}) + a_2\mathbb{E}(X_2|\mathcal{E})$.
  - *Proof*:

$$\int_E (a_1X_1^* + a_2X_2^*)dP = a_1\int_E X_1^* dP + a_2\int_E X_2^* dP = a_1\int_E X_1 dP + a_2\int_E X_2 dP = \int_E (a_1X_1 + a_2X_2)dP. \blacksquare$$

- $\mathbb{E}(XY|\mathcal{E}) = X\mathbb{E}(Y|\mathcal{E})$ if X is $\mathcal{E}$-measurable.
  - *Proof*: We need to show that $\int_E XY^* dP = \int_E XY dP$. First, suppose $X = \mathbb{1}_A$, $A \in \mathcal{E}$. Then,

$$\int_E XY^* dP = \int_{A\cap E} Y^* dP = \int_{A\cap E} Y dP = \int_E XY dP.$$

    Now, suppose X is simple. The result holds by linearity of conditional expectations.
    Next, suppose X is $\mathcal{E}$-measurable. Take simple X_n that are $\mathcal{E}$-measurable such that $|X_n| \le |X|$ and $X_n \to X$. $|X_nY| \le |X|\cdot|Y|$, $X_nY_n \to XY$, $|X_nY^*| \le |X|\cdot|Y^*|$, and $X_nY^* \to XY^*$. Hence,

$$\int_E XY^* dP = \lim_{n\to\infty}\int_E X_nY^* dP = \lim_{n\to\infty}\int_E X_nY dP = \int_E XY dP.$$

    For general X, can apply $X = X^+ - X^-$. $\blacksquare$
- $\mathbb{E}(X_n|\mathcal{E}) \uparrow \mathbb{E}(X|\mathcal{E})$ $X_n \ge 0$, and $X_n \uparrow X$.
  - *Proof*: $X_n^* = \mathbb{E}(X_n|\mathcal{E}) \le X^* = \mathbb{E}(X|\mathcal{E})$, where X_n^* is increasing, so a limit must exist. Let $\tilde{X} \le X^*$ be this limit.

$$\mathbb{E}(\tilde{X}) = \lim_{n\to\infty}\mathbb{E}(X_n^*) \overset{\text{MCT}}{=} \lim_{n\to\infty}\mathbb{E}(X_n) = \mathbb{E}(X). \blacksquare$$

- $\mathbb{E}(X_n|\mathcal{E}) \to \mathbb{E}(X|\mathcal{E})$, where $|X_n| \le Y$, $\mathbb{E}(Y) < \infty$, and $X_n \to X$.
  - *Proof*: WLOG, suppose $X \ge 0$ (can take $X = X^+ - X^-$). Let $Y_n = \sup_{k\ge n} X_k \downarrow X$ and $Z_n = \inf_{k\ge n} X_k \uparrow X$. Thus,

$$\mathbb{E}(X_n|\mathcal{E}) \le \mathbb{E}(Y_n|\mathcal{E}) \downarrow \mathbb{E}(X|\mathcal{E}), \text{ and } \mathbb{E}(X_n|\mathcal{E}) \ge \mathbb{E}(Z_n|\mathcal{E}) \uparrow \mathbb{E}(X|\mathcal{E}) \implies \mathbb{E}(X_n|\mathcal{E}) \to \mathbb{E}(X|\mathcal{E}). \blacksquare$$

- $\mathbb{E}|\mathbb{E}(X|\mathcal{E})| \le \mathbb{E}|X|$.
  - *Proof*: $X \le |X| \implies \mathbb{E}(X|\mathcal{E}) \le \mathbb{E}(|X||\mathcal{E})$. In addition, $\mathbb{E}(-X|\mathcal{E}) \le \mathbb{E}(|X||\mathcal{E})$. Combining, $|\mathbb{E}(X|\mathcal{E})| \le \mathbb{E}(|X||\mathcal{E})$. Integrating wrt P yields the answer. $\blacksquare$
- If $\mathbb{E}(X^2) < \infty$, then for any $\mathcal{E}$-measurable function Y, $\mathbb{E}\left(X - \mathbb{E}(X|\mathcal{E})\right)^2 \le \mathbb{E}(X-Y)^2$.
  - *Proof*: $\mathbb{E}(X^2) < \infty \implies \mathbb{E}(|X|) < \infty$, which means conditional expectation is well-defined.

$$\begin{aligned}
\mathbb{E}(X-Y)^2 =& \mathbb{E}\left[X - \mathbb{E}(X|\mathcal{E}) - (Y - \mathbb{E}(X|\mathcal{E}))\right]^2 \\
=& \mathbb{E}\left[X - \mathbb{E}(X|\mathcal{E})\right]^2 + \mathbb{E}\left[Y - \mathbb{E}(X|\mathcal{E})\right]^2 + 2\mathbb{E}\left[(X - \mathbb{E}(X|\mathcal{E}))(Y - \mathbb{E}(X|\mathcal{E}))\right] \\
=& \mathbb{E}\left[X - \mathbb{E}(X|\mathcal{E})\right]^2 + \mathbb{E}\left[Y - \mathbb{E}(X|\mathcal{E})\right]^2 + 2\mathbb{E}\left\{\mathbb{E}\left[(X - \mathbb{E}(X|\mathcal{E}))(Y - \mathbb{E}(X|\mathcal{E}))\right]|\mathcal{E}\right\} \\
=& \mathbb{E}\left[X - \mathbb{E}(X|\mathcal{E})\right]^2 + \mathbb{E}\left[Y - \mathbb{E}(X|\mathcal{E})\right]^2 + 2\mathbb{E}\left\{(\mathbb{E}(X|\mathcal{E}) - Y)\mathbb{E}[(X - \mathbb{E}(X|\mathcal{E}))|\mathcal{E}]\right\} \\
=& \mathbb{E}\left(X - \mathbb{E}(X|\mathcal{E})\right)^2 + \mathbb{E}[\mathbb{E}(X|\mathcal{E}) - Y]^2 \ge \mathbb{E}\left(X - \mathbb{E}(X|\mathcal{E})\right)^2. \blacksquare
\end{aligned}$$

- $\mathbb{E}[X\mathbb{E}(Y|\mathcal{E})] = \mathbb{E}[\mathbb{E}(X|\mathcal{E})Y]$.
  - *Proof*:

$$\mathbb{E}(XY^*) = \mathbb{E}[\mathbb{E}(XY^*)|\mathcal{E}] = \mathbb{E}(X^*\mathbb{E}[Y|\mathcal{E})] = \mathbb{E}(X^*Y^*).$$

    Similarly, $\mathbb{E}(X^*Y) = \mathbb{E}(X^*Y^*) \implies \mathbb{E}(XY^*) = \mathbb{E}(X^*Y)$. $\blacksquare$

- If $\mathcal{E}$ is a σ-field generated by a finite or countable partition, it can be easier to find $X|\mathcal{E}$.
  - **Example**: $\Omega = [0,1]$, $X(\omega) = \omega$, $P = \lambda$, $X \sim \text{Unif}(0,1)$. $\mathcal{E} = \sigma\langle X^{-1}(0,\frac{1}{4}], X^{-1}(\frac{1}{4},\frac{1}{2}], X^{-1}(\frac{1}{2},\frac{3}{4}], X^{-1}(\frac{3}{4},1]\rangle$. Calculate $X^* = \mathbb{E}(X|\mathcal{E})$.

    X^* is σ-measurable. This is because X is constant on each interval given in $\mathcal{E}$.

$$c_iP(A_i) = \int_{A_i} X dP \implies c_i = \frac{\int_{A_i} X dP}{P(A_i)} \implies \mathbb{E}(X|\mathcal{E}) = c_i\mathbb{1}\{\omega \in A_i\}.$$

Using the fact that $\int_E X^* dP = \int_E X dP$ for all $E \in \mathcal{E}$, where $E = (0, \frac{1}{4}], (\frac{1}{4}, \frac{1}{2}], (\frac{1}{2}, \frac{3}{4}], (\frac{3}{4}, 1]$,

$$c_1 = \frac{\int_0^{1/4} x dx}{1/4} = \frac{1/32}{1/4} = \frac{1}{8}.$$

We repeat on all c_i to get $(c_1, c_2, c_3, c_4) = \left(\frac{1}{8}, \frac{3}{8}, \frac{5}{8}, \frac{7}{8}\right)$. ■

- **Example**: $\Omega = \mathbb{R}$, $\mathcal{R}$, $P = \mathcal{N}(0,1)$, and $X(\omega) = \omega$. $\mathcal{E} = \sigma\langle |X| \rangle = \{|X|^{-1}(B) : B \in \mathcal{R}\} = \{B \cup (-B) : B \in \mathcal{R}_{[0,\infty)}\}$. Show that $\mathbb{E}(e^X|\mathcal{E}) = \frac{e^X + e^{-x}}{2} =: \cosh(X)$.

  f is $\mathcal{E}$-measurable iff f is Borel measurable and symmetric about 0.
  Claim $Y = \cosh X$. This is symmetric and Borel measurable, so it is $\mathcal{E}$-measurable. We nts that $\int_{B\cup(-B)} \frac{e^x + e^{-x}}{2} dP(x) = \int_{B\cup(-B)} e^x dP$.

$$\begin{aligned}
\int_{B\cup(-B)} \frac{e^x + e^{-x}}{2} dP(x) =& \frac{1}{2}\int_B e^x \phi(x)dx + \frac{1}{2}\int_{-B} e^x\phi(x)dx + \frac{1}{2}\int_B e^{-x}\phi(x)dx + \frac{1}{2}\int_{-B} e^{-x}\phi(x)dx \\
=& \frac{1}{2}\int_B e^x \phi(x)dx + \frac{1}{2}\int_{-B} e^x\phi(x)dx + \frac{1}{2}\int_B e^x\phi(x)dx + \frac{1}{2}\int_{-B} e^x\phi(x)dx \\
=& \int_B e^x\phi(x)dx + \int_{-B} e^x\phi(x)dx = \int_{B\cup(-B)} e^x\phi(x)dx. \ \blacksquare
\end{aligned}$$

- To calculate $\mathbb{E}(X|\mathcal{E})$ for a general integrable X and a general σ-field $\mathcal{E}$, we first need to see if we can get a sequence of increasing σ-fields $\mathcal{E}_n$ such that $\mathcal{E} = \sigma\langle \bigcup_{n=1}^\infty \mathcal{E}_n \rangle$, then $\mathbb{E}(X|\mathcal{E}) = \lim_{n\to\infty} \mathbb{E}(X|\mathcal{E}_n)$, and choosing a $\mathcal{E}_n$ generated by a finite partition, then compution $\mathbb{E}(X|\mathcal{E}_n)$.

- **Conditional Probability**: $P(A|\mathcal{E}) = \mathbb{E}(X|\mathcal{E})$, where $X = \mathbb{1}_A$.
  - Properties (these hold a.s.):
    - $P(A|\mathcal{E}) = \mathbb{1}(A)$ if $\mathcal{E} = \mathcal{A}$.
    - $0 \le P(A|\mathcal{E}) \le 1$.
    - $P\left(\bigcup_{n=1}^\infty A_n|\mathcal{E}\right) = \sum_{n=1}^\infty P(A_n|\mathcal{E})$ for disjoint A_n.
    - $A \subset B \implies P(A|\mathcal{E}) \le P(B|\mathcal{E})$.
    - $P\left(\bigcup_{i=1}^n A_i|\mathcal{E}\right) = \sum_{i=1}^n P(A_i|\mathcal{E}) - \sum_{i<j} P(A_i \cap A_j|\mathcal{E}) + \dots$.
    - $A_n \uparrow A \implies P(A_n|\mathcal{E}) \uparrow P(A|\mathcal{E})$.
    - $A_n \downarrow A \implies P(A_n|\mathcal{E}) \downarrow P(A|\mathcal{E})$.
    - $A_n \downarrow A \implies P(A_n|\mathcal{E}) \downarrow P(A|\mathcal{E})$.
    - $P(A = 1) \implies P(A|\mathcal{E}) = 1$.
    - $P(A = 0) \implies P(A|\mathcal{E}) = 0$.
  - **Regular Conditional Probability**: Q such that $\omega \mapsto Q(\omega, A)$ is $\mathcal{E}$-measurable, $\int_B Q(\omega, A)dP(\omega) = P(A \cap B) = \int_B \mathbb{1}(A)(\omega)dP(\omega)$ for every $B \in \mathcal{E}$, and $Q(\omega, \cdot)$ is a probability for every ω, where P is a probability measure.
    - Q is a transition.
    - $\omega \mapsto Q(\omega, A)$ is $\mathcal{E}$-measurable and $\int_B Q(\omega, A)dP(\omega) = P(A \cap B) = \int_B \mathbb{1}(A)(\omega)dP(\omega)$ for every $B \in \mathcal{E}$ is equivalent to saying that $Q(\omega, A) = P(A|\mathcal{E})(\omega)$.
    - Suppose X is an RV and $\mathcal{E}$ is a sub σ-field of $\mathcal{A}$. Then, there exists $Q(\omega, \cdot)$ such that $\omega \mapsto Q(\omega, B)$ is measurable and $P(X \in B|\mathcal{E})$ for all B, $B \mapsto Q(\omega, B)$ is a probability measure for all ω.

- Let X be a random variable on $(\Omega, \mathcal{A}, P)$, and $\mathcal{E}$ be a sub σ-field. Let Q be the conditional distribution of X given $\mathcal{E}$. If $\phi : \mathbb{R} \to \mathbb{R}$ is measurable and integrable, then

$$E[\phi(X)|\mathcal{E}] = \int \phi(x)Q(\omega, dx) \text{ a.s.}$$

  - *Proof*: First, let $\phi = \mathbb{1}(B)$. Then,

$$\mathbb{E}[\mathbb{1}\{X \in B\}|\mathcal{E}) = P(X \in B|\mathcal{E}) = Q(\cdot, B).$$

  By linearity of conditional expectations, this also holds to simple measurable functions.
  For nonnegative ϕ, consider a sequence of simple and measurable $\phi_n \uparrow \phi$. Then, by MCT, result holds.
  For general integrable ϕ, results hold for $\phi = \phi^+ - \phi^-$. ■

- **Conditional Expectation (RVs)**: $\mathbb{E}[\phi(X)|Y](y) = \mathbb{E}[\phi(X)|Y = y] = \int \phi(x)R(y, dx)$, where $R(y, \cdot)$ is a transition probability measure on $\mathbb{R}$.

- **Bayes Theorem**: $p(x|y) = \frac{f(x)g(y|x)}{\int f(t)g(y|t)d\mu(t)}$.

9.9 Exam Problems

Return to Table of Contents

Midterm 2025

- If S stands for all infinite sequences of 0 and 1, and T stands for all finite sequences of 0 and 1 of arbitrary length, then S is uncountable and T is countably infinite.
- If $\mathcal{F}_n$ is a sequence of (not strictly) increasing σ-fields on a sample space Ω, then $\bigcup_{n=1}^{\infty} \mathcal{F}_n$ is a field, but may not be a σ-field.
- On a given sample space, every finite field is a σ-field.
- **Example**: Let $\Omega = \{a, b, c, d, e\}$ and $\mathcal{C} = \{\{a, c, e\}, \{a, b, d\}\}$.
  1. Find the partition generated by $\mathcal{C}$.
  2. Find the σ-field generated by $\mathcal{C}$.

  1. Let $A = \{a, c, e\}$ and $B = \{a, b, d\}$.
  $$\mathcal{P} = \{A \cap B, A \cap B^c, A^c \cap B, A^c \cap B^c\} = \{\{a\}, \{c, e\}, \{b, d\}\}.$$
  Note that we drop $\varnothing = A^c \cap B^c$.
  2. $\sigma\langle\mathcal{C}\rangle$ consists of all unions from $\mathcal{P}$.
  $$\sigma\langle\mathcal{C}\rangle = \{\varnothing, \{a\}, \{c, e\}, \{b, d\}, \{a, b, d\}, \{a, c, e\}, \{b, c, d, e\}, \Omega\}. \blacksquare$$
- **Example**: Let $X_1, X_2, \ldots$ be independent RVs on probability space $(\Omega, \mathcal{A}, P)$ with the common CDF $F(x) = P(X_i \leq x) = (1 + e^{-x})^{-1}$ for all $x \in \mathbb{R}$.
  1. What is $P(X_i \in \mathbb{I} \text{ for all } i)$?
  2. Let Let $A_n = \{X_n \leq -c \log n\}$. show that $P(\limsup_{n\to\infty} A_n) = 0$ if $c > 1$, 1 otherwise.
  3. Show that $P\left(\liminf_{n\to\infty} \frac{X_n}{\log n} = -1\right) = 1$.
  4. Compute $\limsup_{n\to\infty} \frac{X_n}{\log n}$.

  1. Since X_i is continuous,
  $$P(X_i \in \mathbb{Q}) = 0 \implies P\left(\bigcap_{i=1}^{\infty} \{X_i \in \mathbb{Q}^c\}\right) = \lim_{n\to\infty} P\left(\bigcap_{i=1}^{n} \{X_i \in \mathbb{Q}^c\}\right) \overset{\perp}{=} \lim_{n\to\infty} \prod_{i=1}^{n} P(X_i \in \mathbb{I}) = 1.$$
  2.
  $$P(A_n) = P(X_n \leq -c \log n) = (1 + e^{-c \log n})^{-1} = (1 + n^{-c})^{-1} \sim n^{-c}.$$
  This means that $\sum_{i=1}^{\infty} P(A_i) < \infty$ if $c > 1$, otherwise (via convergence of series results). Thus, by the Borel-Cantelli lemmas, $P(\limsup_{n\to\infty} A_n) = 0$ if $c > 1$, 1 otherwise.
  3. $\liminf_{n\to\infty} \frac{X_n}{\log n} = -1 \implies \forall \epsilon > 0$, $\frac{X_n}{\log n} < -1 + \epsilon$ i.o., and $\frac{X_n}{\log n} < -1 - \epsilon$ finitely. $-1 + \epsilon \equiv c < 1$, and $-1 - \epsilon \equiv c > 1$. Thus,
  $$P\left(\frac{X_n}{\log n} < -1 + \epsilon \text{ i.o.}\right) = 1, \text{ and } P\left(\frac{X_n}{\log n} < -1 - \epsilon \text{ i.o.}\right) = 0 \implies P\left(\liminf_{n\to\infty} \frac{X_n}{\log n} = -1\right) = 1.$$
  4. Since $X_n \overset{d}{=} -X_n$ by symmetry,
  $$\limsup_{n\to\infty} \frac{X_n}{\log n} = -\liminf_{n\to\infty} \frac{(-X_n)}{\log n} = -(-1) \text{ a.s.} \implies \limsup_{n\to\infty} \frac{X_n}{\log n} = 1 \text{ a.s.} \blacksquare$$
- **Example**: Let X be a positive RV on probability space $(\Omega, \mathcal{A}, P)$.
  1. Show that $\mathbb{E}(X \log X)$ exists and that $\mathbb{E}(X \log X) > \infty$.
  2. Let X_n be a sequence of positive RVs, $X_n \to X$ pointwise. Show that $\mathbb{E}(X \log X) \leq \liminf_{n\to\infty} \mathbb{E}(X_n \log X_n)$.
  3. If a sequence of positive RVs X_n satisfies the condition that $\sup\{\mathbb{E}(X_n^2) : n \geq 1\} < \infty$, then show that $Y_n = X_n \log X_n$ is uniformly integrable.

1. Let $f(x) = x\log x \implies f'(x) = \log x + 1$, and $f''(x) = \frac{1}{x} > 0$. From this, $x = \frac{1}{e}$ is the unique minimum of $f(x)$. At this value, $f(e^{-1}) = -\frac{1}{e} \implies x\log x \geq -\frac{1}{e} > -\infty$, therefore $\mathbb{E}(X\log X)$ exists, and is $> -\infty$.
2. Let $Y = -\frac{1}{e}$. Since $X_n \to X$, and $X_n \geq Y$, where $\mathbb{E}(Y) > \infty$, then by Fatou's Lemma,
$$\liminf_{n\to\infty} \mathbb{E}(X_n \log X_n) \geq \mathbb{E}(X\log X)$$
3. WLOG, let $C > \frac{1}{e}$.
$$\int_{|Y_n|>C} |Y_n| dP = \int_{Y_n>C} Y_n dP = \int_{X_n>B} x_n \log x_n dP,$$
where $B\log B = C$. Note that $B \to \infty$ iff $C \to \infty$.
$$\int_{X_n>B} x_n \log x_n dP = \int_{X_n>B} x_n^2 \cdot \frac{\log x_n}{x_n} dP \leq \frac{\log B}{B} \int_{X_n>B} x_n^2 dP \leq \frac{\log B}{B} \cdot \mathbb{E}(X_n^2).$$
Since $\mathbb{E}(X_n^2) < \infty$, then as $B \to \infty$, $\frac{\log B}{B} \cdot \mathbb{E}(X_n^2) \to 0 \implies \sup_{n\geq 1} \int_{|Y_n|>C} |Y_n| dP \to 0$ as $C \to \infty$. ■

Final 2023

- **Example**: Let Ω be a set and $\mathcal{C}$ be a class of subsets of Ω. Let $x, y \subset \Omega$ be fixed, where $x \neq y$. Assume that whenever $C \in \mathcal{C}$, either $\{x, y\} \subset C$ or $\{x, y\} \subset C^c$. Show that for any $A \in \sigma\langle\mathcal{C}\rangle$, either $\{x, y\} \subset A$ or $\{x, y\} \subset A^c$.

  Apply the good sets principle. Let $\mathcal{A} = \sigma\langle\mathcal{C}\rangle$, and $\mathcal{G} = \{A \in \mathcal{A} : \{x, y\} \subset A \text{ or } \{x, y\} \in A^c\}$. By construction, $\mathcal{C} \subset \mathcal{G}$. We need to show that $\mathcal{G}$ is a σ-field.

  1. $\varnothing, \Omega \in \mathcal{G}$: Trivially, $\{x, y\} \in \Omega$, and $\varnothing^c = \Omega$ ✓
  2. Closed under countable union/intersection: Suppose $A_n \in \mathcal{G}$. If one of A_n contains $\{x, y\}$, then $\cup A_n$ will as well. If none of A_n contain $\{x, y\}$, then $(\cup A_n)^c = \cap A_n^c$ will contain $\{x, y\}$ ✓
  3. Closed under complementation: trivial ✓

  Therefore, $\mathcal{G} \supset \mathcal{C}$. ■

- **Example**:

  1. If $a < b$ are real numbers, show that there exists a sequence of bounded, continuous $f_n(x)$ such that $f_n(x) \to \mathbb{1}_{(a,b]}(x)$ for all x.
  2. Let X and Y be RVs on $(\Omega, \mathcal{A}, P)$ such that for every bounded, continuous f and g on $\mathbb{R}$, $f(X)$ and $g(Y)$ are independent RVs. Show that
  $$P(a_1 < X \leq b_1, a_2 < Y \leq b_2) = P(a_1 < X \leq b_1)P(a_2 < Y \leq b_2) \ \forall a_1 < b_1, a_2 < b_2.$$
  3. Let X_n and Y_n be two sequences of RVs on $(\Omega, \mathcal{A}, P)$ such that X_n and Y_n are independent for all n, and that $X_n \to \infty$ and $Y_n \to Y$ a.s.

  1. Define $f(x) = \mathbb{1}_{(a,b]}$ and $f_n(x) = \begin{cases} 1 & , a + \frac{1}{n} < x \leq b \\ 0 & , x \leq a \text{ or } x \geq b + \frac{1}{n} \end{cases}$, which are all positive, bounded and continuous. If $x \leq a$ or $x > b$, then $f_n(x) = 0 = f(x)$. If $x \in (a, b]$, then $f(x) = 1$. For large n such that $\frac{1}{n} \approx 0$, if $x \in (a, b]$, then $f_n(x) = 1$. Similarly, for large n, if $x > b$, then $f_n(x) = 0$. Therefore, we can approximate $f(x)$ with $f_n(x)$.
  2. $f_n \leftarrow (a_1, b_1]$ and $g_n \leftarrow (a_2, b_2]$. Thus, $f_n(X)g_n(Y) \to \mathbb{1}\{a_1 < X \leq b_1, a_2 < Y \leq b_2\}$.
  $$P(a_1 < X \leq b_1, a_2 < Y \leq b_2) = \lim_{n\to\infty} \mathbb{E}\left[f_n(X)g_n(Y)\right]$$
  Applying independence,
  $$= \lim_{n\to\infty} \mathbb{E}[f_n(X)]\mathbb{E}[g_n(Y)] = P(a_1 < X \leq b_1)P(a_2 < Y \leq b_2).$$
  3.
  $$\mathbb{E}[f(X_\infty)g(Y_\infty)] = \lim_{n\to\infty} \mathbb{E}[f(X_n)g(Y_n)] = \lim_{n\to\infty} \mathbb{E}[f(X_n)]\mathbb{E}[g(Y_n)] = f(X_\infty)g(Y_\infty). \ ■$$

- **Example**: Let $A_1, A_2, \dots$ be events with $P(A_i \cap A_j) = P(A_i)P(A_j)$ for all $i \neq j$. Let $N_n = \sum_{i=1}^n \mathbb{1}_{A_i}$ and $m_n = \sum_{i=1}^n P(A_i)$. If $\sum_{i=1}^\infty P(A_i) = \infty$, show that $\frac{N_n}{m_n} \xrightarrow{\text{p}} 1$.

  $$\mathbb{E}\left(\frac{N_n}{m_n}\right) = 1, \text{ and } Var\left(\frac{N_n}{m_n}\right) = \frac{1}{m_n^2}\sum_{i=1}^n Var(\mathbb{1}_{A_i}) = \frac{1}{m_n^2}\sum_{i=1}^n P(A_i)(1-P(A_i)) \leq \frac{m_n}{m_n^2} = \frac{1}{m_n} \downarrow 0.$$

  Therefore, by Chebyshev, $\frac{N_n}{m_n} \xrightarrow{\text{p}} 1$, since

  $$P\left(\left|\frac{N_n}{m_n} - 1\right| \geq \epsilon\right) \leq \frac{1}{\epsilon^2} \cdot \frac{Var(N_n)}{m_n^2)} \to 0. \blacksquare$$

- **Example**: Compute $\lim_{n\to\infty} \sum_{k=0}^{2n} e^{-2n}\frac{2^k n^k}{k!}$.

  Let $X_n \sim \text{Pois}(2n)$, and $Z_i \sim \text{Pois}(2)$.

  $$\sum_{k=0}^{2n} e^{-2n} \cdot \frac{2^k n^k}{k!} = P(X_n \leq 2n) = P(Z_1 + \cdots + Z_n \leq 2n) = P\left(\sum_{i=1}^n (Z_i - 2) \leq 0\right).$$

  Applying the CLT,

  $$P\left(\frac{\sum_{i=1}^n (Z_i - 2)}{\sqrt{2n}} \leq 0\right) \to P\left(\mathcal{N}(0,1) \leq 0\right) = \frac{1}{2}. \blacksquare$$

- **Example**: Let $Z_1, Z_2, \dots \overset{\text{iid}}{\sim} (0,1)$. Show that $\sum_{i=1}^n \frac{i^{-1/2}Z_i}{\sqrt{\log n}} \xrightarrow{\text{d}} \mathcal{N}(0,1)$, using the fact that $\sum_{i=1}^n i^{-1} \sim \log n$.

  Let $c_{in} = i^{-1/2}$ for $i = 1, \dots, k_n = n$. $\max_i c_{in}^2 = 1$, and $\sum_{i=1}^n c_{in}^2 = \sum_{i=1}^n \frac{1}{i} \sim \log n$. As a result,

  $$\frac{\max_i c_{in}^2}{\sum_{i=1}^n c_{ni}^2} \sim \frac{1}{\log n} \to 0.$$

  By a corollary for iid sequences, Lindeberg's condition holds for $\{i^{-1/2}Z_i,\, i = 1, \dots, n\}$. Thus, by the CLT,

  $$\frac{\sum_{i=1}^n c_{in}Z_i}{\sqrt{\sum_{i=1}^n c_{in}^2}} \xrightarrow{\text{d}} \mathcal{N}(0,1) \implies \frac{\sum_{i=1}^n i^{-1/2}Z_i}{\sqrt{\log n}} \xrightarrow{\text{d}} \mathcal{N}(0,1) \text{ by Slutsky. } \blacksquare$$

- **Example**: Let $\Omega = (0, \infty)$ be the sample space with the Borel σ-field on it, and let P stand for the standard exponential distribution, with density $p(\omega) = e^{-\omega}$, $\omega > 0$. Let $X(\omega) = \omega$ and $\mathcal{E} = \sigma\langle\mathcal{P}\rangle$, where $\mathcal{P} = \{(n-1, n] : n = 1, 2, \dots\}$. Compute $\mathbb{E}(X|\mathcal{E})$.

  $\mathbb{E}(X|\mathcal{E})(\omega) = c_n$ if $\omega \in (n-1, n]$, where

  $$c_n = \frac{\int_{(n-1,n]} X dP}{P((n-1,n])} = \frac{\int_{n-1}^n \omega e^{-\omega} d\omega}{\int_{n-1}^n e^{-\omega} d\omega} = \cdots = \frac{ne^{-(n-1)} - (n+1)e^{-n}}{e^{-(n-1)}(1-e^{-1})} = \cdots = n - \frac{1}{e-1}. \blacksquare$$

Final 2024

- **Example**: Determine if the following statements are True or False.

  1. The Cantor set is uncountable.
  2. Every Λ-class is also a monotone class.
  3. If P and Q are two arbitrary probability measures on a measurable space $(\Omega, \mathcal{A})$, then $A \mapsto R(A) = \max\{P(A) < Q(A)\}$ is a probability measure.
  4. If ϕ is a CF, then $|\phi|^2$ is also a CF.
  5. If Lindeberg's condition for the CLT holds, then Lyapunov's condition holds.

  1. True by definition.
  2. True.
  3. False, doesn't satisfy countable additivity.
  4. True, $|\phi|^2 = \phi \cdot \bar{\phi}$. So, if X has CF ϕ, then we can take an identical copy of X, denoted as $Y \overset{d}{=} X$. So, $|\phi|^2$ corresponds to X and $-Y$.

5. False. ■

- **Example**: Show that an RV X is integrable iff $\sum_{n=1}^{\infty} P(|X| \geq n) < \infty$.

  First, we establish that $\mathbb{E}|X| = \int_{0,\infty)} P(|X| \geq x)dx$.

  $$\mathbb{E}|X| = \int |x| dP(x) = \int \int_0^{|x|} d\lambda(y) dP(x) \overset{\text{Fub.}}{=} \int \int_{|x| \geq y} dP(x) d\lambda(y) = \int P(|X| \geq y) dy = \sum_{n=1}^{\infty} \int_{n-1}^{n} P(|X| \geq y) dy.$$

  This term is respectively bounded above and below by $\sum_{n=1}^{\infty} P(|X| \geq n-1)$ and $\sum_{n=1}^{\infty} P(|X| \geq n)$. Note

  $$\sum_{n=1}^{\infty} P(|X| \geq n-1) = 1 + \sum_{n=1}^{\infty} P(|X| \geq n).$$

  Since $\sum_{n=1}^{\infty} P(|X| \geq n) < \infty$, the result holds. ■

- **Example**: If X is an RV such that $\mathbb{E}(X^2) = 1$ and $\mathbb{E}|X| \geq a$ for some $a \geq 0$, then show that for any $0 < \lambda < 1$, $P(|X| \geq \lambda a) \geq (1-\lambda)^2 a^2$.

  $$a \leq \mathbb{E}|X| = \int_{|X| < \lambda a} |X| dP + \int_{|X| \geq \lambda a} |X| dP \leq \lambda a \cdot P(|X| < \lambda a) + \int |X| \mathbb{1}\{|X| \geq \lambda a\} dP \leq \lambda a + \int |X| \mathbb{1}\{|X| \geq \lambda a\} dP.$$

  Put the λa on the LHS to get that

  $$(1-\lambda)a \leq \int |X| \mathbb{1}\{|X| \geq \lambda a\} dP \overset{\text{C.S.}}{\leq} \sqrt{\mathbb{E}(X^2)} \cdot \sqrt{P(|X| \geq \lambda a)} = \sqrt{P(|X| \geq \lambda a)}.$$

  Squaring both sides yields the result. ■

- **Example**: Let $X_1, X_2, \ldots \overset{\perp}{\sim} (0,1)$. Let $S_n = \sum_{i=1}^{n} X_n$. Show that

  $$\left| \sum_{n=1}^{\infty} \frac{X_n}{\sqrt{n} \log n} \right| < \infty \text{ a.s., and } \frac{S_n}{\sqrt{n} \log n} \overset{\text{a.s.}}{\to} 0 \text{ as } n \to \infty.$$

  This uses the fact that $\sum_{n=2}^{\infty} n^{-a} (\log n)^{-b} < \infty$ if either $a > 1$, or $a = 1$ and $b > 1$.

  $$Var\left(\frac{X_n}{\sqrt{n} \log n} \right) = \frac{Var(X_n)}{n (\log n)^2} = \frac{1}{n (\log n)^2}.$$

  $\sum_{n=1}^{\infty} Var\left(\frac{X_n}{\sqrt{n} \log n} \right) = \sum_{n=2}^{\infty} \frac{1}{n (\log n)^2}$. We can apply the given fact, where $a = 1$ and $b = 2$, to get that this sum is finite.
  Showing $\frac{S_n}{\sqrt{n} \log n} \overset{\text{a.s.}}{\to} 0$ is a trivial application of Kronecker's lemma using the first result. ■

- **Example**: Let $U_1, U_2, \ldots \overset{\text{iid}}{\sim} \text{Unif}(-1,1)$. Obtain the limiting distribution of $\sqrt{n} \left[\left(\prod_{i=1}^{n} U_i^2 \right)^{1/n} - e^{-2} \right]$.

  Note that $\prod_{i=1}^{n} U_i^2 = \left(\prod_{i=1}^{n} |U_i| \right)^2$. Let $V_i = |U_i|$. We need to find the asymptotic distribution of $\left(\prod_{i=1}^{n} V_i \right)^{2/n}$ This is an application of the Delta method, where $g(x) = e^x$, which has a nonzero first derivative everywhere. Now,

  $$\mathbb{E}(\log V) = \int_0^1 \log v dv = -\int_0^{\infty} y e^{-y} dy = -1,$$

  and

  $$Var(\log V) = \int_0^{\infty} y^2 e^{-y} dy = \Gamma(3) = 2,$$

  so

  $$\log \left(\prod_{i=1}^{n} V_i \right)^{2/n} = \frac{2}{n} \sum_{i=1}^{n} \log(V_i) \sim AN\left(-2, \frac{4}{n} \right).$$

  Therefore, by the Delta Method,

  $$\prod_{i=1}^{n} U_i^2 \sim AN\left(e^{-2}, (e^{-2})^2 \cdot \frac{4}{n} \right). \blacksquare$$

- **Example**: Let X_n be independent RVs uniformly distributed over $\{-n, -n+1, \dots, n-1, n\}$ for $n = 1, 2, \dots$. Show that $3n^{-3/2} \sum_{i=1}^n X_i \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$. Use the fact that for any natural number n, $\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}$, while for any $p > 0$, $\sum_{i=1}^n i^p \sim \frac{n^{p+1}}{p+1}$ as $n \to \infty$.

  We will use Lyapunov's CLT. $X_i \overset{\perp}{\sim} \begin{pmatrix} -i & -i+1 & \dots & i-1 & 1 \\ \frac{1}{2i+1} & \frac{1}{2i+1} & \cdots & \frac{1}{2i+1} & \frac{1}{2i+1} \end{pmatrix}$. $\mathbb{E}(X_i) = 0$ and

$$\mathbb{E}(X_i^2) = \tau_i^2 = 0 + 2 \times \frac{\sum_{j=1}^i j^2}{2i+1} = 2 \cdot \frac{i(i+1)(2i+1)}{6} \times \frac{1}{2i+1} = \frac{1}{3} i(i+1) = \frac{1}{3}(i^2 + i).$$

  Now, we calculate σ_n^2.

$$\sigma_n^2 = \sum_{i=1}^n \tau_i^2 = \frac{1}{3}\left[\frac{n(n+1)(2n+1)}{6} + \frac{n(n+1)}{2}\right] \sim \frac{n^3}{9}.$$

  Next, we verify the Lyapunov condition.

$$E|X_i|^{2+\delta} = 2 \cdot \frac{\sum_{j=1}^i j^{2+\delta}}{2i+1} \lesssim i^{2+\delta}, \text{ so } \lesssim \sum_{i=1}^n \mathbb{E}|X_i|^{2+\delta} \lesssim n^{3+\delta},$$

  where dividing by $\sigma_n^{2+\delta}$ yields an asymptotic value of $\frac{n^{3+\delta}}{(n^{3/2})^{2+\delta}} = n^{-\delta/2}$. ■

- **Example**: Let $\Omega = (1, \infty)$ be the sample space equipped with the Borel σ-field on it, and probability measure P, where P has the density $p(x) = 2x^{-3}$, $x > 1$. Consider a sub-σ-field $\mathcal{E}$ generated by the countable partition $\{(n, n+1] : n = 1, 2, \dots\}$. Let $X(\omega) = \omega$ for all $\omega \in \Omega$. Compute $\mathbb{E}(X|\mathcal{E})$.

  $\mathbb{E}(X|\mathcal{E})(\omega) = \sum_{n=1}^\infty c_n \mathbb{1}\{\omega \in (n, n+1]\}$, where

$$c_n = \frac{\int_{E_n} X dP}{P(E_n)} = \frac{\int_n^{n+1} \omega \cdot 2\omega^{-3} d\omega}{\int_n^{n+1} 2\omega^{-3} d\omega} = \frac{\frac{\omega^{-1}}{-1}|_n^{n+1}}{\frac{\omega^{-2}}{-2}|_n^{n+1}} = n + \frac{n}{2n+1}.$$ ■

- **Example**: Let X and Y be square integrable RVs such that $\mathbb{E}(X|Y) = Y$ and $\mathbb{E}(Y|X) = X$. Show that $X = Y$ a.s.

  Sufficient to show that $\mathbb{E}[(X-Y)^2] = 0$.

$$\begin{aligned}
\mathbb{E}[(X-Y)^2] =& \mathbb{E}(X^2) - \mathbb{E}(XY) - \mathbb{E}(XY) + \mathbb{E}(Y^2) = \mathbb{E}(X^2) - \mathbb{E}\left[\mathbb{E}(XY|X)\right] - \mathbb{E}\left[\mathbb{E}(XY|Y)\right] + \mathbb{E}(Y^2) \\
=& \mathbb{E}(X^2) - \mathbb{E}\left[X \cdot \mathbb{E}(Y|X)\right] - \mathbb{E}\left[Y \cdot \mathbb{E}(X|X)\right] + \mathbb{E}(Y^2) \\
=& \mathbb{E}(X^2) - \mathbb{E}\left[X \cdot X\right] - \mathbb{E}\left[Y \cdot Y\right] + \mathbb{E}(Y^2) = 0. \ \blacksquare
\end{aligned}$$

10 Random Math Stuff

Return to Table of Contents

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

11 Distributions

Return to Table of Contents

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

Discrete Distributions

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

Continuous Distributions

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

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

11.1 Equivalences

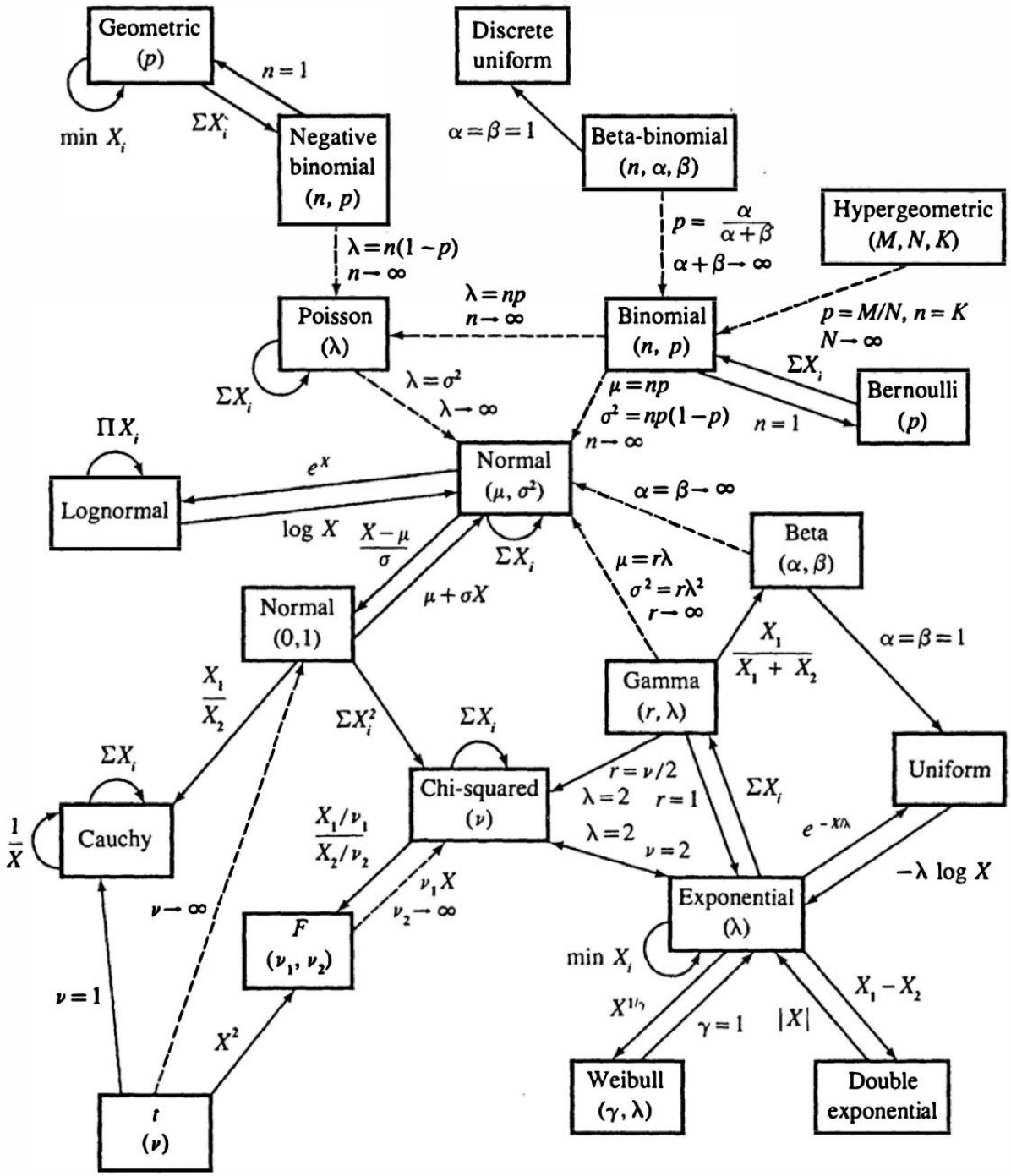
Return to Table of Contents

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

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

Memorylessness

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



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

Source: Casella and Berger, *Statistical Inference*.