# PhD Econometrics Notes

Paul Bousquet

2021

## Acknowledgements

These are notes largely compiled from PhD Econometrics classes at the University of Alabama and Duke University (from Dr. Junsoo Lee, Dr. Traviss Cassidy, Dr. Matt Masten, and Dr. Adam Rosen). The general content will be the simple statement of key definitions, theorems/lemmas, proofs, and propositions (with little extraneous discussion). The structure will follow a review on statistics/asymptotics and then notes on the core topics seen in these classes, such as OLS, IV,m estimation, and hypothesis testing.

## Formatting

- First two sections will have their own formatting, reading more like a traditional textbook
- (otherwise) Definitions will be bolded; Lemmas will be underlined; Theorems will be bolded and underlined
- Any lemma/theorem formatted such as $(\cdot)^*$ will be proven in the appendix.
- This is meant to make the most sense as a retrospective overview. Therefore, the ordering may seem a bit counterintuitive from an instructing/learning perspective at some points (e.g. how much content is covered before OLS is explicitly covered).

# Math/Stats/Measure Theory Review

Informally, we consider the **measure** of a set to be its length. For example, intuitively the measure of the union of disjoint intervals $[a, b], [c, d]$ should be $b-a+d-c$. To extend this across more non-trivial examples with subsets of $\mathbb{R}$, more formally consider that we want a function, call it a measure on $\mathbb{R}$, $\mu : \mathcal{A} \to [0, \infty]$ ($\mathcal{A}$ is a subset of the power set of $\mathbb{R}$), such that (s.t) we can measure every subset of $\mathbb{R}$, for a given collection of mutually disjoint sets we have $\mu(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i)$, two sets have the same $\mu(\cdot)$ if they can be made equivalent by basic transformations (translation, rotations, and reflections), and $\mu([0, 1]) = 1$ (unit interval). Unfortunately no such function exists. The solution is instead to create a measure function with a domain that in some sense excludes problematic subsets of $\mathbb{R}$. The most common "solution" is to focus on a particular type of subset known as a *Borel* $\sigma-$*algebra*, a type of $\sigma-$*algebra*. The next series of definitions serves to lay the groundwork for measurably and Borel algebras.

Specifically, $\mathcal{A}$, subset of the power set of X, is an **algebra** of sets on X if $A, B \in \mathcal{A} \implies A \cup B \in \mathcal{A}$, $B \in \mathcal{A} \implies B^c \in A$ (complement with respect to $X$), anf $X \in \mathcal{A}$. If $\mathcal{A}$ also has the property that any countable family in $\mathcal{A}$ satisfies $\cup_{n=1}^{\infty} A_n \in \mathcal{A}$ then it is a **$\sigma$-algebra**. Consider that the intersection of a collection of $\sigma-$algebras is a $\sigma-$algebra. This fact allows for a process to generate a unique, most parsimonious $\sigma-$algebra that contains $\mathcal{A}$. Accordingly, define $\sigma(\mathcal{A})$, the $\sigma-$*algebra generated by* $\mathcal{A}$, as the $\sigma-$algebra that satisfies $\mathcal{A} \subseteq \sigma(\mathcal{A})$, (uniqueness) $\mathcal{B}$ is a $\sigma-$algebra with $\mathcal{B} \subseteq \mathcal{A} \implies \sigma(\mathcal{A}) \subseteq \mathcal{B}$, and which is precisely equal to the intersection of all $\sigma-$algebras containing $\mathcal{A}$. Combined with a Topology, this yields a construction of the Borel $\sigma-$algebra[1]. A *topology* on X is a collection $\mathcal{T}$ of subsets of $X$ such that

1. $\emptyset, X \in \mathcal{T}$
2. $\mathcal{T}$ is closed under arbitrary unions (if $\{A_{i \in I}\}$ is a collection of sets in $\mathcal{T} \implies \cup_{i \in I} A_i \in \mathcal{T}$)
3. $\mathcal{T}$ is closed under finite intersections (if $\{A_i\}_{i=1}^{n}$ is a finite collection of sets in $\mathcal{T} \implies \cap_{i \in I}^{n} A_i \in \mathcal{T}$)

Call $(X, \mathcal{T})$ a topological space. Given $(X, \mathcal{T})$, $\mathcal{B}(X)$, the **Borel $\sigma$-algebra** on $X$, is simply $\sigma(\mathcal{T})$. Note

$$\mathcal{B}(\mathbb{R}) = \sigma(\{(a, b] : -\infty \leq a \leq b < \infty\}) = \sigma(\{(a, b) : -\infty \leq a \leq b \leq \infty\}) = \sigma(\{(-\infty, b] : b \in \mathbb{R}\})$$

A **measure space** is a triple $(X, \mathcal{A}, \mu)$, where X is a set, $\mathcal{A}$ is a $\sigma-$algebra on X, and $\mu : \mathcal{A} \to [0, \infty]$ is a *measure* on the space where $\mu(\emptyset) = 0$ and $\mu(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mu(E_i)$ if $\{E_i\}_{i=1}^{\infty}$ is a collection of mutually disjoint subsets. We consider the sets in $\mathcal{A}$ to be measurable. A function $f : X \to Y$ is *Borel measurable* if, given a set $A$ open in $Y$, $f^{-1}(A)$ is a Borel set ($f^{-1}(A) = \{x : f(x) \in A\}$). Applying these concepts to probability theory, consider $\Omega$, with $\sigma$-algebra $\mathcal{F}$, to be our sample space, where subsets of $\Omega$ can be thought of as events. We say a measure $\mu$ for $(\Omega, \mathcal{F})$ is a *probability measure* if $\mu(\Omega) = 1$. Here, it's standard to use $\mathbb{P}$ to denote a probability measure. Thus, we have a *probability space* by $(\Omega, \mathcal{F}, \mathbb{P})$. With respect to this

[1] Some definitions have the $\sigma-$algebra generated by the family of compact sets in X (not equivalent to our definition)

probability space, for $A, B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, the *conditional probability*[2] *of* $A$ *given* $B$ is $\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}$ and $A$ and $B$ are **independent** ($\mathbb{P}(A|B) = \mathbb{P}(A)$) if and only if (iff) $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$. Note that if we have two Borel measurable sets $f$ and $g$, independence of $A$ and $B$ implies that $f(A)$ and $g(B)$ are independent by $\mathbb{P}(f(A) \in X, g(B) \in Y) = \mathbb{P}(A \in f^{-1}(X), B \in g^{-1}(Y)) = \mathbb{P}(A \in f^{-1}(X)) \cdot \mathbb{P}(B \in g^{-1}(Y))$.

A *random element* is a measurable function on a probability space. If $X : \Omega \to \mathbb{R}^K$ is a random element (codomain being the measurable space $(\mathbb{R}^K, \mathcal{B}(\mathbb{R}^K))$, for $k \geq 2$ X is a **random vector** and for **k=1** its a **random variable**. This is how "data" can be thought of under a math-based paradigm. We can consider single realization $\omega \in \Omega$ (the data) with $n$ observations $X_1(\omega), \ldots, X_n(\omega)$. If $\{X_1, \ldots X_n\}$ are independent random variables with the same $\mathbb{P}$, then they are an **iid** (independent/identically distributed) sample. We can also use Borel measurability to create well-defined environments for analysis. If $X$ is a random variable and $g : \mathbb{R} \to \mathbb{R}$ is Borel measurable on $\mathbb{R}$, then defining $Y(\omega) = g(X(\omega))$ (where $Y : \Omega \to \mathbb{R}$) we have that Y is a random variable. Additionally, all continuous[3] functions are Borel measurable. It should be apparent why we have had all this discussion: Borel is the smallest $\sigma$-algebra such that all continuous functions are measurable.

If $A \in \mathbb{R}$ and we have random variables $X$ and $Y$ satisfying $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$, then $X$ and $Y$ have the same distribution. If they only differ on a set which occurs with probability zero, then $X$ and $Y$ are *equivalent*. We can infer from the $\mathcal{B}(\mathbb{R})$ result that we are only concerned with probabilities of the form $\mathbb{P}(X \in (-\infty, b])$. This motivates the standard, fundamental statistical definitions. Let $X$ be a random variable and define $F_X : \mathbb{R} \to [0, 1]$ by $F_X(x) = \mathbb{P}(X \leq x)$ (with $x \in \mathbb{R}$). We call $F_X$ the **cumulative distribution function (CDF)** of $X$. When the CDF is continuous at $x$, $\mathbb{P}(X = x) = 0$. If $F_X$ is absolutely continuous, then define $f_X$, the **probability distribution function (PDF)**, by $F_X(b) - F_X(a) = \int_a^b f_X(x)dx$ and if $f_X$ is continuous at $x$ an equivalent definition is $f_X(x) = F'_X(x)$. The *support* of $X$, call it $S$, is the smallest closed set such that $\mathbb{P}(X \in S) = 1$. The **expectation**[4] of a random variable $X$ with respect to (w.r.t) the space $(\Omega, \mathcal{F}, \mathbb{P})$ is $\mathbb{E}[X] = \int_\Omega X d\mathbb{P}$. Equivalently, if $\mathbb{E}[X]$ exists, $\mathbb{E}[X] = \int_\mathbb{R} x dF_X(x)$, and if $X$ is absolutely continuously distributed then $\mathbb{E}[X] = \int_\mathbb{R} x f_X(x)dx$. We refer to $\mathbb{E}[X^n]$ as an *nth moment* and $\mathbb{E}[(X - \mathbb{E}[X])^n]$ as an *nth central moment*. The **variance** of $X$ is its second central moment, or by linearity $\text{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. For $a, b \in \mathbb{R}$, $\text{var}(aX + b) = a^2 \text{var}(X)$. When $X$ is a random vector, $\text{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])']$ and the $(i, j)$ element of $\text{var}(X)$ is the *covariance*: $\text{cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$. Also for a square matrix $B$ (with same dimension as $X$), $\text{var}(BX) = B\,\text{var}(X)B'$.

[2] Recall/note the inclusion-exclusion formula: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

[3] For topological spaces $X$ and $Y$, a function $f : X \to Y$ is continuous if for every $V$ open in $X$, $f^{-1}(V)$ is open in X

[4] For information sets $\mathcal{F}_1 \subseteq \mathcal{F}_2, \mathbb{E}[\mathbb{E}[X|\mathcal{F}_2]|\mathcal{F}_1] = \mathbb{E}[X|\mathcal{F}_1]$ (*L.I.E - Law of Iterated Expectations*).
Also $\mathbb{E}[\mathbb{E}[X|\mathcal{F}_1]|\mathcal{F}_2] = \mathbb{E}[X|\mathcal{F}_1]$ since $\mathbb{E}[X|\mathcal{F}_1]$ is know w.r.t $\mathcal{F}_2$

When we say $X \sim \mathcal{Q}(\mu, \sigma^2)$, this typically means that $X$ is distributed along the "$\mathcal{Q}$"-distribution with a first moment (*mean*) of $\mu$ and variance $\sigma^2$. There are exceptions: for instance, the Cauchy distribution does not have a first moment. The most commonly used distribution is the *normal*: $X \sim \mathcal{N}(\mu, \sigma^2)$ with PDF $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-(x-\mu)^2}{2\sigma^2})$. The *standard normal* distribution follows $X \sim \mathcal{N}(0,1)$. Any normal can be converted to a standard normal (asymptotically) by dividing its difference from the mean by its *standard deviation*, the square-root of its variance. The normal distribution is *symmetric*: $f_X(\mu + x) = f_X(\mu - x)$. Accordingly, the normal distribution has the nice visual property of 68.2% of observations being within one standard deviation away from the mean (34.1% on either side of $\mu$), 95.4% of observations being away from two standard deviations, and so on. This property helps illustrate the underlying important of standard deviations/variance: it's how much you should reasonably expect an observation to deviate from the mean (in absolute value). However, for all of these well-behaved characteristics of the normal distribution, there's only so much slack with manipulation. Consider $X \sim \mathcal{N}(0,1)$ with CDF $\Phi(\cdot)$. Then for some odd integer $a$, the CDF of $Y = X^a$ is the piecewise function $F_Y(x) = \Phi(x^{\frac{1}{a}})\ (x \geq 0), 1 - \Phi(|x|^{\frac{1}{a}})\ (x < 0)$. Raising $X$ to an even power is even more complicated (consider how to deal with $x < 0$).

Now that we have introduced notions of asymptotic behavior, we can talk about how to "limit towards" or arrive at results that we want, and precisely characterize how this happens. A sequence of random variables $\{X_n\}$ converges **almost surely** to a random variable $X$ ($X_n \xrightarrow{a.s} X$) if $\mathbb{P}(\lim_{n \to \infty} ||X_n - X|| = 0) = 1$.
Similarly, $X_n$ converges to X **in probability** ($X_n \xrightarrow{p} X$) if $\forall \varepsilon > 0, \mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$.
$X_n$ converges to X **in distribution** ($X_n \xrightarrow{d} X$) if $X_n \sim F_n$,$X \sim F \implies F_n(x) \to F(x)\ (n \to \infty)\ \forall x$ s.t $F(x)$ is continuous. We can also say $X_n \xrightarrow{d} X$ if $\lim_{n \to \infty} \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x)$ for all x where continuity holds. This is also sometimes called *weak convergence* and denoted ($X_n \rightsquigarrow X$). *Uniform convergence* requires that some $N > 0$ will assure that given $n > N, X_n - X$ is arbitrarily small (i.e the difference is less in magnitude than any $\epsilon > 0$, w.r.t the given convergence paradigm). We can also introduce ("O") notation to generalize asymptotic behavior. We say $X_n$ is *bounded in probability* ($O_p(1)$) if $\forall \varepsilon > 0\ \exists B \in \mathbb{R}$ s.t $\sup_n(\mathbb{P}(|X_n| > B)) < \varepsilon$. More generally, if there exists a sequence $\{r_n\}$ of positive numbers and some $M \in \mathbb{R}$ such that $|\frac{X_n}{r_n}| < M\ (\forall n)$ then $X_n = O_p(r_n)$. If $\frac{X_n}{r_n} \xrightarrow{p} 0\ (n \to \infty)$ then we say $X_n = o_p(r_n)$.

Next, we can consider some important results using these properties.
(**Continuous Mapping Theorem**) If $X \mapsto g(x)$ continuous on a set C with $\mathbb{P}(x \in C)$= 1 then $X_n \xrightarrow{p} X$ implies that $g(X_n) \xrightarrow{p} g(x)$ and $X_n \rightsquigarrow X \implies g(X_n) \rightsquigarrow g(X)$
(Slutsky's Lemma:) $X_n \rightsquigarrow X$, $Y_n \xrightarrow{p} c \in \mathbb{R}$ implies $X_n + Y_n \rightsquigarrow X + c$, $X_n Y_n \rightsquigarrow cX$, and $Y_n^{-1} X_n \rightsquigarrow X/c$.

(**Markov's Inequality**) Suppose $\mathbb{E}[|X|] < \infty$. Then for any $\lambda > 0$, $\mathbb{P}(|X| \geq \lambda) \leq \lambda^{-1}\mathbb{E}[|X|]$

**Proof:** Taking expectations of the LHS and RHS of the following gives the result:

$$1 \cdot \mathbb{1}(\lambda^{-1}|X| \geq 1) \leq \lambda^{-1}|X| \cdot \mathbb{1}(\lambda^{-1}|X| \geq 1) \leq \lambda^{-1}|X|$$

(**Chebychev's Inequality**) If $\mathbb{E}[|X|] < \infty$ and $\operatorname{var}(X) < \infty$, then for $\lambda > 0$, $\mathbb{P}(|X - \mathbb{E}[X]| \geq \lambda) \leq \lambda^{-2} \operatorname{var}(X)$

**(Delta Method)** If $r_n(x_n - \mu) \rightsquigarrow X$, $r_n \to \infty$, and $g(\cdot)$ is differentiable at $\mu$, then $r_n(g(X_n) - g(\mu)) \rightsquigarrow g'(\mu)X$

**Proof:** More general proof follows by assuming $g(\cdot)$ is twice continuously differentiable.
For some $\widetilde{X_n}$ such that $|\widetilde{X_n} - \mu| \leq |X_n - \mu|$:

$$g(X_n) = g(\mu) + g'(\mu)(X_n - \mu) + (X_n - \mu)^2 \cdot g''(\widetilde{X_n})/2$$

Since $r_n(X_n - \mu)$ is $O_p(1), X_n - \mu = r_n^{-1}O_p(1) = o_p(1)O_p(1) = o_p(1)$. Therefore,

$$r_n(g(X_n) - g(\mu)) = g'(\mu)(X_n - \mu) + o_p(1) \rightsquigarrow g'(\mu)X$$

Another important result for Metrics (and Macro) is Jensen's Inequality, which combines properties of expectations with properties of functions. But first we must understand what convex and concave means. A set $C$ is *convex* if given two points in the set, a line segment joining those points is contained in the set. More formally, $x, x' \in C$ and $\alpha \in [0, 1] \implies \alpha x + (1 - \alpha)x' \in C$. A more intuitive way to think about this definition is considering a line segment as an interval. WLOG[5] let $x < x'$, meaning the interval $[x, x']$ could be defined as $\{\alpha x + (1 - \alpha)x'\}$ (consider the extreme values $\alpha$ could take). If this interval is a subset of our set of interest, then the set is convex. Similarly, a function $f(\cdot)$ is convex on the interval $I$ if the set $\{(x, y) : x \in I, y \geq f(x)\}$ is convex. $f(\cdot)$ is *concave* if $\{(x, y) : x \in I, y \leq f(x)\}$ is convex. Informally, this definition implies that a line segment connecting two points on a convex function will lie above the graph. Consider the midpoint $(.5f(a) + .5f(b))$ of the line-segment connecting $f(a)$ and $f(b)$): if $f(.5a + .5b)$ lies above the midpoint the function is convex (you can extend this for $\alpha \neq .5$). Now we are ready to state the main result which intuitively follows almost directly from the definition of convex/concave

(**Jensen's Inequality**) Suppose $f : \mathbb{R} \to \mathbb{R}$ is measurable and that $\mathbb{E}[X]$ and $\mathbb{E}[f(X)]$ exist. If $f(\cdot)$ is convex then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$. If $f(\cdot)$ is concave then $f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$.

**Proof:** Assume $f(\cdot)$ is a well-defined convex function; the concave case follows largely by flipping the inequality signs. Because $f(\cdot)$ is convex, if it's not differentiable everywhere, worst-case scenario it's not differentiable on some countable set of points[6], call this set $\mathcal{D} \subseteq \mathbb{R}$. Then if $\mathbb{E}[X] \notin \mathcal{D}$, by the definition

[5] Without Loss of Generality, meaning you could flip the definition (it would not affect the result if notation were switched)

[6] Intuitively, this is because you can show that the set of all jump discontinuities is a subset of a union of several sets constructed using sufficiently large neighborhoods of rational numbers (there exists some rational number inside the gap of each "jump"). Thus it's a subset of a countable set, so it's countable

of convex[7] we have $f(X) \geq f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(X - \mathbb{E}[X])$. Taking an expected value yields

$$\mathbb{E}[f(X)] \geq \mathbb{E}[f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(X - \mathbb{E}[X])] = f(\mathbb{E}[X]) + f'(\mathbb{E}[X])\mathbb{E}[(X - \mathbb{E}[X])] = f(\mathbb{E}[X])$$

For the multivariate case, use the transpose of the gradient in place of the derivative ($\nabla f(\mathbf{X})^T = [\frac{\partial f(\mathbf{X})}{\partial X_1}, \ldots, \frac{\partial f(\mathbf{X})}{\partial X_n}]$). If the $\mathbb{E}[X] \notin \mathcal{D}$ assumption is undesirable, consider an inductive proof to the alternative form of Jensen's inequality: $\lambda_1 f(z_1) + \cdots + \lambda_n f(z_n) \geq f(z_1\lambda_1 + \ldots z_n\lambda_n)$, where $\lambda_i \in [0,1]$ can be considered weights. Using $y = \sum_{k=1}^{n-1} \frac{\lambda_k z_k}{1-\lambda_n}$, the inductive step here is given by

$$f(z_1\lambda_1 + \ldots z_n\lambda_n) = f((1-\lambda_n)y + \lambda_n z_n) \leq (1-\lambda_n)f(y) + \lambda_n f(z_n) \leq (1-\lambda_n)\sum_{k=1}^{n-1} \frac{\lambda_k}{1-\lambda_n} f(z_k) + \lambda_n z_n = \sum_{k=1}^{n} \lambda_k f(z_k)$$

Now we finally have enough context start talking about arguably the two most major results in the realm of "background math" that will be used ad nauseum in Econometrics (and elsewhere in academia). They are the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT). The version of the LLN of interest to us says that as long as we can assure iid and the existence of a first moment, the sample mean is a great approximation (converges in probability) to the expected value. These are relatively weak assumptions so this is a very nice result to have in our toolkit. The CLT more or less exploits the behavior of the sample mean and says that asymptotically the difference between the sample mean and the population mean is very well behaved (normal distribution) when scaled by $\sqrt{n}$. We will first state and prove LLN, then give a bit more background (Characteristic Functions) needed to state and prove CLT.

**(LLN - Strong)** If $\mathbb{E}[|X|] < \infty$ and $X_1, \ldots, X_n$ iid copies of X, then $\overline{X}_n \xrightarrow{p} \mathbb{E}[X]$

**Proof:** WLOG let $\mathbb{E}[X] = 0$. Pick $M > 0$ and arbitrary $\epsilon > 0$. Then we can define

$$\begin{gathered} Y_i = X_i \cdot \mathbb{1}(|X_i| > M) - \mathbb{E}[X_i \cdot \mathbb{1}(|X_i| > M)] \\ Z_i = X_i \cdot \mathbb{1}(|X_i| \leq M) - \mathbb{E}[X_i \cdot \mathbb{1}(|X_i| \leq M)] \\ \therefore X_i = Z_i + Y_i \ \underline{\textbf{and}} \ \overline{X}_n = \overline{Y}_n + \overline{Z}_n \end{gathered}$$

If we can show $\overline{Y}_n, \overline{Z}_n \xrightarrow{p} 0$ we are done. By Markov's/Chebychev's Inequality and the definition of variance

$$\mathbb{P}(|\overline{Z}_n| > \epsilon) \leq \epsilon^{-2} \operatorname{var}(\overline{Z}_n) = \epsilon^{-2} n^{-1} \mathbb{E}[Z_i^2]$$

By the triangle and Jensen's inequalities

$$\begin{aligned} &|Z_i| \leq |X_i| \cdot \mathbb{1}(|X_i| \leq M) + \mathbb{E}[|X_i| \cdot \mathbb{1}(|X_i| \leq M)] \leq 2M \\ &\implies \epsilon^{-2} n^{-1} \mathbb{E}[Z_i^2] \leq \epsilon^{-2} n^{-1} 4M^2 \xrightarrow{p} 0 \end{aligned}$$

Observe that $|X| \cdot \mathbb{1}(|X| > M) \leq |X|$ and since $\mathbb{E}[|X|] < \infty$, $|X_i|\mathbb{1}(|X_i| > M) \to 0$ as M blows up. Thus

$$\mathbb{P}(|\overline{Y}_n| > \epsilon) \leq \epsilon^{-1}\mathbb{E}[|\overline{Y}_n|] \leq \epsilon^{-1}\mathbb{E}[|Y_i|] \leq 2\epsilon^{-1}\mathbb{E}[|X_i| \cdot \mathbb{1}(|X_i| > M)] \to 0$$

[7] The function values over an interval are above the values of the tangent line. Here's a quick sketch of why this is: convex & $\alpha \in [0,1] \implies f((1-\alpha)x + \alpha y) - f(y) \leq (1-\alpha)(f(x) - f(y))$. Divide by $(1-\alpha)x + \alpha y - y = (1-\alpha)(x-y) > 0$ and note that we get the definition of the derivative as $\alpha \to 1$

Let $X$ be a random variable with PDF $f(x)$. Then the **characteristic function** is defined as

$$\phi_X(t) \equiv \mathbb{E}[\exp(itX)] = \int_{-\infty}^{\infty} \exp(itx) f(x) dx$$

If $\int |\phi_X(t)| dt < \infty$ then $f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(itx)\phi_X(t)dt$

(**Levy Continuity Theorem**) Let $T_n$ have CDF $F_n$ with characteristic function $\phi_n(t)$. Suppose there exists $\phi(t) : \mathbb{R} \to \mathbb{C}$ such that $\phi_n(t) \to \phi(t)$ and $\phi(\cdot)$ is continuous at 0. Then $\phi(\cdot)$ is the characteristic function of some $T$ with CDF $F(\cdot)$ and $T_n \xrightarrow{d} T$

We are now ready to state one of the most important results in statistics that allows for describing distribution with relatively mild assumptions (assuming iid and first/second moments is almost an afterthought in most modeling environments and papers).

(**Lindeberg-Levy CLT for Variables**) Let $\{X_i\}$ be a sequence of iid variables with $\mathbb{E}[X_i] = \mu < \infty$ and $\text{var}[X_i] = \sigma^2 < \infty$

$$\implies \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0,1)$$

**Proof:** Let $Y_i = \frac{X_i - \mu}{\sigma}$ and $T_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$. Then the characteristic function of $T_n$ can be defined as $\phi_n(t) = \mathbb{E}[\exp(\frac{it}{\sqrt{n}} \sum_{i=1}^n Y_i)]$. Let $\psi(t) = \mathbb{E}[\exp(itY)]$ (the characteristic function of Y). Then it follows that $\phi_n(t) = \mathbb{E}[\prod_{i=1}^n \exp(\frac{it}{\sqrt{n}} Y_i)] = \psi(\frac{t}{\sqrt{n}})^n$. Now let $m(t) = \log(\psi(t)) \implies \log(\phi_n(t)) = n \cdot m(\frac{t}{\sqrt{n}})$. By established properties of the taylor series expansion around 0

$$\log(\phi_n(t)) = n\bigg(m(0) + \frac{t}{\sqrt{n}}m'(0) + \frac{t^2}{2n}m''(0) + o(n^{-1})\bigg)$$

By our previous definition of $m(\cdot)$, $m'(t) = \frac{\psi'(t)}{\psi(t)}$ and $m''(t) = \frac{-\psi'(t)}{\psi(t)^2} + \frac{\psi''(t)}{\psi(t)}$. Further note that $\psi(0) = 1$ and $\psi'(0) = i\mathbb{E}[Y] = 0$ because $\mathbb{E}[X_i] - \mu = 0$. So $\log(1) = 0$ eliminates the first two terms of the taylor series. Finally, $\psi''(0) = i^2\mathbb{E}[Y^2] = -1$ because $\text{var}(X_i) = \sigma^2$ and $\mu$ is a constant, so $1 = \frac{\sigma^2}{\sigma^2} = \text{var}(Y_i) = \mathbb{E}[Y^2]$ (by the definition of variance since $\mathbb{E}[Y_i] = 0$). Now we have $m''(0) = \frac{\psi''(t)}{\psi(t)} = -1$. Combining everything yields

$$\log(\phi_n(t)) = n\bigg(-\frac{t^2}{2n} + o(n^{-1})\bigg) \implies \phi_n(t) \to \mathbb{E}[\exp(-\frac{t^2}{2})]$$

Which is the characteristic function of the standard normal distribution. Note $\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} = \frac{\sqrt{n}}{n} \sum_{i=1}^n Y_i = T_n$. Therefore by the Levy Continuity theorem we have our desired result. ■

Note the importance of $\sqrt{n}$. Say for $T_n$ we used $\frac{1}{n^a}$ (for $a$ not necessarily equal to $\frac{1}{2}$). Then the proof would collapse to $\log(\phi_n(t)) = n(-\frac{t^2}{2n^{2a}} + o(\frac{1}{n^{2a}})) = -\frac{t^2}{2}n^{1-2a} + o(n^{-2a})$. Therefore if $a > \frac{1}{2}$, the limit of $\phi_n(t)$ is 0, which is the characteristic function of a degenerate variable. If $a < \frac{1}{2}$, the limit of $\phi_n(t)$ diverges.

(Cramer-Wold Device) Let $\{X_n\}$ be a sequence of random vectors in $\mathbb{R}^K$ and let $X \in \mathbb{R}^K$.
Then $X_n \xrightarrow{d} X$ iff $\lambda' X_n \xrightarrow{d} \lambda' X$ (for all $\lambda \in \mathbb{R}^K$)

Now we can prove the **Lindeberg-Levy CLT for vectors**.

**Proof:** Let $\{X_i\}$ be iid vectors in $\mathbb{R}^K$ with $\mathbb{E}[X_i] = \mu$ and $\text{var}(X_i) = \Omega < \infty$ (assume $\Omega$ is non-singular). Let $\lambda \in \mathbb{R}^K$ be an arbitrary fixed vector. By $\mathbb{E}[\lambda' X_i] = \lambda'\mu$ and $\text{var}(\lambda' X_i) = \lambda'\Omega\lambda$, from the univariate CLT result we have $\frac{1}{\sqrt{n}}\sum_{i=1}^n (\lambda' X_i - \lambda'\mu) \xrightarrow{d} \mathcal{N}(0, \lambda' X_i \Omega\lambda) = \lambda'\mathcal{N}(0,\Omega)$. Since we made $\lambda$ arbitrary, by the Cramer-Wold device $\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu) \xrightarrow{d} \mathcal{N}(0,\Omega)$ ■.

Finally, there are some linear algebra concepts which are helpful to be aware of. Let $A$ be an $m \times n$ matrix. The *transpose* of $A$, denoted by $A^T$, is the matrix whose $i$-th column is the $i$-th row of $A$, or equivalently, whose $j$-th row is the $j$-th column of $A$. Notice that $A^T$ is an $n \times m$ matrix. We will write $A^T = (a^T_{ji})$ where $a^T_{ji} = a_{ij}$. Notice that the $ji$-entry of $A^T$ is the $ij$-entry of $A$. This tells us that the main diagonals of a matrix and its transpose are the same and that entries of $A^T$ are the entries of $A$ reflected about the main diagonal. Here are a couple of examples.

**Properties of the Transpose**: Let $A$ and $B$ be appropriately sized matrices and $r \in \mathbb{R}$.
Then $(A^T)^T = A$,$(A+B)^T = A^T + B^T$,$(rA)^T = rA^T$,$(AB)^T = B^T A^T$.

A *positive definite matrix* is a symmetric matrix with all positive eigenvalues. More specifically, let $A$ and $x$ be $n \times n$ and $n \times 1$ matricies (respectively). A is positive definite (pd) matrix if $\forall x, x'Ax > 0$. A is *positive semi-definite* (psd) if $\forall x, x'Ax \leq 0$. Notationally, for two matricies $A$ and $B$, $A \geq B$ iff $A - B$ is psd. Let $X$ be $n \times k$ with full rank. We also have some useful matrices $P = X(X'X)^{-1}X'$ and $M = I_n - P$, the *projection* and *annihilator* matrices (respectively). Both $P$ and $M$ are *symmetric* ($P = P'$) and *idempotent* ($PP = P$). We also see that $PX = X, MX = 0,$ and $MY$ produces the residuals of OLS.

# Hypothesis Testing

**Wald's Unifying Theory of Statistical Data Analysis:** Consider $\mathbb{P}$ the "state of the world": the true probability distribution. Different actions have different relative value, and one can consider a utility function that provides an implicit ranking of these actions based on $\mathbb{P}$. Therefore, a **decision rule** is a function $d(\cdot)$ mapping our data to an action that is taken. This is the paradigm under which *hypothesis testing* comes into play: our decision rule rejects or fails to reject a hypothesis based on some piecewise/threshold of relative probabilistic performance. Broadly, we want to solve $\max_{d(x)} \mathcal{U}(\mathbb{P}, d(x))$, but we can't because $\mathbb{P}$ is unknown. There are different ways to deal with this issue.

**Subjective Bayesian Solution:** Impose some belief about $\mathbb{P}$ prior to observing any data. This manifests in a **prior distribution**, $\pi(\cdot)$, a probability measure. Once we see a realization of the data, the prior is updated to $\pi(\cdot|x)$, a new probability measure. Now, expected utility can be determined by integrating over $\mathbb{P}$ given a choice of decision rule and a known prior: $\max_{d(x)} \int \mathcal{U}(\mathbb{P}, d(x)) d\pi(\mathbb{P}|x)$, resulting in the Bayesian optimal decision. However, the process of choosing this prior is subjective by design, which immediately creates doubt as to the reasonableness of the choice. Also these models perform much better when $\mathbb{P}$ is known to being in a parametric class of probability distributions (e.g. normal distribution with unknown mean and variance). The Bayesian approach is *ex post* because it takes a realization ($x$ of $X$) as given (what is seen in the data).

**Frequentist Solution:** An *ex ante* approach that ignores the realization of $X$. Instead, multiple possible realizations are considered, via considering $d(\cdot)$ as a sampling distribution. The goal here is to learn about the distribution of $d(x)$ that arises from repeated sampling. Consider a *welfare* function: $W(\mathbb{P}, d) = \int \mathcal{U}(\mathbb{P}, d(x)) d\mathbb{P}(x)$, which does not depend on the data. Therefore, instead of maximizing over a single number ($d(x)$) the entire function needs to be picked by $\max_{d(\cdot)} W(\mathbb{P}, d)$, but this cannot be done without dealing with $\mathbb{P}$. One approach is to essentially average the average: $\max_{d(\cdot)} \int W(\mathbb{P}, d) d\pi(\mathbb{P})$, where $\pi$ is a prior on the set of possible $\mathbb{P}$ ($\mathcal{P}$). Another is to look at worst case scenarios for $\mathbb{P}$ and pick what does best when the worst case happens: $\max_{d(\cdot)} \min_{\mathbb{P} \in \mathcal{P}} W(\mathbb{P}, d)$. A similar approach – these are thought of as *minimax* approaches – is to look at minimizing "regret" from choosing something non-optimal. Let $d^*(\mathbb{P}) \in \text{argmax}_{d(\cdot)} W(\mathbb{P}, d)$ be the decision rule we'd pick if we knew the true state of the world (infeasible since we don't). Then define MaxRegret (a function of $d$ by $W(\mathbb{P}, d^*(\mathbb{P})) - W(\mathbb{P}, d)$. Then we can have a reasonable choice of d by simply choosing the d that minimizes maximum regret. The potentially problematic aspects of the Frequentist approach come from the need to produce repeated samples (i.e. how do we learn about the sampling distribution). One obvious way to deal with this is making many assumptions about $\mathbb{P}$. A more nuanced approach is the to use approximations for the sampling distribution, for instance using the CLT for large sample sizes.

**Wald Hypothesis Testing:** Suppose we want to characterize a parameter $\theta = \theta(\mathbb{P})$. Consider $\Theta$ the set of all possibilities of $\theta$. For testing, $\Theta$ is partitioned into two disjoint circumstances: **null hypothesis ($H_0$)** that $\theta \in \Theta_0$ and the **alternative hypothesis ($H_a$)** $\theta \in \Theta \backslash \Theta_0$. Common examples include $\Theta_0 = \{\theta_0\}$ (singleton) and $\Theta_0 = \{\theta \in \Theta : \theta \geq \theta_0\}$ (one-sided). Hypothesis testing imposes a binary decision rule: 1 for rejecting the null and 0 for not rejecting. Consider a *loss* function $\mathcal{L} = -\mathcal{U}$ and *risk* function $R = -W$. Then the risk function of a hypothesis test decision rule $\phi_n$ is $R(\mathbb{P}, \phi_n) = \mathcal{L}(\mathbb{P}, 1)\mathbb{P}(\phi_n(X) = 1) + \mathcal{L}(\mathbb{P}, 0)\mathbb{P}(\phi_n(X) = 0)$. So the risk only depends on $\mathbb{P}(\phi_n(X) = 1)$, which is called the *Power* function. Generally, the loss function takes a value of 1 if the decision rule is wrong (*type 1 error*: null is true but is rejected or *type 2 error*: null is false but we did not reject) and 0 otherwise. Another option is weighting the loss of false positives and false negatives differently: giving a loss of $K$ if there is a type 1 error. A standard choice is $K = 19$, which implies a false positive is 19 times worse than a false negative. A standard test for $H_0 : \theta \leq \theta_0$ is to reject if the realization lies beyond the $\frac{K}{1+K}$ quantile (.95 for $K = 19$) of the given distribution.

**Neyman-Pearson Hypothesis Testing:** Largely a special case of Wald. To find an optimal test, this approach restricts possible tests and then tries to find the most powerful among the group. First, note that a test's *size* is essentially the maximum likelihood of a false positive: Size $= \sup_{\mathcal{P} \ni \mathbb{P} : \theta(\mathbb{P}) \in \Theta_0} \mathbb{P}(\phi_n(X) = 1)$. So given $\alpha \in (0, 1)$ a *level $\alpha$ test* follows Size$(\phi_n) \leq \alpha$. A uniformly most powerful test means that any other test can't be more powerful (i.e if $\theta(\mathbb{P}) \in \Theta_0$, the value of the power function for any other test will not be greater). So the idea would be restricting to level alpha then finding the unconditionally most powerful test. However, proving a test is uniformly most powerful is pretty infeasible. One workaround is restricting the potential tests even further (i.e. being most powerful among a smaller group). For instance, defining the group of unbiased tests as ones where the power of a test with $\theta(\mathbb{P}) \notin \Theta_0$ is greater than or equal to a test with $\theta(\mathbb{P}) \in \Theta_0$.

**Applications:** Recall $\alpha$ is the size of a test. Let $\beta$ denote the probability of a false negative ($\mathbb{P}_1(\phi_n(X) = 0)$). The *sensitivity* of a test, or the probability of a true positive, is $1 - \beta$; the *specificity*, $1 - \alpha$, is the probability of a true negative. Let the *rejection region* be defined as the realizations of $X$ leading to a rejection ($\mathcal{R} = \{x \in \text{supp}(X) : \phi_n(x) = 1\}$), so the probability of rejection, or power, is simply

$\mathbb{P}(\phi_n(X) = 1) = \mathbb{P}(X \in \mathcal{R})$. One of the most popular types of tests is a cutoff test with a decision rule based on a test statistic ($\mathbb{1}[T(x) > c]$), where c is called a *critical value*. *Randomized tests* are measurable functions from the sample to $[0, 1]$. Essentially this acts as a test statistic, where repeated samples can be drawn under the assumption that the null is true. For example, we take a draw from a Bernoulli trial, and if we restrict the codomain of the test to be $(0, 1) \subseteq [0, 1]$, then the decision to reject the null is some fixed probability ($q$) or in other words resembles a potentially unevenly weighted coin flip (formally, if the codomain is not restricted and is instead $\{0, q, 1\}$, we can consider the case where $f_{X|\theta(\mathbb{P})}(x|\theta \notin \Theta_0) = K \cdot f_{X|\theta(\mathbb{P})}(x|\theta_0)$ being resolved by this "coin flip"). These ideas set the table for a discussion on **p-values**, which have their own paradigms.

**Neyman-Pearson p-values:** The Neyman-Pearson p-value is calculated by first restricting attention to non-randomized test, then defining the p-value as the largest size out of all possible tests under rejection. More formally, first let $\Phi$ be the set of all possible non-randomized tests of a given null. Let $x \in \text{supp}(X)$. Call $p(\Theta_0, x) = \inf_{\Phi \ni \delta_n : \delta_n(x)=1} \text{Size}(\delta_n)$ the *(N-P) p-value.* This definition becomes a bit clearer when applied to a cutoff test. Consider the case where the critical value is defined by a strictly increasing function $c = cv(\cdot)$, then a level $\alpha$ test would be $c = cv(1 - \alpha)$. The smaller $\alpha$ gets, the higher the critical value becomes, meaning the bar to reject is raised, which should follow intuitively. As an arbitrary example, consider $T(x) = 1.7 \text{and} cv(1 - \alpha) = 1.64$. This means there is a magnitude of .06 worth of "slack". We can think about the p value as defined where there is no slack, or more simply when the critical value function is exactly equal to the test statistic. Within the $1 - \alpha$ paradigm, we can say the p-value solves $T(x) = cv(1 - p) \implies p = 1 - cv^{-1}(T(x))$, which isn't exactly equivalent to the earlier definition since this is the special case of cutoff tests. This paradigm can be narrowed even further. Suppose $cv^{-1}(t) = \mathbb{P}_0(T(X) \leq t)$, where $\mathbb{P}_0$ is the distribution of $X$ under the null. Then $cv^{-1}(\cdot)$ is simply a CDF of the sampling distribution of the test statistic under the null. This leads to a p-value function of $p(x) = \mathbb{P}_0(T(X) > T(x))$, which is possibly the most used definition. One immediate application of p-values is to compare them to $\alpha$. Formally, we can say that $\mathbb{1}[p(x) \leq \alpha]$ is a level $\alpha$ test of the null of $\mathbb{P}_0$ against the alternative that $\mathbb{P} = \mathbb{P}_0$. Many people want to extend p-values beyond a simple extension of level $\alpha$ tests, relating to false positives, and make a more nuanced claim that p-values can be a proxy for the level of evidence, with a low p-value meaning strong evidence the null is untrue.

**Fisherian p-values** Fisher's approach is not really derived from statistical decision theory, unlike Bayesian and frequentist approaches. Specifically in contrast to the ex ante frequentist approach, there is not a desire to have certain properties relating to error rates across repeated samples. These procedures are supposed to work around a single dataset. The approach can be thought of methodically as follows. First, make enough assumptions such that the sampling distribution of some test statistic is known. Then, observe a realization $x$ of $X$ and subsequently a value for $T(x)$. Compare this realization to the distribution of $T(X)$ under the null. Formally, this gives the *Fisherian p-value*: $P_0(T(X) > T(x))$. The interpretation for this p-value is its a measure of "abnormality", so large p-values mean $T(x)$ conforms to expectations and low $p - values$ are unlikely to be seen if the null is true. So we can make a cutoff for rejecting the null (e.g. $p < .05$). Note that a formal alternative hypothesis is not specified, and an important criticism is that the data not being conformable to a null hypothesis doesn't necessarily imply its more conformable to a different hypothesis. Although the interpretation of the p-value is different from N-P, we can generalize the Fisherian approach and show numerical equivalence. Suppose $\Phi$ is a collection of hypothesis test indexed by increasing propensity to reject the null (indexing holds for all $x \in \text{supp}(X)$) and let $p(X)$ represent the N-P p-value for $\Phi$. Define the relation $\preceq$ on $\text{supp}(X)$ by $x \preceq x'$ if and only if $p(x) \geq p(x')$. For Fisher with a distribution $\mathbb{P}$ of $X = (X_1, \ldots, X_n)$, we think of $x \preceq x'$ as a way of saying ("weak ordering") that $x$ is more consistent with the hypothesis $\mathbb{P}$ than $x'$, and have a p-value in a *pure significance* setting (significance probability of the data relative to the weak order) by $\mathbb{P}(\{x' \in \text{supp}(X) : x \preceq x'\})$. This is numerically equivalent to the previously derived N-P p-value under the $\preceq$ class.

**Clarity:** P-values are extremely nuanced, but their popularity erodes some of the intricacies of interpretation. For instance, some falsely claim p-values are probabilities the null holds. This is a Bayesian statement: this would be derived a test of $\theta \in \Theta_0$ based on a-priori assumptions about the world, but people usually make such statements after using a frequentist approach. P-values are also not type-1 error rates. This is false namely because it varies across repeated samples; $\alpha$ is simply used as a barometer fixed beforehand. Further, as Hubbard and Bayarri (2003) sum up well, often researchers will use N-P methodology but Fisherian philosophy. P-values are not measures of the strength of evidence under N-P. It's also important to note the implications of the lack of alternative hypothesis and the supporting evidence paradigm. As stated earlier, the null being weakly supported doesn't imply an alternative is more supported. Further, a null is not as necessarily as "supported" as a hypothesis that includes the null and other possibilities, which is a bit counter-intuitive. In a practical sense, we expect that $\mathbb{P}(A) \leq \mathbb{P}(A \cup B)$. And formally, if p-values were truly a metric for supporting evidence, we would expect that $\Theta_0 \subseteq \Theta_0' \implies p(\Theta_0, x) \leq p(\Theta_0', x) \forall x \in \text{supp}(X)$. But Fisherian p-values do not satisfy this requirement. There are often critiques of the hypothesis testing status quo in general. For instance, a common null is $\theta_0 = 0$, which is largely expected to be false and is a rather specific point estimate. In addition, there is a difference between statistical significance and whether something is meaningfully different that can get lost in the weeds. Loss functions have also been criticized as being too structurally rigid, with a 0-1 paradigm not accounting for the relative distance between $\theta\,\text{and}\,\theta_0$. There is also the problem of multiple testing; the procedures are built on running a single tests and multiple can skew the results (e.g. running 20 tests on 20 independent samples), leading to the phenomena known as "p-hacking" where researchers run simulations until they get what they want and therby do not commit to procedures/properties hypothesis testing has based on the ex ante/"go in blind and commit" foundations.

**Confidence Sets:** A *level $\alpha$ confidence set* is a function $\mathcal{C}(X)$ from the sample to a subset of $\Theta$ such that $\forall \mathbb{P} \in \mathcal{P}, \{x \in \text{supp}(X) : \mathcal{C}(x) \ni \theta(\mathbb{P})\}$ is $\mathbb{P}$-measurable and $1 - \alpha \leq \mathbb{P}(\theta(\mathbb{P}) \in \mathcal{C}(X))$. These are usually known as confidence intervals when $\theta \in \mathbb{R}$. A common misinterpretation of confidence sets is that they contain the true probability with probability $1 - \alpha$, and as alluded to in the preceding subsection, this is a Bayesian way of thinking that is usually misapplied. In fact, the confidence set itself is the random element. A good analogy is the difference between archery and ring toss. A mistake would be to think of a confidence set like an archery target, where its sufficiently big enough that the true value (the arrow) will land inside. Instead, the true value is fixed (like a post in ring toss) and we are "throwing" a confidence set in its direction trying to capture it. Returning from metaphor land, $1-\alpha$ gives a large, ex ante probability the confidence set will contain the true parameter, but once the data is drawn it either contains the true parameter or doesn't. This discussion primarily falls under the N-P school of thought. Fisherian sets, which are still analogous, can be thought of as $\{\theta \in \Theta : \mathbb{P}_{\theta_0}(T(X) \geq T(x)) \geq \alpha\}$, the parameter values not rejected by a test comparing Fisherian p-values to $\alpha$. This does not rely upon a notion of repeated sampling. Practically, we can attain confidence sets through finite sample properties, like if we know data to be normally distributed it has certain properties, or use asymptotics (usually through the CLT) to construct intervals centered around the idea of deviation in the limit.

Proofs relevant to the applied performance of the sampling distribution and minimax approaches will be in the appendix.

# Identification and Causality

(This section is especially a work in progress)

Say we have population parameters $\{(Y_i, X_i, U_i) \subseteq \mathbb{R}^3 : ii \in \mathcal{I}\}$ with a corresponding distribution $F_{Y,X,U} = P^{\text{true}}$. We want to learn about some (population) parameter that's a function of the distribution, call it $\theta(\cdot)$. However, usually we don't observe an entire population; we only observe population data, call it $h(\cdot)$ with realization $P^{\text{data}} = h(P^{\text{true}})$. Let $\Theta$ be the possible values of $\theta(\cdot)$. Then the *identified set* for $\theta$ is $\Theta_I = \{\theta \in \Theta : \theta(P)\text{for some } P \in \mathcal{P} \text{ s.t } h(P) = P^{\text{data}}\}$. These are all the parameter values that satisfy our assumptions and are observationally equivalent to the distribution of realized values. If $\Theta_I$ is singleton, then it's *point identified*, and if $\Theta_I$ is some non-singleton subset of $\Theta$ it's *partially identified.* If $\Theta_I$ is empty, the assumptions made are too strong and the model is **falsified**. When we can't tell whether an assumption is correct or not, we say its not *falsifiable.* Models can be falsifiable because of an assumption made on the observe random variables (e.g. assuming a normal distribution) or assumptions made on unobserved variables. The unobserved variables are important when trying to learn about causality because we often want to learn about outcomes we do not see in the data. For example, if we define the function $Y(x) = p(x)'\beta + U$ as an approximation to $Y$, this is falsifiable because it gives an approximation for $\mathbb{E}[Y|X = x]$ (by taking the conditional expectation of the function) which we observe in the data. This informs how we can create models for learning about causality and gauging the validity of them. The big ideas are that we want to learn "how A affects B" *ceteris paribus* - holding everything else equal, but the fundamental problem of causal inference is that we only observe one realization in the data, whereas we would like to have observe two disjoint outcomes to compare the difference and extrapolate a causal effect.

A *unit level causual model* consists of *dependent* variables ($y$) determined inside the model, covariates/outcome variables ($x \, \text{and} \, u$) determined outside of the model, and functional relations between them known as *structural functions.* There are also restrictions on the relationships and values of these variables.

Consider the "all causes" special case: $y \in \mathbb{R}, x \in \mathcal{X} \subseteq R,$ and $u$ are a vector of unobservables that can can take on any value in some set $\mathcal{U}$. For some function $g : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$, $y = g(x, u)$. This is an "all causes" model because if we know the values of $g(\cdot)$'s inputs we can perfectly predict y. This gives a basic paradigm for terminology to develop

Define the *unit level causal effect* of x on y to be $g(x_1, u_0) - g(x_0, u_0)$, in other words the change in $y$ given the change in $x$ holding $u$ constant. We can also define the *marginal level causal effect* of x on y to be $[\nabla_x g](x_0, u_0) = \frac{\partial g(x_0,u_0)}{\partial x}$

The *treatment response function* $Y_i(\cdot) : \mathcal{X} \to \mathbb{R}$ is given by $Y_i(x) = g(x, U_i)$, where for any fixed value of

$x$ $Y_i(x)$ is a potential outcome. So a unit level causal effect for $i$ is simply $Y_i(x_1) - Y_i(x_0)$, and learning about this relationship is the main goal of causal analysis. In the all causes model, $Y_i$ already has a realized outcome of $Y_i = Y_i(X_i)$. So the value of the potential outcome function when its input is different from its realization (i.e. $x \neq X_i$) are called **counterfactual** outcomes. And the difference between a counterfactual and a realized outcome can be thought of as a "status quo" *treatment effect*

In the all causes model and in general, we say that when the unit level causal/marginal effects vary with differences in $U_i$, the causal effects are *heterogeneous.* So it would then follow that there is a distribution of causal effects. Variation in $X$ among individuals is considered observed heterogeneity, and variation in $U$ is unobserved. In general, we can define $\delta(x_0 \to x_1) = Y(x_1) - Y(x_0)$. Implicit in this notation is that $Y(x) = g(x, U)$ without the individual subscript is a random variable, as its not based on an individual. This gives us the *distribution of treatment effects* by $\text{DTE}(t, x_0 \to x_1) = F_{Y(x_1)-Y(x_0)}(t) = \mathbb{P}(Y(x_1) - Y(x_0) \leq t)$

Suppose $X$ is binary. Then $\mathbb{P}(Y(1) > Y(0)) = 1 - \text{DTE}(t, x_0 \to x_1)$ is proportion who benefit from being treated and $\mathbb{E}[\delta(x_0)] = \mathbb{E}_U([\nabla_x g](x_0, U))$ is the average partial effect of X on Y. The ATE - *average treatment effect* (by linearity) is $\mathbb{E}_U[g(x_1, U)] - \mathbb{E}_U[g(x_0, U)] = \mathbb{E}[Y(x_1)] - \mathbb{E}[Y(x_0)]$, with the *average structural function* (ASF) $\mathbb{E}_U[g(x, U)] = \mathbb{E}[Y(x)]$. Note $\text{ASF}(x) \neq \mathbb{E}[g(x, U)|X = x]$. We also can define *average treatment effect on treated* (ATT) and *average treatment effect on the untreated* (ATU) by $\mathbb{E}[Y(1) - Y(0)|X = 1]$ and $\mathbb{E}[Y(1) - Y(0)|X = 0]$ (respectively), yielding $\text{ATE} = \text{ATT}\mathbb{P}(X = 1) + \text{ATU}\mathbb{P}(X = 0)$. Suppose we also observe covariates $W$, then we can define *conditional ATE* (CATE) by $\mathbb{E}[Y(1) - Y(0)|W = w]$. There is also an additively separable, where $Y(x) = m(x) + U$. Letting $U_i$ be fixed, this is a *homogeneous* treatment effects model since $Y_i(x_1) - Y_i(x_0) = m(x_1) - m(x_0)$. The marginal effect also doesn't depend on $U_i$.

QTE section

The rest of our analysis will assume that $X$ is binary (so realizations can be formulated by $Y_i = Y_i(1)X_i + Y_i(0)(1 - X_i i)$) and the possible treatment values ($x_0$ and $x_1$) are degenerate (i.e. we only observe one value for each if it occurs). Then if $\mathcal{Y}$ is the set of logically possible values for potential outcomes, the identified set for $Y_i(x)$ for $x \neq X_i$ and $x \in \mathcal{X}$ is $\mathcal{Y}$. This is a formalization of the fundamental problem of causal inference: basically the data is completely unhelpful on its own for determining counterfactual outcomes.

To get some more information, we will look at marginal and joint distributions of potential outcomes. We can write the marginal distribution (i.e. CDF of potential outcomes)) as the sum of the relative probability of treatment "$x$" times the conditional CDF of the realization and the probability non-treatment "$x$" times the conditional CDF of the potential outcomes, in other words

$$\mathbb{P}(Y(x) \leq y) = \mathbb{P}(Y \leq y|X = x)\mathbb{P}(X = x) + \mathbb{P}(Y(x) \leq y|X \neq x)\mathbb{P}(X \neq x)$$

Note that the only object we don't observe in the data is $\mathbb{P}(Y(x) \leq y | X \neq x)$, a distribution of counterfactual outcomes. So this is essentially a missing data problem. Similarly, the identified set for the joint distribution of two different potential outcomes $F_{Y(x_1),Y(x_0)}(\cdot,\cdot)$ (given treatments $x_1, x_0 \in \mathcal{X}$) is simply the set of all joint CDFs whose marginal CDFs are defined above. This implies that we can't point identify the CDF of potential outcomes for all treatments: the more we learn about the distribution of $\mathbb{P}(Y(x) \leq y)$ the less we know about the distribution of other potential outcomes. The simple and highly relevant question of whether outcomes change with (counterfactual) treatment cannot be answered, therefore what we have before us in both the marginal and joint distributions is essentially a missing data problem.

Some context on the missing data issue: let $Z$ be a binary variable reflecting whether we observe $Y$ or not, since observability is not necessarily independent of the outcome (otherwise we can basically just pretend like there is no missing data). Assume we observe the distribution of $Y|(Z = 1)$ and marginal distribution of $Z$. Then the identified set for $\theta \equiv \mathbb{P}(Y = 1)$ is $[\mathbb{P}(Y = 1|Z = 1)\mathbb{P}(Z = 1), \mathbb{P}(Y = 1|Z = 1)\mathbb{P}(Z = 1) + \mathbb{P}(Z = 0)]$. This is because by the LIE $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 1|Z = 1)\mathbb{P}(Z = 1) + \mathbb{P}(Y = 1|Z = 0)\mathbb{P}(Z = 0)$, and since $\mathbb{P}(Y = 1|Z = 0)$ is completely not identified from the data, the bounds on this object are $[0, 1]$. Thus, $\mathbb{P}(Y = 1)$ is point identified if $\mathbb{P}(Z = 0) = 0$ and partially IDd when $\mathbb{P}(Z = 0) \in (0, 1)$. A common solution to deal with this is to impute missing observations with sampling from the observed data. However, in the case of causal inference, there is an important difference between the distribution of observed outcomes and distributions of hypothetical, and in turn $\mathbb{P}(Y \leq y|X = x) - \mathbb{P}(Y(x) \leq y)$ can be thought of as *selection bias.*

Now applying these concepts to the world of potential outcomes, define $Y^{\text{norm}}(x) = \frac{Y(x) - y_{\min}}{y y_{\max} - y_{\min}}$. This allows us to bound any given $Y$ by $[0, 1]$. Using normalized $Y$, the identified set for $\mathbb{E}[Y(x)]$ is $\mathbb{E}[\text{LB}(x), \text{UB}(x)] = [\mathbb{E}(Y = 1|X = x)\mathbb{P}(X = x), \mathbb{E}(Y|Z = 1)\mathbb{P}(X = x) + \mathbb{P}(X \neq x)]$. Thus, the identified set for the ATE is $[\text{LB}(x_1) - \text{UB}(x_0), \text{UB}(x_1) - \text{LB}(x_0)]$. Note that the best (tightest) width we can get is 1, meaning 0 will lie in the interval and we can't definitively say one treatment is better than the other from the data. More generally we have the bound $[\mathbb{E}[Y|X = x_1]\mathbb{P}(X = x_1) - \mathbb{E}[Y|X = x_0]\mathbb{P}(X = x_0) + 0 \cdot \mathbb{P}(X \neq x_1) - 1 \cdot \mathbb{P}(X \neq x_0), \mathbb{E}[Y|X = x_1]\mathbb{P}(X = x_1) - \mathbb{E}[Y|X = x_0]\mathbb{P}(X = x_0) + 1 \cdot \mathbb{P}(X \neq x_1) - 0 \cdot \mathbb{P}(X \neq x_0)]$, so the width is $\mathbb{P}(X \neq x_1) + \mathbb{P}(X \neq x_0) = 2 - (\mathbb{P}(X = x_1) + \mathbb{P}(X = x_0))$

Since data alone doesn't help very much, we have to consider other alternatives, the first of which is *random assignment*: $X \perp\!\!\!\perp \{Y(x) : x \in \mathcal{X}\}$, which point identifies $F_{Y(x)}$ if the joint distribution $(X, Y)$ is known. The joint distribution of potential outcomes is still only partially identified because we can't simultaneously observe a pair of outcomes. Another relevant issue is how covariates affect treatment assignments, or more specifically the propensity by researchers to assume $X \perp\!\!\!\perp \{Y(x) : x \in \mathcal{X}\}|W$.

## M-Estimation

Suppose we have an iid sample, $Z_1, \dots, Z_n$. $\widehat{\theta}$ is called an **m-estimator** if it solves

$$\max_{\theta \in \Theta} \widehat{Q}_n(\theta) = \max_{\theta \in \Theta} Q(Z_1, \dots, Z_n; \theta)$$

Where $Q(\cdot)$ is the objective function and $\Theta$ is a family of parameters. This definition is commonly rewritten using a sample estimator of an objective function, where the estimator is derived by maximizing a known, real-valued function $m_\theta(\cdot)$ by

$$\widehat{Q}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_\theta(z_i)$$

Inarguably the most commonplace m-estimator is OLS, where $m_\theta(z_i) = -(y_i - x_i^{'}\theta)^2$ (with $z_i = (y_i, x_i)$). Suppose the data ($Z = (Z_1, \dots, Z_n)$) has a joint distribution with $\theta$ (i.e. a pdf) of $p_\theta(z_i) = f(z_i; \theta)$, where $\int f(z; \theta) dz = 1$. This parameter $\theta$ indexes members of a family of distributions, such as $\{p_\theta : \theta \in \Theta\}$. Then we can extrapolate another common m-estimator, MLE, as $m_\theta(z_i) = \log(p_\theta(z_i))$.

**(Consistency of m-estimators)** If there exists $Q_0(\theta)$ such that i) $Q_n(\theta)$ is uniquely maximized at $\theta_0$ ii) $\Theta$ compact iii) $Q_0(\theta)$ is continuous iv) $\widehat{Q}_n(\theta)$ converges uniformly to $Q_0(\theta)$. Then $\widehat{\theta} \xrightarrow{p} \theta_0$.

**Proof:** For any $\epsilon > 0$, since $\widehat{\theta}_n$ maximizes $Q_0(\theta)$, $\widehat{Q}_0(\widehat{\theta}) > \widehat{Q}_n(\theta) - \epsilon/3$. By the uniform convergence property, with probability approaching 1 (WPA 1) we have $Q_0(\theta) > \widehat{Q}_0(\widehat{\theta}) - \epsilon/3$ and $\widehat{Q}_0(\theta_0) > Q_0(\theta_0) - \epsilon/3$.

Therefore, WPA 1

$$Q_n(\widehat{\theta}) > \widehat{Q}_n(\theta) - \epsilon/3 > \widehat{Q}_0(\widehat{\theta}) - 2\epsilon/3 > Q_0(\theta_0) - \epsilon$$

Now we have the issue of establishing separability: we have shown that the objective function using $\theta_0$ and $\widehat{\theta}$ will converge, but as presently summarized we can't rule out that $\widehat{\theta}$ isn't "far away from" $\theta_0$ and happens to produce a comparable value of the objective function in the limit. Let $\mathcal{N}$ be any open subset of $\Theta$ such that $\theta_0 \in \mathcal{N}$ (think of this as an arbitrary neighborhood). Define $A = \Theta \cap \mathcal{N}^c$. Since A is compact and by the unique maximization and continuity conditions, for some $\theta^* \in A$: $\sup_{\theta \in A} Q_0(\theta) = Q_0(\theta^* < Q_0(\theta_0)$. Choose $\epsilon = Q_0(\theta_0) - Q_0(\theta^*)$. Then it follows WPA 1 $Q_n(\widehat{\theta}) > Q_0(\theta^*)$, so $\widehat{\theta} \in \mathcal{N}$. Since we made $\mathcal{N}$ arbitrary, $\widehat{\theta}$ is in every neighborhood of $\theta_0$ (with respect to the probability limit). ■

The first condition is known as an identification condition. This is because $\boldsymbol{\theta_0}$ is **identified** if for all $\theta \neq \theta_0$ we have $p_\theta \neq p_{\theta_0}$. To see why this is important consider $X_i$ under a Bernoulli distribution with parameter $\theta = (\beta, \delta)$. Then $p_\theta(x_i) = (\beta + \delta)_i^x (1 - (\beta + \delta))^{1 - x_i}$. Then $\theta = (.1, .2)$ and $\theta = (0, .3)$ gives the same result. So estimation of this parameter would not be unique or informative. Returning to the first condition consider

(MLE ID Lemma) If $\theta_0$ is identified and $\mathbb{E}[|\log(f(y|x;\theta))|] < \infty$ for all $\theta$, then $Q_0 = \mathbb{E}[\log(f(y|x;\theta))]$ has a unique max at $\theta_0$. (Think of $f(y|x;\theta) = f(z;\theta)$)

**Proof:** Unique max exists if for all $\theta \neq \theta_0$, $Q_0(\theta_0) - Q_0(\theta) > 0$. By the strictness of Jensen's Inequality

$$Q_0(\theta_0) - Q_0(\theta) = \mathbb{E}[-\log\left(\frac{f(y|x;\theta)}{f(y|x;\theta_0)}\right)] > -\log(\mathbb{E}[\frac{f(y|x;\theta)}{f(y|x;\theta_0)}]) = -\log(\int \frac{f(y|x;\theta)}{f(y|x;\theta_0)} f(y|x;\theta_0)dz) = -\log(\int f(y|x;\theta)dz) = 0$$

Another type of m-estimation is GMM: Generalized Method of Moments. Suppose we have an $m_\theta(\cdot)$-type function $g(\cdot,\cdot) : \mathbb{R}^K \times \mathbb{R}^L \to \mathbb{R}^L$ that satisfies $g_0(\theta_0) = \mathbb{E}[g(z,\theta_0)] = 0$ (where $\theta_0$ is the "true parameter"). Suppose $\widehat{W}$ is a psd weight matrix, where we assume it converges in probability to some constant psd matrix $(W)$ by LLN. Then $\widehat{\theta}$ is the **GMM estimator** if it maximizes

$$\widehat{Q}_n(\theta) = -\left(\frac{1}{n}\sum_{i=1}^{n} g(z_i,\theta)\right)^{'} \widehat{W}\left(\frac{1}{n}\sum_{i=1}^{n} g(z_i,\theta)\right) = -(\widehat{g}_n(\theta))^{'}\widehat{W}\widehat{g}_n(\theta)$$

(GMM ID Lemma) Using GMM conditions, if $Wg_0(\theta) \neq 0$ for $\theta \neq \theta_0$ then $\theta_0$ uniquely maximizes $-g_0(\theta)^{'}Wg_0(\theta)$.

**Proof:** Define $R$ such that $W = R^{'}R$. So if $\theta \neq \theta_0$ then $R^{'}Rg_0(\theta) \neq 0$. Therefore by pre-multiplication

$$Rg_0(\theta) \neq 0 \implies -g_0(\theta)^{'}Wg_0(\theta) = -(Rg_0(\theta))^{'}(Rg_0(\theta)) < 0 = -g_0(\theta_0)^{'}Wg_0(\theta_0)$$

Now that we have established how to identify the true parameters for some of the m-models, we need to think about showing the usefulness of the estimators of the true parameters. The following lemma makes it easier to formally complete the consistency theorem tenets to show that the the estimators are consistent.

(Convergence Lemma) If the data are iid, $\Theta$ compact, a function $a(z_i,\theta)$ continuous at each $\theta \in \Theta$ with probability 1, and there exists a function $d(\cdot)$ such that $||a(z,\theta)|| < d(z)$ for all $\theta$ and $\mathbb{E}[d(z)] < \infty$. Then $\mathbb{E}[a(z,\theta)]$ is continuous and $\sup_{\theta\in\Theta} ||\frac{1}{n}\sum_{i=1}^{n} a(z_i,\theta) - \mathbb{E}[a(z,\theta)]|| \xrightarrow{p} 0$

**(MLE Consistency)** Suppose the data $Z$ are iid with pdf $f(z_i;\theta_0)$ and i) $\theta_0$ identified ii) $\Theta$ compact with $\theta_0 \in \Theta$ iii) $\log((f(z_i;\theta))$ continuous at each $\theta \in \Theta$ with probability 1 iv) $\mathbb{E}[\sup_{\theta\in\Theta} |\log((f(z;\theta))|] < \infty$. Then $\widehat{\theta}_{MLE} \xrightarrow{p} \theta_0$.(Proven from the convergence lemma completing the m-estimator consistency theorem).

**(GMM Consistency)** Assume iid data and that the previous GMM conditions and identification hold. If i) $\Theta$ compact with $\theta_0 \in \Theta$ ii) $g(z,\theta)$ continuous at each $\theta \in \Theta$ with probability 1 iii) $\mathbb{E}[\sup_{\theta\in\Theta} ||g(z,\theta)||] < \infty$, then $\widehat{\theta}_{GMM} \xrightarrow{p} \theta_0$

**Proof:** Let $\widehat{g}_n(\theta) - g_0(\theta) = GG_n$.By the triangle and Cauchy-Schwartz inequalities (also that $\widehat{W}$ is symmetric)

$$\begin{aligned}\widehat{Q}_n(\theta) - Q_0(\theta) &\leq |[GG_n]'\widehat{W}GG_n| + |g_0(\theta)'(\widehat{W} + \widehat{W}')GG_n| + |g_0(\theta)'(\widehat{W} - W)g_0(\theta)| \\ &\leq ||GG_n||^2||\widehat{W}|| + 2||\widehat{W}||||g_0(\theta)||||GG_n|| + ||g_0(\theta)||^2||(\widehat{W} - W)||\end{aligned}$$

where simplifying the first line yields the triangle inequality.Using the convergence lemma, $\sup_{\theta \in \Theta} |\widehat{Q}_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0$ from properties of $\widehat{W}, g_0(\theta)$;the lemma/given info complete the remaining consistency tenets $\blacksquare$.

(**<u>Asymptotics of m-estimators</u>**) Suppose $\widehat{\theta}$ is consistent for $\theta_0$ and that i) $\theta_0 \in \text{Int}(\Theta)$, ii) $\widehat{Q}_n(\theta)$ is twice continuously differentiable in a neighborhood ($\mathcal{N}$) of $\theta_0$, iii) $\sqrt{n}\frac{d\widehat{Q}_n(\theta_0)}{d\theta} \xrightarrow{p} \mathcal{N}(0, \Sigma)$, iv) there exists $H(\cdot)$ continuous at $\theta_0$ with $\sup_{\theta \in \mathcal{N}} ||\frac{d^2\widehat{Q}_n(\theta)}{d\theta d\theta'} - H(\theta)|| \xrightarrow{p} 0$, v) $H = H(\theta_0)$ is non-singular.
Then $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{p} \mathcal{N}(0, H^{-1}\Sigma H^{-1})$

**Proof:** Since $\widehat{\theta}$ is consistent, there exists an open convex set $\mathcal{N}$ with $\theta_0 \in \mathcal{N}$ such that $\theta \in \mathcal{N}$ WPA 1. Thus, also WPA 1 $\frac{d\widehat{Q}_n(\widehat{\theta})}{d\theta} = 0$ by the F.O.C and i)-iii). By MTV, there exists $\theta_n^*$ in a properly defined interval of $\widehat{\theta}$ and $\theta$ such that

$$0 = \frac{d\widehat{Q}_n(\widehat{\theta})}{d\theta} = \frac{d\widehat{Q}_n(\theta_0)}{d\theta} + \frac{d^2\widehat{Q}_n(\theta_n^*)}{d\theta d\theta'}(\widehat{\theta} - \theta)$$

$$\implies \sqrt{n}(\widehat{\theta} - \theta) = \left(\frac{d^2\widehat{Q}_n(\theta_n^*)}{d\theta d\theta'}\right)^{-1}\left(-\sqrt{n}\frac{d\widehat{Q}_n(\theta_0)}{d\theta}\right) = (\widehat{H}(\theta_n^*))^{-1}\left(-\sqrt{n}\frac{d\widehat{Q}_n(\theta_0)}{d\theta}\right)$$

We have a condition to take care of the second RHS term, but need to be careful about how we deal with the first (see the discussion of MLE asymptotics for more detail). Remember that $\theta_n^*$ is defined in an interval, and by consistency this interval is shrinking as n grows,so $\theta_n^* \xrightarrow{p} \theta_0$.By iv) and the triangle inequality,WPA 1

$$||\widehat{H}(\theta_n^*) - H|| \leq ||\widehat{H}(\theta_n^*) - H(\theta_n^*)|| + ||H(\theta_n^*) - H|| \leq \sup_{\theta \in \mathcal{N}} ||\widehat{H}(\theta) - H(\theta)|| + ||H(\theta_n^*) - H|| \xrightarrow{p} 0$$

Therefore by iii), the CMT, and Slutsky $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{p} H^{-1}\mathcal{N}(0, \Sigma) = \mathcal{N}(0, H^{-1}\Sigma H^{-1})$ $\blacksquare$.

For MLE asymptotics, we will use some different notation and go through a more informal proof (that will break down some of the steps in the last proof), assuming conditions as we go along and stating them formally at the end. We will temporarily use $\theta$ for the true parameter. Let $\dot{l}_\theta = \frac{\partial}{\partial\theta}\log p_\theta$ (known as the score function) and $\ddot{l}_\theta = \frac{\partial}{\partial\theta}\dot{l}_\theta$. The MLE solves $\sum_{i=1}^n \dot{l}_\theta(x_i) = 0$. By the MVT,for some $\tilde{\theta}$ between $\widehat{\theta}_n$ and $\theta$

$$0 = \sum_{i=1}^n \dot{l}_{\widehat{\theta}}(x_i) = \sum_{i=1}^n \dot{l}_\theta(x_i) + \sum_{i=1}^n \ddot{l}_{\tilde{\theta}}(x_i)(\widehat{\theta}_n - \theta)$$

$$\implies \sqrt{n}(\widehat{\theta}_n - \theta) = -\left(\frac{1}{n}\sum_{i=1}^n \ddot{l}_{\tilde{\theta}}(x_i)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{l}_\theta(x_i)$$

We can't use the LLN because $\ddot{l}_{\tilde{\theta}}(x_i)$ is a function of $\{x_i\}$ and thus not independent across i. So to establish the convergence of $\frac{1}{n}\sum_{i=1}^n \ddot{l}_{\tilde{\theta}}(x_i)$ note that

$$\frac{1}{n}\sum_{i=1}^n \ddot{l}_{\tilde{\theta}}(x_i) - \mathbb{E}[\ddot{l}_\theta] = \frac{1}{n}\sum_{i=1}^n \ddot{l}_{\tilde{\theta}}(x_i) - \mathbb{E}[\ddot{l}_{\tilde{\theta}}] + \mathbb{E}[\ddot{l}_{\tilde{\theta}}] - \mathbb{E}[\ddot{l}_\theta] = (*) + (**)$$

Note that $\mathbb{E}[\ddot{l}_{\tilde{\theta}}]$ can be viewed as a mapping of $\theta$ to $\mathbb{E}[\ddot{l}_\theta]$ that integrates with respect to randomness in $X_i$. Since $\tilde{\theta}$ is in a de-facto shrinking neighborhood, consistency for $\widehat{\theta}_n$ implies that $\tilde{\theta}$ is also consistent for $\theta$. If we assume continuity, then $\mathbb{E}[\ddot{l}_{\tilde{\theta}}] \xrightarrow{p} \mathbb{E}[\ddot{l}_\theta]$ so $(**) \xrightarrow{p} 0$.

And by the uniform LLN (assuming compactness of set of estimators, continuity, and that there exists a function with finite first moment that is greater than $|\ddot{l}_\theta(x_i)|$ for each $x_i$).

$$|*| \leq \sup_\theta |\frac{1}{n}\sum_{i=1}^n \ddot{l}_\theta(x_i) - \mathbb{E}[\ddot{l}_\theta]| \xrightarrow{p} 0$$

Therefore $\frac{1}{n}\sum_{i=1}^n \ddot{l}_{\tilde{\theta}}(x_i) \xrightarrow{p} \mathbb{E}[\ddot{l}_\theta]$ so by the CMT $(\frac{1}{n}\sum_{i=1}^n \ddot{l}_{\tilde{\theta}}(x_i))^{-1} \xrightarrow{p} (\mathbb{E}[\ddot{l}_\theta])^{-1}$.

To deal with the other term, note that $\int p_\theta dx = 1$. Differentiating both sides of this expression with respect to $\theta$ (assuming $p_\theta$ is differential in $\theta$, for example that the support of $X_i$ doesn't depend on $\theta$) yields

$$0 = \frac{\partial}{\partial\theta}\int p_\theta dx = \int \dot{p}_\theta dx = \int \ddot{p}_\theta dx$$

We also have $\dot{l}_\theta = \frac{\partial}{\partial\theta}\log p_\theta = \frac{\dot{p}_\theta}{p_\theta}$ and by extension $\ddot{l}_\theta = \frac{\ddot{p}_\theta}{p_\theta} - \left(\frac{\dot{p}_\theta}{p_\theta}\right)^2$. Then $\mathbb{E}[\dot{l}_\theta] = \int \frac{\dot{p}_\theta}{p_\theta} p_\theta dx = \int \dot{p}_\theta dx = 0$. Since $\int \ddot{p}_\theta dx = 0$

$$\mathbb{E}[\ddot{l}_\theta] = \mathbb{E}[\frac{\ddot{p}_\theta}{p_\theta}] - \mathbb{E}[\left(\frac{\dot{p}_\theta}{p_\theta}\right)^2] = -\mathbb{E}[(\dot{l}_\theta)^2]$$

So for $\theta$ that's not a scalar, define the fisher information matrix by

$$-\mathbb{E}[\ddot{l}_\theta] = \mathbb{E}[\dot{l}_\theta \dot{l}_\theta'] \equiv I_\theta$$

So by CLT $\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{l}_\theta(x_i) \xrightarrow{d} \mathcal{N}(0, I_\theta)$ and therefore by Slutsky and $(\mathbb{E}[\ddot{l}_\theta])^{-1} = (I_\theta)^{-1}$

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} I_\theta^{-1}\mathcal{N}(0, I_\theta) = \mathcal{N}(0, I_\theta^{-1})$$

Let $\mathbb{E}[T_n] = \theta$. To show efficiency among unbiased estimators ($\text{var}[T_n] \geq I_\theta$), note that

$$\mathbb{E}[T_n \dot{l}_\theta] = \int T_n \frac{\dot{p}_\theta}{p_\theta} p_\theta = \int T_n \dot{p}_\theta = \frac{\partial}{\partial \theta} \int T_n p_\theta = \frac{\partial}{\partial \theta} T_n = \frac{\partial}{\partial \theta} \theta = 1$$

By $\mathbb{E}[\dot{l}_\theta] = 0$ and the Cauchy-Schwartz inequality

$$(\mathbb{E}[T_n \dot{l}_\theta])^2 = (\mathbb{E}[(T_n - \mathbb{E}[T_n])\dot{l}_\theta)])^2 \leq \mathbb{E}[(T_n - \mathbb{E}[T_n])^2]\mathbb{E}[(\dot{l}_\theta)^2] = \text{var}[T_n] I_\theta$$

Combining these results, $1 \leq \text{var}[T_n] I_\theta \implies I_\theta^{-1} \leq \text{var}[T_n]$, so MLE is asymptotically efficient

Now we will state explicitly all the conditions needed for these results (using $\theta_0$ as true parameter)

**(Asymptotics of MLE)** Suppose all the earlier conditions for consistency of MLE are satisfied. Also that i) $\theta_0 \in \text{Int}(\Theta)$, ii) $f(z;\theta)$ is twice continuously differentiable in $\theta$ and $f(f;\theta) > 0$ in a neighborhood ($\mathcal{N}$) of $\theta_0$, iii) $\int \sup_{\theta \in \mathcal{N}} ||\frac{df(z;\theta)}{d\theta}||dz < \infty$ and $\int \sup_{\theta \in \mathcal{N}} ||\frac{d^2 f(z;\theta)}{d\theta d\theta'}||dz < \infty$, iv) $J \equiv \mathbb{E}\big[\frac{d\log(f(z;\theta_0))}{d\theta}\frac{d\log(f(z;\theta_0))}{d\theta'}\big]$ exists and is non-singular, and v) $\mathbb{E}\big[\sup_{\theta \in \mathcal{N}} ||\frac{df(z;\theta)}{d\theta}||\big] < \infty$. Then $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, J^{-1})$ ■.

(MLE Variance Estimation Lemma) Using the conditions about for MLE asymptotics and the score function, define $\widehat{J}_1 = \frac{1}{n}\sum_{i=1}^n \dot{l}_{\widehat{\theta}}(z_i)\dot{l}_{\widehat{\theta}}(z_i)^{'}$; $\widehat{J}_2 = \frac{1}{n}\sum_{i=1}^n \ddot{l}_{\widehat{\theta}}(z_i)$; $\widehat{J}_3 = \frac{1}{n}\sum_{i=1}^n J(x_i, \widehat{\theta})J(x_i, \widehat{\theta})^{'} = \mathbb{E}[\dot{l}_{\widehat{\theta}}(z)\dot{l}_{\widehat{\theta}}(z)^{'}|x]$
Then $(\widehat{J}_2)^{-1} \xrightarrow{p} J^{-1}$. If there exists a neighborhood $\mathcal{N}$ of $\theta_0$ such that $\mathbb{E}[\sup_{\theta \in \mathcal{N}} ||\dot{l}_\theta||^2] < \infty$
then $(\widehat{J}_1)^{-1} \xrightarrow{p} J^{-1}$. If $J(x, \theta)$ is continuous at $\theta_0$ with probability 1 and $\mathbb{E}[\sup_{\theta \in \mathcal{N}} ||J(x, \theta)||] < \infty$
then $(\widehat{J}_3)^{-1} \xrightarrow{p} J^{-1}$. If the model is mispecified, then $(\widehat{J}_2)^{-1}\widehat{J}_1(\widehat{J}_2)^{-1}$ is consistent for the (asym) variance.

**(Asymptotics of GMM)** Suppose the GMM definition hold, the GMM consistency conditions hold, and we have iid data. Then if i) $\theta_0 \in \Theta$, ii) $g(z, \theta)$ is continuously differentiable in a neighborhood of $\theta_0(\mathcal{N})$, iii) $\mathbb{E}[||g(z, \theta_0)||^2] < \infty$, iv) $\mathbb{E}[\sup_{\theta \in \mathcal{N}} ||\frac{dg(z,\theta)}{d\theta}||] < \infty$, v) $G_w = G^{'}WG$ is nonsingular, where $G \equiv \mathbb{E}[\frac{dg(z,\theta_0)}{d\theta}]$

$$\implies \sqrt{n}(\theta - \theta_0) \xrightarrow{d} \mathcal{N}(0, (G_w)^{-1}GW\Omega WG(G_w)^{-1})$$

where $\Omega = \mathbb{E}[g(z, \theta_0)g(z, \theta_0)^{'}]$
**Proof:** Using the usual sample analog, by the FOC, symmetry, MVT (for some $\tilde{\theta}$ between $\theta_0$ and $\widehat{\theta}$)

$$0 = 2\widehat{G}_n^{'}(\widehat{\theta})\widehat{W}\widehat{g}_n(\widehat{\theta}) \implies 0 = \widehat{G}_n^{'}(\widehat{\theta})\widehat{W}\widehat{g}_n(\theta_0) + \widehat{G}_n^{'}(\widehat{\theta})\widehat{W}\widehat{G}_n(\tilde{\theta})(\widehat{\theta} - \theta_0)$$

$$\implies \sqrt{n}(\widehat{\theta} - \theta_0) = -\left(\widehat{G}_n^{'}(\widehat{\theta})\widehat{W}\widehat{G}_n(\tilde{\theta})\right)^{-1} \widehat{G}_n^{'}(\widehat{\theta})\widehat{W}\sqrt{n}\widehat{g}_n(\theta_0)$$

By CLT, $\sqrt{n}\widehat{g}_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, \Omega)$. To see details for the other term, look at the MLE Asymptotic proof. Note that by the triangle inequality, $||\widehat{G}_n^{'}(\widehat{\theta})\widehat{W}\widehat{G}_n(\tilde{\theta}) - G_w|| \leq ||\widehat{G}_n^{'}(\widehat{\theta})\widehat{W}\widehat{G}_n(\tilde{\theta}) - \widehat{G}_n^{'}(\theta_0)\widehat{W}\widehat{G}_n(\theta_0)|| + ||\widehat{G}_n^{'}(\theta_0)\widehat{W}\widehat{G}_n(\theta_0) - G_w||$, and the RHS converges in probability to 0 by iv), $\tilde{\theta} \xrightarrow{p} \theta_0$, continuity, LLN, and properties of of W. Therefore, by the CMT and Slutsky we have our desired result ■.

(**Consistency of GMM Variance**)* If the conditions for GMM Asymptotics hold, if $g(z,\theta)$ is continuous at $\theta_0$ and $\mathbb{E}[\sup \theta \in \mathcal{N} ||g(z,\theta)||^2] < \infty$ then V, a term using the sample analogs all the same components of the asymptotic variance of GMM, is consistent for $(G_w)^{-1}GW\Omega WG(G_w)^{-1}$.

(GMM Variance Efficiency Lemma) Assume the aforementioned necessary GMM conditions. The weight as $\Omega^{-1}$ yields the most efficient GMM Variance (asymptotic variance collapses to $(G^{'}\Omega^{-1}G)^{-1}$ ).

**Proof:** For $W \neq \Omega^{-1}$,can be seen by distributing each individual line that

$$\begin{aligned}(G_w)^{-1}GW\Omega WG(G_w)^{-1} - (G^{'}\Omega^{-1}G)^{-1} &= (G_w)^{-1}G^{'}W\Omega W(G_w)^{-1} - G_w(G^{'}\Omega^{-1}G)^{-1}G_w)(G_w)^{-1}\\ &= (G_w)^{-1}G^{'}W\Omega^{.5}(I - \Omega^{-.5}G(G^{'}\Omega^{-1}G)^{-1}G^{'}\Omega^{-.5})\Omega^{.5}WGG_w)^{-1}\\ &= A(I-B)A^{'}\end{aligned}$$

Where A and B are appropriately defined matricies. Since B is symmetric and idempotent, so is I-B. Thus

$$A(I-B)A^{'} = A(I-B)(I-B)^{'}A^{'} = (A(I-B))(A(I-B))^{'}$$

which is psd ■.

However, $\Omega$ depends on $\theta_0$, which is obviously problematic for estimation. Instead, we will look at at "2-step" GMM, leading to a generalization of "k-step" GMM. This process basically involves using several GMM estimations to yield desirable asymptotic results.

(2-Step and K-Step GMM) For 2-Step GMM, first let $\widehat{W}$ be the identity matrix. Let $\tilde{\theta}^*$ solve the GMM with this weight matrix. Then set $\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(z_i\tilde{\theta}^*)g(z_i\tilde{\theta}^*)^{'}$. Note that this is consistent for $\Omega$ (shown in previous results). Then construct a new GMM estimation $(\widehat{\theta})$ using $\widehat{W} = \Omega^{-1}$, and this leads to the efficient result of $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, (G^{'}\Omega^{-1}G)^{-1})$. The 2-step estimator also satisfies $n\widehat{g}(\widehat{\theta})^{'}\widehat{W}\widehat{g}(\widehat{\theta}) \xrightarrow{d} \chi^2_{dim(\theta_0)}$. 3-Step GMM would be using the 2-Step estimator to construct a new estimation of $\Omega$, and by iteration you can define K-Step GMM. As $K \to \infty$, this becomes the continuously uppdated estimator (CUE) for GMM which solves $\min_{\theta \in \Theta} \widehat{g}(\theta)^{'}\widehat{\Omega}^{-1}(\theta)\widehat{g}(\theta)$

Let $m_\theta = \mathbb{E}[\frac{dg(z,\theta_0)}{d\theta}]^{'}W\frac{dg(z,\theta_0)}{d\theta} = G^{'}W\frac{dg(z,\theta_0)}{d\theta}$. Similarly let $m = G^{'}Wg(z,\theta_0)$ and recall that $J = \mathbb{E}[\dot{l}_{\theta_0}\dot{l}^{'}_{\theta_0}]$.

(**MLE vs GMM Asymptotics**) Given MLE/GMM asymptotics conditions (notably, observational distribution is in family $f(X;\theta)$), if for all $\theta \in \mathcal{N}$ (a neighborhood of $\theta_0$) $\int \sup_{\theta\in\mathcal{N}} ||g(z,\theta)||^2 f(z;\theta)dz$ is bounded then the difference in the asymptotic variance for GMM and MLE $(\mathbb{E}[m_\theta]^{-1}\mathbb{E}[mm^{'}]\mathbb{E}[m_\theta]^{-1} - J^{-1})$ is psd.

**Proof:** Note that $\mathbb{E}[g(z,\theta_0)] = 0$ iff $0 = \int g(z,\theta)f(z;\theta)dz\Big|_{\theta=\theta_0}$. Thus by premultiplying $G^{'}W$

$$0 = \frac{d}{d(\theta)}\int g(z,\theta)f(z;\theta)dz\Bigg|_{\theta=\theta_0} = \left(\int \frac{dg(z,\theta)}{d(\theta)}f(z;\theta)dz + \int g(z,\theta)\frac{df(z;\theta)}{d(\theta)}dz\right)\Bigg|_{\theta=\theta_0} = \mathbb{E}[m_\theta] + \mathbb{E}[m\dot{l}^{'}_{\theta_0}]$$

$$\implies \mathbb{E}[m_\theta]^{-1}\mathbb{E}[mm^{'}]\mathbb{E}[m_\theta]^{-1} - \mathbb{E}[\dot{l}_{\theta_0}\dot{l}^{'}_{\theta_0}]^{-1} = \mathbb{E}[m\dot{l}^{'}_{\theta_0}]^{-1}\mathbb{E}[mm^{'}]\mathbb{E}[m\dot{l}^{'}_{\theta_0}]^{-1} - \mathbb{E}[\dot{l}_{\theta_0}\dot{l}^{'}_{\theta_0}]^{-1} = \mathbb{E}[m\dot{l}^{'}_{\theta_0}]^{-1}\mathbb{E}[UU^{'}]\mathbb{E}[m\dot{l}^{'}_{\theta_0}]^{-1}$$

which is psd (where $U = m - \mathbb{E}[m\dot{l}^{'}_{\theta_0}]\mathbb{E}[\dot{l}_{\theta_0}\dot{l}^{'}_{\theta_0}]^{-1}\dot{l}_{\theta_0}$). So the result follows by distributing terms ■.

# Bootstrap

The **Bootstrap** is a model that embodies frequentest philosophy that involves estimating the distribution of an estimator or a test statistic by resamping the data. First, we will define some useful notation. Let $F_0$ denote the population distribution (population CDF) of (Y,X). We say that we are interested in the "true parameter" $\theta(F_0) = \theta_0$. Typically, we use $F_n$, and estimator of $F_0$, to estimate the true parameter by $\theta(F_n) = \widehat{\theta}_n$. We can think of the test statistic $T_n = T_n(X_1, \ldots, X_n)$ as being some function of $\widehat{\theta}_n$. This allows for this formulation of a finite sample CDF of $T_n$: $G_n(t, F_0) \equiv \mathbb{P}[T_n \leq t]$. Thus, we can generalize this to $G_\infty(\cdot, F)$ as the asymptotic distribution from F. We say that $T_n$ is **asymptotically pivotal** if its asymptotic distribution doesn't depend on F (define pivotal similarly for $G_n$).

Executing the bootstrap can be generalized by the following three steps

**Step 1** Draw a random sample of observations $\{X_i^*\}_{i=1}^n$ from $F_n$

**Step 2** Compute $T_n^*$ from $\{X_i^*\}_{i=1}^n$

Repeat Steps 1 and 2 $B$ times (Monte Carlo procedure)

**Step 3** Choose $c^*$ to be the $1-\alpha$ quartile of all the calculated $T_n^*$ and create a confidence interval

In terms of how $F_n$ is chosen it can broadly be non-parametric ($\frac{1}{n}\sum_{i=1}^n \mathbb{1}[X_i \leq x]$) or parametric ($F(\cdot; \widehat{\theta}_n)$), where both are estimates of $F_0(\cdot) = F(\cdot; \theta_0)$. Roughly speaking, we say that $G_n(\cdot, F_n)$ is consistent if it converges to $G_\infty(\cdot, F_0)$. More formally

**(Consistency of Bootstrap definition)** Let $P_n$ denote the joint probability distribution of $\{X_i\}_{i=1}^n$. The bootstrap estimator $G_n(\cdot, F_n)$ is consistent for $G_\infty(\cdot, F_0)$, where $F_0 \in \mathcal{F}$ (finite dimensional family indexed by $\theta$) if for all $\epsilon > 0$ we have $\lim_{n\to\infty} P_n[\sup_t |G_n(t, F_n) - G_\infty(t, F_0)|] = 0$.

**(<u>Sufficient Bootstrap Consistency</u>)** $G_n(\cdot, F_n)$ is consistent if for each $\epsilon > 0, F_0 \in \mathcal{F}$, and metric $\rho$ on $\mathcal{F}$ i) $\lim_{n\to\infty} P_n[\rho(F_n, F) > \epsilon] = 0$, ii) $G_\infty(t, F)$ is continuous in t for each $F \in \mathcal{F}$, and iii) for any t and sequence $\{\tilde{F_n}\} \in \mathcal{F}$ such that $\lim_{n\to\infty} \rho(\tilde{F_n}, F_0) = 0$, $G_n(t, \tilde{F_n}) \to G_\infty(t, F_0)$.

**(<u>T-Statistic Bootstrap Consistency</u>)** Let $\{X_i\}_{i=1}^n$ be an iid sample and $\{X_i^*\}_{i=1}^n$ be the bootstrap sample. For a sequence of functions $g_n$ and a sequence of numbers $t_n$ and $\sigma_n$, let $\overline{g}_n = \frac{1}{n}\sum_{i=1}^n g_n(X_i)$ with $T_n = \frac{\overline{g}_n - t_n}{\sigma_n}$ and $T_n^* = \frac{\overline{g}_n - \overline{g}_n^*}{\sigma_n}$ (using $\overline{g}_n^* = \frac{1}{n}\sum_{i=1}^n g_n(X_i^*)$). With $P_n^*$ being the probability distribution induced by the bootstrap, then $G_n^*(t) = P_n^*(T_n^* \leq t)$ is consistent for $G_n(t) = P_n(T_n \leq t)$ iff $T_n \xrightarrow{d} \mathcal{N}(0,1)$.

In some settings $|G_n(\cdot, F_n) - G_n(\cdot, F_0)|$ converges faster than $|G_\infty(\cdot, F_n) - G_n(\cdot, F_0)|$, which can be problematic for how we are conceptualizing the ordering of convergence. To address this, we will define the Smooth Function Model (**SFM**) by letting $T_n = \sqrt{n}[H(\overline{Z}) - H(\mu_Z)]$, where $\overline{Z} = (\overline{Z}_1, \ldots, \overline{Z}_J)'$, $\mu(Z) =$

$\mathbb{E}[Z]$, and a function $H$ (and $Z$) satisfying some smoothness conditions[8] (each $\overline{Z}_J$ is the sample mean of $Z_j(X)$). These conditions are more than sufficient to show that $T_n = \partial H(\mu_Z)' \sqrt{n}(\overline{Z} - \mu_Z) + o_p(1)$ (where $\partial H(z) = \partial H(z)/\partial z$). By the CLT $T_n$ follows a standard normal asymptotic CDF, so by the Berry-Esséen theorem $\sup_t |G_n(t, F_0) - G_\infty(t, F_0)| = O(\frac{1}{\sqrt{n}})$. Also, under SFM we have bootstrap consistency by the last theorem, and we can also show the stronger result that $\sup_t |G_n(t, F_n) - G_\infty(t, F_0)| \xrightarrow{a.s} 0$.

(**SFM Bootstrap Approximation**) Under the Cramér condition[9], from a Taylor-Edgeworth expansion

$$G_n(t, F_0) = G_\infty(t, F_0) + n^{-.5} g_1(t, F_0) + n^{-1} g_2(t, F_0) + n^{-1.5} g_3(t, F_0) + O(n^{-2})$$
$$G_n(t, F_n) = G_\infty(t, F_n) + n^{-.5} g_1(t, F_n) + n^{-1} g_2(t, F_n) + n^{-1.5} g_3(t, F_n) + O(n^{-2})$$

uniformly in t, where the $g_i(\cdot)$ are various functions of moments

By $\Delta g_i = g_i(t, F_n) - g_i(t, F_0)$, this gives us $G_n(t, F_n) - G_n(t, F_0) = G_\infty(t, F_n) - G_\infty(t, F_0) + n^{-.5}\Delta g_1 + O(n^{-1.5})$ almost surely uniformly over t. From the continuity of $G_\infty$ and $F_n - F_0 = O(n^{-.5})$ almost surely uniformly over the support of $F_0$, the bootstrap has an error of size $O(n^{-.5})$. This means the bootstrap is as good of an approximation of the finite sample distribution as the asymptotic distribution.

(Intuition about the Bootstrap) Here is some general guidance on how/when to implement the bootstrap.

**1.** Use bootstrap to estimate distribution of asymptotically pivotal statistic or its critical value whenever available. Don't use bootstrap on a non-asymptotically pivotal statistic when something asymptotically pivotal is available.

**2.** Bootstrap models on dependent data, semi-/non-parametric estimators, and/or non-smooth estimators require extra caution.

[8] H(z) is 6-times continuously partially differentiable with respect to any mix of elements in z in a neighborhood of $\mu_Z$, $\partial H(z)/\partial z \neq 0$, and the expectation of the product of any 16 components of Z exists ("16" from the Edgeworth-Taylor expansion)

[9] $\limsup_{||t|| \to \infty} |\mathbb{E}[\exp(it'Z)]| < 1$

## Non-Parametric Estimation

First, some more trivial matters to establish for total clarity. We consider $U$ in these settings to be unobserved heterogeneity. In many of the following settings, we will be interested in a moment function $Y = m(X, U) = m(X, \theta, U)$, where $m$ and the conditional distribution of $U|X$ are known for $\theta \in \Theta$. Also, note that for conditional distributions given $X$ and $U \perp\!\!\!\perp X$, the distributions are not equivalent for different realizations of $X$. Finally, $\mathbb{E}[U|X] \equiv \mathbb{E}[U|X = x]\ \forall\, x \in \operatorname{supp}(X)$.

Referring to our definition of $m(\cdot)$ above, if the parameter space $\Theta$ is infinite dimensional then the model is **non-parametric**. If instead $\Theta = M \cup I$, where $M$ is finite dimensional and $I$ is infinite dimensional, then the model is **semi-parametric**. To make these definitions a bit more concrete, consider the following examples. Assuming a linear, separable model $Y = X\beta + U, U \perp\!\!\!\perp X,$ and $U \sim \mathcal{N}(\beta, \sigma^2)$, then we can learn about the $U|X$ distribution by the finite parameter set $(\beta, \sigma^2)$, so the model is **parametric**. If instead we have a linearly separable model but the only other think we know is $\mathbb{E}[U|X] = 0$, then the model is semi-parametric. One final example will motivate how we circumvent potential application issues. Consider a partially linear model of $Y = X\beta + \lambda(Z) + U$, where $\mathbb{E}[U|X, Z] = 0$ and $\lambda(\cdot)$ is a parametric function. Then by L.I.E we have $\mathbb{E}[Y|Z] = \mathbb{E}[X|Z]\beta + \lambda(Z) + 0$ so therefore $Y - \mathbb{E}[Y|Z] = (X - \mathbb{E}[X|Z])\beta + U$. Denote this expression by (*). Our "idea" is to run non-parametric estimation of $\mathbb{E}[Y|Z]$ and $\mathbb{E}[X|Z]$, then plug this into $(*)$ and then run linear regression. This would manifest in the following way

$$\begin{aligned} \textbf{X discrete} \quad &\implies \mathbb{E}[Y|X = x] = \frac{\mathbb{E}[X \cdot \mathbb{1}_{X=x}]}{\mathbb{P}(X = x)} = \frac{\mathbb{E}[X \cdot \mathbb{1}_{X=x}]}{\mathbb{E}[\mathbb{1}_{X=x}]} \\ \textbf{X continuous} \quad &\implies \mathbb{E}[Y|X = x] = \int x f(y|x) dy = \int \frac{y f(y|x) f(x)}{f(x)} dy = \int \frac{y f(y, x)}{f(x)} dy \end{aligned}$$

Assume $X$ is univariate continuously distributed with density $f(\cdot)$. We want an estimate $\widehat{f}(x_0) \xrightarrow{p} f(x_0)$. We can exploit the fundamental theorem of calculus and mean value theorem ($\exists c \in [a, b]$ s.t $f'(c)(b - a) = f(b) - f(a)$), specifically that for small $b - a$ its midpoint will be an acceptable approximation of $c$, and get the *naive estimate* for the density by

$$\begin{aligned} \mathbb{P}(X \in [a, b]) = \int_a^b f(x)dx &\implies \mathbb{P}(X \in [x_0 \pm h]) = \int_{x_0-h}^{x_0+h} f(x)fx \approx 2h \cdot f(x_0) \\ &\implies \widehat{f}(x_0) = \mathbb{P}(X \in [x_0 \pm h])/2h = \frac{1}{2h}\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i \in [x_0 \pm x]} \end{aligned}$$

However, the true form is $f(x_0) = \lim\limits_{h \to 0} \mathbb{P}(X \in [x_0 \pm h])/2h$. In the naive estimate, $h$ is fixed. We need $h$ to be a converging term. More specifically, to satisfy the LLN, we our "bandwidth object" $h_n$ to converging to 0 at a more measured pace relative to the growth of $n$ (i.e $h_n \cdot n \to \infty$). So $\widehat{f}(x_0) = \frac{1}{h_n n} \sum_{i=1}^{n} \mathbb{1}_{x_i - x_0 \in [-h_n, h_n]}$ can be considered as an alternative. More generally, we can consider a class of **kernel** functions to replace the indicator function that take the form $K(\frac{x_i - x_0}{h_n})$. Often, we will want to consider more restricted classes of estimators, for instance $K(\cdot)$ that satisfy $\int \boldsymbol{K(\psi)d\psi = 1, \int \psi^2 K(\psi)d\psi < \infty}, |\psi| \cdot K(\psi) \to 0\,(\psi \to \infty), \sup_{\psi}, \text{and} \int K^2(\psi)d\psi < \infty$. We will show that the MSE $\to 0$ with the help of the first two assumptions along with $K(\cdot)$ being symmetric (i.e. these are sufficient conditions for consistency)

**Proof:** Want to show (WTS) that MSE$(\widehat{f}(x_0), f(x_0)) = \text{bias}^2 + \text{var} \to 0$. We have

$$\text{Bias} = \mathbb{E}[\frac{1}{h_n}K(\frac{x_i - x_0}{h_n})] - f(x_0) = \int \frac{1}{h_n}K(\frac{x - x_0}{h_n})f(x)dx - f(x_0) = \int K(\psi)f(x_0 + \psi h_n)d\psi - f(x_0)$$

using a change of variables with $\psi = \frac{x - x_0}{h_n}$ (so $d\psi = dx/h_n$). By a Taylor expansion

$$\int K(\psi)[f(x_0) + h_n\psi f'(x_0) + \frac{1}{2}(h_n\psi)^2 f''(x_0) + \dots]d\psi = \int K(\psi)f(x_0)d\psi + \int K(\psi)h_n\psi f'(x_0)d\psi + \int K(\psi)\frac{1}{2}(h_n\psi)^2 f''(x_0)d\psi + \dots$$

If we assume $f(\cdot)$ 2nd order cts diff at $x_0$ with $f''(\cdot) < \infty$, and the initial assumptions[10], the above becomes

$$f(x_0) + f'(x)h_n \int K(\psi)\psi d\psi + \frac{1}{2}h_n^2 f''(x_0)\int K(\psi)\psi^2 d\psi + \dots = f(x_0) + 0 + O(h_n^2) + O(h_n^4) + \dots$$

By substituting the simplification of the Taylor expansion, bias$^2 = O(h_n^4)$ ($h_n^p > h_n^q \;\forall$ int. $p > q > 0$). Also,

$$\text{var}(\widehat{f}(x_0)) = \frac{1}{nh_n}f(x_0)\int K^2(\psi)d\psi + o(\frac{1}{nh_n}) \implies \text{var}(\widehat{f}(x_0)) = O(\frac{1}{nh_n})$$

by a similar calculation. Thus, MSE$= O(h_n^4) + O(\frac{1}{nh_n}) \to 0$ by our assumptions on $h_n$ $\blacksquare$.

Implicit in the proof is the result that $h_n^2\sqrt{nh_n} \to 0 \implies \sqrt{nh_n}(\widehat{f}(x_0) - \mathbb{E}[\widehat{f}(x_0)]) \xrightarrow{d} \mathcal{N}(0, f(x_0)\int K^2(\psi)d\psi)$

If we define *risk* as MSE, then we showed it's $O(h_n^4) + O(\frac{1}{nh_n})$. Let risk at a point $(x)$ and integrated risk be

$$R_x = \frac{1}{4}\sigma_k^4 h_n^4 (f''(x))^2 + f(x)\int K^2(x)dx + O(\frac{1}{n}) + O(h_n^6)$$

$$\underline{\textbf{and}}\;\; R = \int R_x dx = \frac{1}{4}\sigma_k^4 h_n^4 \int (f''(x))^2 dx + \int K^2(x)dx + O(\frac{1}{n}) + O(h_n^6)$$

where $\sigma_k^2 = \int x^2 K(x)dx$ and we assume $\int f(x)^2 dx$ is bounded.

Everything til now has assumed that $X$ is univariate. If we relax this assumption and allow $X \in \mathbb{R}^d, d \geq 1$, if we look in a "ball" (multi-dimensional neighborhood) of $x_0$, we now have an analogous result that began with this analysis by $\mathbb{P}(x \in B_n(x_0)) \approx ch_n^d f(x_0) \implies \frac{1}{nh_1 \dots h_d}\sum_{i=1}^n \Pi_{j=1}^d K(\frac{x_{ij} - x_{0j}}{h_j})$. So it's useful to define $K_d = \Pi_{j=1}^d K(\cdot)$. Thus, $f(x_0, y) = \frac{1}{nh_d}\sum_{i=1}^n K(\frac{y_i - y_0}{h_n})K_d(\frac{x_i - x_0}{h_n})$

[10] bolded assumption 1, symmetry, and bolded assumption 2 for the first, second, and third term respectively

## Causal Graphs

Consider a graph with a disjoint partition that includes the set of nodes $X_C$ and $X_E$. Consider the object $\mathbb{P}(X_E|\text{do}(X_c = x_c)$, where this represents a probability object conditional on fixing/forcing the sub-population to be $x_c$ (creating a sub-set, not selecting one from the default). If we wanted to think of this in a visual/graph sense, this transformation can be thought of as first eliminating any arrows coming into nodes in $X_c$, then setting their values to $x_c$ (accordingly) and calculating the resulting distribution of $X_E$. If using $x' \neq x_c$ yields a different value for this object, then say $X_C$ has a *causal effect* on $X_E$. To be explicitly clear about notation, $\mathbb{P}(X_E|X_c = x_c)$ can be considered probabilistic conditioning; this is simply a population object (the conditional distribution). $\mathbb{P}(X_E|\text{do}(X_c = x_c)$ is causal conditioning; this is a *counterfactual* object (something we do not observe in the data).

Explicitly define $\mathbb{P}(Y|\text{do}(T = t)) = \sum_x \mathbb{P}(Y|T = t, \text{Pa}(T) = x)\mathbb{P}(\text{Pa}(T) = x)$, where $\text{Pa}(\cdot)$ is parents[11] of $X$

To make use of this object, which we can think of as the *causal effect* of $T$ on $Y$, we need to define some additional terminology so we can represent the causal effect in a way that works in applied settings. For all of these definitions, assume that $G$ is an acylclic, directed graph with vertex set $V$ that includes single-nodes $T, Y$, and let $W$ contain all other verticies ($W = V \cap (T \cup Y)$. Define this as *causal graph* (CG) vertex partitioning. A *collider* is a vertex on a path that has at least two incoming edges (on that path). Say that $T$ and $Y$ are **backdoor d-connected** by a vertex set $Z \subset W$iff $\exists$ a simple, undirected path $U \equiv T \leftarrow \{u_i\}_{i=1}^{n} \rightarrow Y$ such that for $\mathcal{C}$, the set of verticies that are colliders with respect to $U$ **i)** $\forall j$ s.t $u_j \in \mathcal{C}$, either $u_j$ or a descendent of $u_j$ is in $Z$ and **ii)** $\forall j$ s.t $u_j \notin \mathcal{C} \implies u_j \notin Z$. Say that $T$ and $Y$ are **backdoor d-separated** by $Z$ iff they are not backdoor d-connected by $Z$. We also can say that if $T$ and $Y$ and d-separated by $Z$, then they are *blocked* by $Z$ (the converse does not necessarily hold in general, but for our purposes we only care when it does). Finally, $\big((Y \perp\!\!\!\perp T) \mid C\big)_{\text{BD}}$ if $C$ *blocks* every path from $T$ to $Y$ that includes an arrow into $T$. Now we present the main result, which keep in mind relies on all previous definitions of objects (e.g. $Z \subseteq W$). This result is important because it allows for the mathematical formulation of the causal effect of $T$ on $Y$ without using counterfactual objects.

**Theorem: (Backdoor Criterion)** If $T$ and $Y$ are backdoor d-separated by $Z$ and no node in $Z$ is a descendent of $T$ then

$$\mathbb{P}(Y|\text{do}(T = t)) = \sum_z \mathbb{P}(Y|T = t, Z = z)\mathbb{P}(Z = z)$$

[11] Parents are direct ancestors (i.e. they are directly adjacent)

**Proof (#1):** We have by the definition $\mathbb{P}(Y|\text{do}(T=t)) = \sum_x \mathbb{P}(Y|T=t, \text{Pa}(T)=x)\mathbb{P}(\text{Pa}(T)=x)$. Denote this (1). To simplify notation a bit, consider instead $\text{Pa}(T) = S$. Further, consider that we don't want to be restricted to the entire parent set. Instead, we would like to condition on a different set, $Z$. Therefore, we first can rewrite (1) to get a simpler, equivalent[12] definition (2)

$$\mathbb{P}(Y|\text{do}(T=t)) = \sum_{x \in S} \mathbb{P}(Y|t,x)\mathbb{P}(x)$$

Consider the following two assumptions: **a)** $(Y \perp\!\!\!\perp S)|(T,Z)$ and **b)** $(T \perp\!\!\!\perp Z)|S$. If a) and b) hold, we can then say (2) is equivalent to the following

$$\mathbb{P}(Y|\text{do}(T=t)) = \sum_{x \in S} \mathbb{P}(x) \sum_{z \in Z} \mathbb{P}(Y|t,z)\mathbb{P}(z|t,x) \implies \mathbb{P}(Y|\text{do}(T=t)) = \sum_{z \in Z} \mathbb{P}(Y|t,z)\mathbb{P}(z)$$

because the first equation follows from conditioning on $Z$ using assumption a) and the second equation follows because assumption b) implies $\mathbb{P}(z|t,x) = \mathbb{P}(z|x)$. Note that the second equation, which we can denote (3), is equivalent to what we are trying to prove. So if we can prove that the back-door criteria conditions implies a) and b), we are done! b) follows plainly from the "no descendent of $T$" assumption because after conditioning on the parents of $T$, $Z$ will not be able to have a dependent relationship on $T$ unless it was a descendant (this is also known as the Markov property). To see that the backdoor d-separated assumption implies a), consider a brief proof by contradiction. First, assume there is a path between a parent of $T$ and $Y$ that is not blocked by $T$ or $Z$. It cannot be a direct path because then it would be blocked by $T$ (it would need to go through $T$ to get to $Y$ since its a parent of $T$). However, it cannot also be a backdoor path because $Z$ blocks every path going into $T$ (which would start with a node in the parent set). Since there are no other path types, our original assumption must have been invalid. Therefore, the backdoor criteria conditions imply a), and we have sufficient conditions for equivalence of (1) and (3) ■.

**Proof (#2):** This proof is similar but the approach is a bit different. However, we will rely on notation and results from the previous proof to make it easier. Again $\text{Pa}(T) = S$. Also, recall assumptions a) and b), and that we proved that the backdoor criteria implies that these hold. Finally, consider again the equation (1). Now, consider that if we have a probability object $P(A)$, for a non-empty vertex set $B$ this is equivalent to $\sum_{b \in B} P(A,b)$ because its simply diving up the probability object into parts. So therefore (1)

[12] Note that $Y$ here could be written as $Y = y$, and further that all lowercase lettering imply something similar. For instance, lowercase $z$ relates to the larger set $Z$. If there is any confusion, compare the result of the theorem to (3) and note they are equivalent

is equivalent to

$$\mathbb{P}(Y|\text{do}(T=t)) = \sum_x \left[\mathbb{P}(S=x)\sum_z \left(\mathbb{P}(Y, Z=z|T=t, S=x)\right)\right]$$

By the definition of joint probability

$$\mathbb{P}(Y, Z|T=t, S=x) = \mathbb{P}(Y|T=t, S=x, Z=z)\mathbb{P}(Z=z|T=t, S=x)$$

which means

$$\mathbb{P}(Y|\text{do}(T=t)) = \sum_x \left[\mathbb{P}(S=x)\sum_z \left(\mathbb{P}(Y|T=t, S=x, Z=z)\mathbb{P}(Z=z|T=t, S=x)\right)\right]$$

invoking assumption a)

$$\mathbb{P}(Y|\text{do}(T=t)) = \sum_x \left[\mathbb{P}(S=x)\sum_z \left(\mathbb{P}(Y|T=t, Z=z)\mathbb{P}(Z=z|T=t, S=x)\right)\right]$$

and invoking assumption b)

$$\mathbb{P}(Y|\text{do}(T=t)) = \sum_x \left[\mathbb{P}(S=x)\sum_z \left(\mathbb{P}(Y|T=t, Z=Z)\mathbb{P}(Z=z|S=x)\right)\right]$$

Since $\sum_x \mathbb{P}(S=x)\left(\mathbb{P}(Y|T=t, Z=Z)\mathbb{P}(Z=z|S=x)\right) = \mathbb{P}(Z=z)$

$$\mathbb{P}(Y|\text{do}(T=t)) = \sum_z \mathbb{P}(Y|T=t, Z=z)\mathbb{P}(Z=z)$$

as desired ■.

*A digression on cycles and simple paths*: These results rely on the definitions that assume a simple path. But one may wonder why we are able to do this. We will consider a formal and informal proof of this fact. Consider first a proof by contradiction. That is, given G, a DAG with CG partitioning, there is a non-simple, undirected path $T \leftarrow \{u\}_{i=1}^{n} \rightarrow Y$ that is not blocked by $Z \subset W$, which is a vertex set that backdoor d-separates $T$ and $Y$. The implication of this is that this path gets around the blocking induced by $Z$ by repeating vertices. But this is a contradiction; because $G$ is a DAG, such a path would either have to eventually arrive at a vertex in $Z$ or already be blocked off at an earlier point by a vertex in $Z$, which is backdoor d-separating. In other words, in a DAG, every non-simple undirected path that starts with an arrow into $T$ and ends in an arrow into $Y$ has an an analogous simple path, and by analogous we mean that the cycle is blocked off by whatever vertex completes the cycle (for example if a path that contains the sequence $A \to B \to C$ is blocked, then $A \to B \leftarrow D \to E \leftarrow B \to C$ is also blocked, otherwise there would be a contraction). An informal proof can be given in the form of an analogy of driving. If one is trying to drive from point $A$ and point $C$, lets say they have a choice along the way of turning left and going into a "one horse town" that has one street which is a dead end or turning right and going onto the highway to point $C$. If they turn left and go into the town, they will have to turn around and come back. So there is no point in considering that route. Further, if the right turn onto the highway is blocked (let's say by construction), turning left will not lead the car into a route where you can get to the destination. To distill this issue even further, the aspect of simple paths comes down to **distance** vs. **displacement**. For the backoor criteria, we do not care about the distance of the path. We only care about the displacement: the *net* movement. Because we assume a DAG, considering cycles can only add to the distance, but the displacement can never be different if we only consider simple paths.

# References

**Math Review**

Pretty much everything from Dr. Masten. Dr. Cassidy influenced the ordering of some topics

Markov's Inequality from Dr. Cassidy

Delta Method from Dr. Cassidy

Some convexity background from Po-Shen Loh (CMU)

Proof of Jensen's Inequality from Kai-Seng Chou (CUHK) and Lei Mao

Proof of LLN from Dr. Cassidy

L-L CLT univarite from Dr. Rosen

L-L CLT multivariate from Dr. Masten and Dr. Rosen

Transpose from Jeff Hefferon (Saint Michael's)

Positive definite and projection stuff from Dr. Cassidy

**Hypothesis Testing**

Pretty much everything from Dr. Masten

**M-Estimation**

Consistency of m-estimators from Newey-McFadden; intuition from Dr. Cassidy and Dr. Rosen

MLE ID Lemma from Dr. Masten and Dr. Rosen

GMM ID Lemma from Newey-McFadden

GMM Consistency from Newey-McFadden

Asymptotics of m-estimators from Dr. Rosen and Newey-McFadden

MLE Asymptotics from Dr. Cassidy (conditions from Newey-McFadden)

GMM Asymptotics from Dr. Rosen and Newey-McFadden

Consistency of GMM Variance from Newey-McFadden

GMM Variance Efficency Lemma from Wikipedia page on GMM

2-Step and K-Step GMM from Dr. Cassidy and Dr. Rosen

MLE vs GMM Asymptotics from Dr.Rosen and Newey-McFadden

**Bootstrap**

Every major tenet is from Bootstrap Consistency Horowitz (2001) *Handbook of Econometrics* Bootstrap chapter. Specifically Definition 2.1 and Theorems 2.1,2.2,and 3.1.

Dr. Rosen influenced the selection/ordering of material and provided more context about the SFM