The Population and Personalized AUCs

August 2, 2022

Abstract.
keywords: AUC, Simpson's paradox, clustered data, confounding, causal inference

1 Introduction

The AUC is a way to evaluate a predictor of a binary outcome. The AUC is the probability that the value of a random sampled predictor from one of the outcome classes is less than an independent sample of the predictor from the other outcome class. There are several ways to generalize the AUC to accommodate clustered data. What we refer to as the "population AUC" appears to be the most commonly studied. The population AUC evaluates the predictor's typical effect on an entire population, in a sense clarified further below.

While the population AUC is an important part of understanding the usefulness of a predictor, the medical field has lately focused on personalizing treatment. For example, in 2018 the National Academy of Medicine concluded: "The individuality of the patient should be at the core of every treatment decision. One-size-fits-all approaches to treating medical conditions are inadequate; instead, treatments should be tailored to individuals based on heterogeneity of clinical characteristics and their personal preferences."

We examine an "individual AUC" in conjunction with the population AUC. These two evaluations may give different accounts of the usefuless of a marker. In the extreme case, the phenomenon known as Simpson's paradox may occur: The individual AUC may be nearly uninformative while the population AUC is nearly completely predictive, or vice versa. Modern accounts of Simpson's paradox, working in the framework of causal inference, delineate situations in which the individual AUC is appropriate, and other situations in which the population AUC is appropriate.

[[literature review.]] [6] proposes a nonparametric estimator for the variance of an estimator for the population AUC. We give an alternate derivation here. We also clarify the statistical model and target of inference. [5] analyzes population and individual versions of the ROC curve and suggests parametric and nonparametric estimators. In principle, these may be used to obtain estimates of the population and individual AUCs discussed here. The analysis here is nonparametric and avoids the inefficiency introduced by first estimating the entire ROC curve.[[don't want journal to think this is a small extension of that work. don't want referee to require simulation comparing the two estimation strategies.]][[add more to literature review]]

[[give outline of rest of paper]]

2 Main

2.1 Setting and Notation

Given are scalar observations on two classes of individuals $X_1, X_2, \ldots, X_M$ IID as F_X and $Y_1, Y_2, \ldots, Y_N$ IID as F_Y. Define the function $\psi : (x, y) \mapsto \{x < y\} + \frac{1}{2}\{x = y\}$. The AUC is defined as

$$\theta = E(\psi(X, Y)) = P(X < Y) + \frac{1}{2}P(X = Y)$$

where X and Y are independent draws from F_X and F_Y. We refer to the function ψ as the kernel. The AUC is often used to evaluate how effectively the markers distinguish the 2 classes. The AUC is close to $1/2$ when the distinction is poor. In the extreme case that $F_X = F_Y$, $\theta = 1/2$. The AUC is farther from $1/2$ when the distinction is better. In the extreme, there is a number $c \in \mathbb{R}$ such that always $X < c$ and $Y > c$, and then $\theta = 1$. We informally refer to the two classes as "control" and "case", though in a given application they might represent any other binary classification, e.g., non-diseased and diseased. Switching the observations designated "control" and "case" reflects the AUC across $1/2$, $AUC \mapsto 1 - AUC$, so $|AUC - \frac{1}{2}|$ is often of greater interest than the AUC itself.

We extend the AUC to accommodate 1) vectors of case and control observations and 2) dependence between case and control observations. Examples of data of this type are:

1. The predictors are longitudinal measurements of tumour antigens (CEA, CA15-3, TPS), and the outcomes are progression or non-progression of breast cancer [2].

2. The predictors are longitudinal measurements of levels of vascular enothelial growth factor and a soluble fragment of Cytokeratin 19, and the outcomes are progression or non-progression of non-small cell lung cancer [8].

3. give [[pre/post treat example. maybe a social sciences example]]

Let (X, Y, M, N) be a random vector with joint distribution P such that X is a vector of length M and Y is a vector of length N, where M and N are counting numbers.

$$\begin{aligned} &(X, Y, M, N) \sim P \\ &X = (X_1, \ldots, X_M), Y = (Y_1, \ldots, Y_N) \\ &M, N \in 1, 2, 3, \ldots. \end{aligned} \tag{1}$$

Extend the AUC kernel to vector arguments as

$$\psi(x, y) = \psi((x_1, \ldots, x_m), (y_1, \ldots, y_n)) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\{x_i < y_j\} + \frac{1}{2}\{x_i = y_j\}\right). \tag{2}$$

We define the individual AUC as

$$\theta_{11}(P) = E\left(\frac{\psi(X, Y)}{MN}\right). \tag{3}$$

With $(X_1, Y_1, M_1, N_1), (X_2, Y_2, M_2, N_2)$ being two independent draws from P, we define the population AUC as

$$\theta_{12}(P) = \frac{E\psi(X_1, Y_2)}{E(M_1)E(N_2)} \tag{4}$$
$$(X_1, Y_1, M_1, N_1), (X_2, Y_2, M_2, N_2) \overset{\text{IID}}{\sim} P.$$

The individual AUC may be undefined if M or N can take the value 0 with positive probability, which is the reason for restricting them to counting numbers. The population AUC may still be well-defined and when analyzed without regard to the individual AUC the possibility of $M = 0$ or $N = 0$ may be allowed [6]. In an applications where $M = 0$ or $N = 0$ is possible, our analysis is therefore conditional on $M > 0, N > 0$, a sub-population in which all clusters have at least 1 case and 1 control observation.

For estimation, suppose a sample is given,

$$(x_1, y_1, m_1, n_1), \ldots, (x_I, y_I, m_I, n_I) \overset{\text{IID}}{\sim} P.$$

An unbiased estimator of θ_{11} is

$$\hat{\theta}_{11} = \frac{1}{I}\sum_{i=1}^{I} \frac{\psi(x_i, y_i)}{m_i n_i}.$$

A consistent estimator of θ_{12} is

$$\hat{\theta}_{12} = \frac{\sum\sum_{i \neq j} \psi(x_i, y_j)}{\sum_i m_i \sum_i n_i}. \tag{5}$$

Both the population and individual AUC, like the usual AUC, are bounded between 0 and 1, $\frac{1}{2}$ represents poor predictiveness, and distance from $\frac{1}{2}$ represents increasing predictiveness. However, they describe distinct measures of informativity. It is possible for one to be informative and therefore far from 1/2, while the other is non-informative, or close to 1/2. Whereas the individual AUC is the AUC of a typical cluster, the population AUC is, setting aside ties in the data, the probability that a typical control observation in the population is less than a typical case observation. The following proposition makes this description precise. The consistency of $\hat{\theta}_{12}$ follows from Corollary 8.

Proposition 1. *1. Let $(X_1, Y_1, M_1, N_1), \ldots, (X_I, Y_I, M_I, N_I)$, be a random sample of size I IID according to P. Let P_I be the joint distribution of two independent random selections from among the elements of $X_1, \ldots, X_I$, and $Y_1, \ldots, Y_I$, and let $(\xi_I, \eta_I) \sim P_I$. Then $\theta(P_I) = Pr(\xi_I < \eta_I) + \frac{1}{2}Pr(\xi_I = \eta_I) \to \theta_{12}(P)$ as $I \to \infty$.*

2. Let $(X_1, Y_1, M_1, N_1), \ldots,$ be an infinite random sequence sampled IID according to P. Let P_∞ be the joint distribution of two independent random selections from among the elements of $X_1, \ldots,$ and $Y_1, \ldots,$ and let $(\xi_\infty, \eta_\infty) \sim P_\infty$. Then $\theta(P_\infty) = Pr(\xi_\infty < \eta_\infty) + \frac{1}{2}Pr(\xi_\infty = \eta_\infty) = \theta_{12}(P)$.

[[proofs, lemmas etc to appendix]]

The definition of the population AUC (4) allows for dependence between (M, N) and (X, Y) in capturing a population-level AUC in the sense of Proposition 1. Practical reasons to avoid assuming $(X, Y) \perp\!\!\!\perp (M, N)$ include informative censoring, imbalanced designs, [[add more reasons.]] As an alternative definition of the population AUC, consider

$$\theta'_{12} = E\left(\frac{\psi(X_1, Y_2)}{M_1 N_2}\right). \tag{6}$$

This parameter is formally a closer counterpart to the individual AUC (3), but does not take into account different cluster sizes, with a small cluster contributing as much as a large cluster to this measure of the population AUC. Therefore, this estimator would not represent the predictiveness of a typical pair of control and case observations, except in case $(X, Y) \perp\!\!\!\perp (M, N)$.

Similar to the population AUC estimator (5), [6] presents the estimator

$$\hat{\theta}'_{12} = \frac{\sum\sum_{i,j} \psi(x_i, y_j)}{\sum_i m_i \sum_i n_i} = \hat{\theta}_{12} + \frac{\sum_i \psi(x_i, y_i)}{\sum_i m_i \sum_i n_i}. \tag{7}$$

This estimator differs from ours only in including the diagonal terms, an asymptotically negligible $O(1/I)$ bias. The definition (4) was chosen in part as the probability limit of (7). Though [6] does not enunciate a clear statistical model, the analysis of (7) rather than the simpler (6) perhaps suggests that [6] contemplates $(X, Y) \not\perp\!\!\!\perp (M, N)$.

The population AUC, which appears more prominently in past research, may lay a claim to being the more natural generalization of the usual AUC since it equals the usual AUC when $M = N = 1$. Below we argue that in general the population and individual AUCs are both important, complementary tools in evaluating an estimator. In the other direction, we give inequalities that may be used in some situations to relate the two cluster AUCs.

2.2 Examples

We illustrate the population and individual AUCs and their differences using a generic random effects model with a location shift parameter. We show that the location shift can be used to control the individual AUC while the random effect can be used separately to control the population AUC.

Let the distribution of (X, Y, M, N) given M, N be

$$\begin{aligned} X \mid M, N &\sim Z(M, N) + \xi^x_i, i = 1, \ldots, M \\ Y \mid M, N &\sim Z(M, N) + \xi^y_j + \Delta, j = 1, \ldots, N \end{aligned} \tag{8}$$

Here, $\Delta > 0$ is a non-random location shift between the control and case values, Z is a random, cluster-level effect, and $\xi^x_i, \xi^y_j, i = 1, \ldots, M, j = 1, \ldots, N$, are IID individual effects. The within-cluster dependence is induced by Z. The individual effects ξ^x_i, ξ^y_j are assumed to be independent of (M, N), but Z is not assumed to be so. To keep things simple, we assume continuous densities are available, and so $\psi(x, y) = \{x < y\}$.

The individual AUC is

$$\begin{aligned}\theta_{11} &= E\frac{\psi_{11}}{M_1N_1} = E\left(\frac{1}{M_1N_1}\sum_{i=1}^{M_1}\sum_{j=1}^{N_1}\{X_{1i} < Y_{1j}\}\right)\\ &= E\left(\frac{1}{M_1N_1}\sum_{i=1}^{M_1}\sum_{j=1}^{N_1}\{Z_1+\xi_i^x < Z_1+\xi_j^y+\Delta\}\right)\\ &= E\left(\frac{1}{M_1N_1}\sum_{i=1}^{M_1}\sum_{j=1}^{N_1}P(\xi_i^x-\xi_j^y<\Delta \mid M_1,N_1)\right)\\ &= P(\xi_1-\xi_2<\Delta). \end{aligned} \tag{9}$$

Lemma 9 was used to pull the conditional expectation inside the double sum.

The population AUC is

$$\begin{aligned}\theta_{12} &= \frac{1}{E(M)E(N)}E\left(\sum_{i=1}^{M_1}\sum_{j=1}^{N_2}\{X_{1i} < Y_{2j}\}\right)\\ &= \frac{1}{E(M)E(N)}E\left(\sum_{i=1}^{M_1}\sum_{j=1}^{N_2}P(Z_1+\xi^x < Z_2+\xi^y+\Delta \mid M_1,N_1,M_2,N_2)\right)\\ &= \frac{1}{E(M)E(N)}E\left(M_1N_2P(Z_1+\xi^x < Z_2+\xi^y+\Delta \mid M_1,N_1,M_2,N_2)\right)\\ &= E\left(\frac{M_1N_2}{E(M)E(N)}\{Z_1-Z_2+(\xi^x-\xi^y)<\Delta\}\right)\end{aligned} \tag{10}$$

The last expression is a covariance-like term lying between 0 and 1.

2.2.1 Informative individual AUC, uninformative population AUC

From (9), $\theta_{11} \to 1$ as $\Delta \to \infty$.

Letting Z be independent of (M, N), $\theta_{12} = P(Z_1 + \xi^x - (Z_2 + \xi^y) < \Delta)$. As a difference of two IID random variables, $Z_1 + \xi^x - (Z_2 + \xi^y)$ is symmetric about 0, and $\theta_{12} = P(Z_1 + \xi^x - (Z_2 + \xi^y) < \Delta) = 1 - P(Z_1 + \xi^x - (Z_2 + \xi^y) \geq \Delta) = 1 - 1/2P(|Z_1 + \xi^x - (Z_2 + \xi^y)| \geq \Delta)$. For continuous densities as are being considered here, $P(|Z_1 + \xi^x - (Z_2 + \xi^y)| \geq \Delta) \to 1$, and therefore $\theta_{12} \to 1/2$, when $|f_{Z+\xi}|_\infty \to 0$, in turn implied by $|f_Z|_\infty \to 0$. For example, suppose Z belongs to a scale family, $f_Z = f_{Z_0}(Z/\sqrt{\mathrm{Var}(Z)})/\sqrt{\mathrm{Var}(Z)}$ for a fixed density $f_{Z_0}, |f_{Z_0}|_\infty < \infty$, and $\mathrm{Var}(Z) \to \infty$.

Therefore, for $\Delta = E(Y_{11}) - E(X_{11})$ large enough, θ_{11} is arbitrarily close to 1, while for any fixed Δ, for $\mathrm{Var}(Z)$ large enough, θ_{12} is arbitrarily close to $1/2$.

2.2.2 Informative population AUC, uninformative individual AUC

From (9), $\theta_{11} \to 1/2$ as $\Delta \to 0$, $\xi^x - \xi^y$ being symmetric about 0.

Let Δ be fixed. The covariance-like term (10) approaches 1 when there is a large negative covariance between M, N and $Z_1 - Z_2$, i.e., a large negative covariance bewteen M and Z

Figure 1: Two visualizations contrasting the individual and population AUCs. Each gives rug plots of ten clusters of data, each cluster sampled IID according to a binormal model, with the unclustered data at the bottom. Case observations are represented with "−" and control observations with "|". On the left, the individual AUC is informative and the population AUC uninformative. The reverse situation is presented on the right.

or large positive covariance between N and Z, or both.

Figure 1 presents a simulation using gaussian data to demonstrate the discussed differences between the population and individual AUCs. The normal model is an example of the random effects model (8) and is discussed further in Section 3. Though a large location shift can push the individual AUC close to 1, large inter-cluster variance relative to intra-cluster variance keeps the population AUC uninformative. Similarly, if the number of case observations relative to control is positively associated with the observation values, the population AUC may approach 1 irrespective of the individual AUC. [[maybe concrete examples?]]

2.3 Simplifications when $(X, Y) \perp\!\!\!\perp (M, N)$

Under some conditions, the cluster AUC parameters θ_{12} and θ_{11} may simplify to the $M = N = 1$ case. An example is given in Section 2.2, where the exchangeable cluster structure enables the simplification.

Proposition 2. *Given $(X, Y, M, N) \sim P$, suppose (X, Y) is independent of (M, N) and that $E\psi(X_{1k}, Y_{1l})$ does not depend on k, l. Then $\theta_{11}(P) = E\psi(X_{11}, Y_{11})$ and $\theta_{12}(P) = E\psi(X_{11}, Y_{21})$.*

In order for $\hat{\theta}_{12} \to 1$ while $\hat{\theta}_{11} \not\to 1$ in the random effects model discussed in Section 2.2, it was necessary that $(X, Y) \not\perp\!\!\!\perp (M, N)$. Theorem 3 bounds θ_{12} by θ_{11} under one case of

$(X, Y) \perp\!\!\!\perp (M, N)$, namely, when M and N are each constant.

We introduce the bound in a simple case. Each cluster contributes just one control and one case observation each, and their joint distribution P is supported on finitely many points in the plane:

$$\begin{aligned} P &= \sum_{i=1}^{B} p_i \delta_{(x_i, y_i)} \\ &(x_i, y_i) \in \mathbb{R}^2 \text{ and } 0 \le p_i \le 1, i = 1, \ldots, B \\ &p_1 + \ldots + p_B = 1. \end{aligned}$$

For this simple example, assume further that all the x_i and y_i are distinct, so $\psi(x, y) = \{x < y\}$.

The individual AUC is

$$\theta_{11}(P) = P(X < Y) = \sum_{i: x_i < y_i} p_i.$$

The population AUC depends on the marginal distributions of X and Y, say, $P_{\perp\!\!\!\perp}$,

$$\theta_{12}(P) = P_{\perp\!\!\!\perp}(X < Y).$$

Since all the x-coordinates of the support points are distinct, the marginal distribution of X is simply $P_{\perp\!\!\!\perp}(X = x) = \sum_i p_i \delta_{x_i}(x)$. Similarly, $P_{\perp\!\!\!\perp}(Y = y) = \sum_i p_i \delta_{y_i}(y)$. The product measure is therefore a sum of B^2 atoms, $P_{\perp\!\!\!\perp}(X = x, Y = y) = \sum_{i,j} p_i p_j \delta_{(x_i, y_j)}(x, y)$. We give a lower bound for the population AUC $P_{\perp\!\!\!\perp}(X < Y)$. An atom of P lying in $\{x < y\}$ of mass p contributes p_i^2 to the mass of $P_{\perp\!\!\!\perp}(X < Y)$. Each pair of atoms of P lying in $\{x < y\}$ of mass p and q contributes at least pq and possibly $2pq$ to the mass of $P_{\perp\!\!\!\perp}(X < Y)$. See Figure 2. Therefore

$$\begin{aligned} \theta_{12}(P) = P_{\perp\!\!\!\perp}(X < Y) &\ge \sum_{i: x_i < y_i} p_i^2 + \sum_{i: x_i < y_i} \sum_{\substack{j: x_j < y_j \\ i < j}} p_i p_j \\ &= \frac{1}{2} \left(\sum_{i: x_i < y_i} p_i \right)^2 + \frac{1}{2} \sum_{i: x_i < y_i} p_i^2 \\ &\ge \frac{1}{2} \left(\sum_{i: x_i < y_i} p_i \right)^2 + \frac{1}{2|\{i : x_i < y_i\}|} \left(\sum_{i: x_i < y_i} p_i \right)^2 \\ &= \frac{1}{2}(1 + |\{i : x_i < y_i\}|^{-1}) P(X < Y)^2 \\ &= \frac{1}{2}(1 + |\{i : x_i < y_i\}|^{-1}) \theta_{11}(P)^2. \end{aligned}$$

The first inequality is tight when each each pair i, j such that $x_i < y_i$ and $x_j < y_j$ contributes exactly $p_i p_j$, i.e., when the square given by x_i, x_j and y_i, y_j has exactly one corner in $\{x < y\}$.

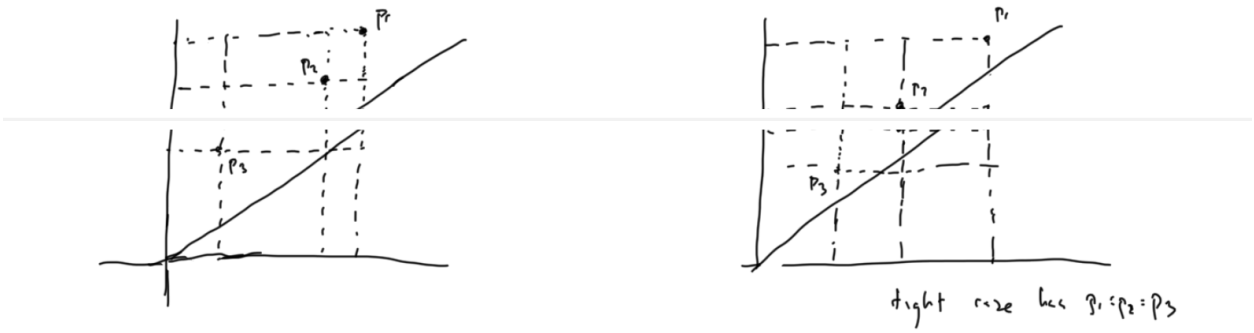

Figure 2: The case $M = N = 1$ and finitely supported (X, Y). When the distance between the atoms p_1 and p_2 is small relative to their distances to the line $x = y$, they contribute p_1p_2 to the product of the marginals. When it is relatively large, they contribute $2p_1p_2$.

See Figure 2. Assuming $x_i < x_j$, this occurs when $y_i - x_i < x_j - x_i$. The second inequality is Cauchy-Schwarz, and is tight when all the atoms in $\{x < y\}$ have the same mass.

By symmetry,

$$P_{\perp\!\!\!\perp}(X > Y) \geq \frac{1}{2}(1 + |\{i : x_i > y_i\}|^{-1})P(X > Y)^2,$$

leading to an upper bound

$$\theta_{12} \leq 1 - \frac{1}{2}(1 + |\{i : x_i > y_i\}|^{-1})(1 - \theta_{11})^2.$$

Simplifying and combining these bounds,

$$\frac{1}{2}\theta_{11}^2 \leq \theta_{12} \leq 1 - \frac{1}{2}(1 - \theta_{11})^2,$$

or equivalently,

$$1 - \sqrt{2(1 - \theta_{12})} \leq \theta_{11} \leq \sqrt{2\theta_{12}}.$$

When the individual AUC is completely uninformative, $\theta_{11} = 1/2$, the informativity of the population AUC is limited, $1/8 \leq \theta_{12} \leq 7/8$. However, when the population AUC is completely uninformative, $\theta_{12} = 1/2$, the above bounds on the individual AUC, which are tight, are vacuous, $0 \leq \theta_{11} \leq 1$. Situations as described in Section 2.2, where the population AUC $\to 1$ while the individual AUC $\to 1/2$, appears to require some dependence between M, N and X, Y.

Theorem 3 extends the inequality to an arbitrary probability measure on the plane P so long as M and N are constant.

Theorem 3. *Let $(X, Y, M, N) \sim P$ be given as in (1). Assume further that $M = m$ and $N = n$ are constant. Then*

$$\frac{1}{2}\left(\theta_{11} + \frac{\sum_{k,l} P(X_{1k} = Y_{1l})}{2mn}\right)^2 \leq \theta_{12} \leq 1 - \frac{1}{2}\left(1 - \theta_{11} + \frac{\sum_{k,l} P(X_{1k} = Y_{1l})}{2mn}\right)^2$$

Lemma 4. *Given a pair of scalar random variables (X, Y) with joint distribution P, let $P_{\perp\!\!\!\perp}$ be the product measure of the marginals, i.e., for all real a, b,*

$$P_{\perp\!\!\!\perp}(\{x < a\} \cap \{y < b\}) = P(\{x < a\})P(\{y < b\}).$$

Then

$$\frac{1}{2}(P(X < Y) + P(X = Y))^2 \leq P_{\perp\!\!\!\perp}(X < Y) + \frac{1}{2}(X = Y) \leq 1 - \frac{1}{2}(1 - P(X < Y))^2.$$

With the random vector $(X, Y, M, N) \sim P$, with constant $M = N = 1$ and so P may be regarded as the joint distribution of (X, Y), the conclusion of the Lemma is

$$\frac{1}{2}(\theta_{11}(P) + \frac{1}{2}P(X = Y))^2 \leq \theta_{12}(P) \leq 1 - \frac{1}{2}(1 - \theta_{11}(P) + \frac{1}{2}P(X = Y))^2. \tag{11}$$

Proof of Theorem 3. With

$$\theta_{11} = \frac{1}{mn}E(\psi_{11}) = \frac{1}{mn}\sum_{i,j}(P(X_{1i} < Y_{1j}) + \frac{1}{2}P(X_{1i} = Y_{1j}))$$

Lemma 4 gives

$$\begin{aligned}
\theta_{12} = \frac{1}{mn}E(\psi_{12}) &= \frac{1}{mn}\sum_{i,j}(P(X_{1i} < Y_{2j}) + \frac{1}{2}P(X_{1i} = Y_{2j})) \\
&\geq \frac{1}{mn}\sum_{i,j}\frac{1}{2}(P(X_{1i} < Y_{1j}) + P(X_{1i} = Y_{1j}))^2 \\
&\geq \frac{1}{2}\left(\frac{1}{mn}\sum_{i,j}(P(X_{1i} < Y_{1j}) + P(X_{1i} = Y_{1j}))\right)^2 \\
&= \frac{1}{2}\left(\theta_{11} + \frac{1}{2mn}\sum_{i,j}P(X_{1i} = Y_{1j})\right)^2.
\end{aligned}$$

The second inequality is Jensen's inequality, which is tight with the pairwise AUCs are all equal.

The other bound follows similarly. □

2.4 Relation to Simpson's Paradox

[[needs revision]] Simpson's paradox, understood broadly, refers to situations where data is clustered, exhibits a consistent trend at each cluster, but exhibits a contrary trend when the unclustered data is analyzed. The situation in Section 2.2 is an example of this phenomenon. The individual and population AUCs are clustered and unclustered analyses that can yield opposite conclusions about the quality of the predictor. Contemporary analyses of Simpson's paradox show the importance of considering both the individual and population AUCs.

The original Simpson's paradox concerns contingency table data. The tables in [[ref]] are drawn from [4]. One finds that $E(f(Y) \mid T, M) > E(f(Y) \mid \overline{T}, M)$ for all levels of M and still

$$\int E(f(Y) \mid T, m) f_{M|T}(m) = E(f(Y) \mid T) < E(f(Y) \mid \overline{T}) = \int E(f(Y) \mid \overline{T}, m) f_{M|\overline{T}}(m).$$

Mathematically the situation should is not surprising, for as long as $\min_m E(f(Y) \mid T, m) < \max_m E(f(Y) \mid \overline{T}, m)$ one can find $f_{M|T}, f_{M|\overline{T}}$ bringing about the reversal. The paradoxical nature comes when relating the statistics to an application. Under one interpretation, [[ref table]] presents the recovery rates for a treatment broken down by gender. The recovery rates suggest that the treatment is harmful for the male and female subpopulations, but beneficial for the population as a whole. It would seem that a doctor faced with any given patient, male or female, ought to advise the patient to avoid the treatment, despite its apparently beneficial effects for the population as a whole. Under a second interpretation, [[ref table]] presents the yields for two strains of flowers, white or black, broken down by two levels of height, short or tall. In this situation, [4] suggest, one might prefer the "treatment" that performs best at the population level, namely, white flowers.

Working in the framework of causal inference, [7] argues that the confusion arises from the complicated relationship between causal intervention and statistical conditioning. The correct analysis in any given situation, whether the clustered or unclustered analysis, requires information about the underlying causal relationships between the treatment, outcome, and clustering variable. The inclination to favor the clustered treatment under the first interpretation is that gender is a pretreatment confounder, and it ought to be controlled for. In the second, by contrast, the clustering variable, height, might lie on the causal pathway between the "treatment" of flower species and the outcome of yield, e.g., the white flower species may be both taller and in turn higher yielding. A causal analysis would be obscured by conditioning on height.

The analysis applies as well to the individual and population AUCs. [[examples. could use the same as lindley-novick? Ie, drug treatment (say dosage level) as a predictor of recovery, with gender a pre-treatment confounder]] The main point is that both needed in different situations. In general, if underlying causal structure is unknown, both may be studied.

2.5 Asymptotic Distribution of $(\theta_{12}, \theta_{11})$

Theorem 5 gives the asymptotic joint distribution of the individual and population AUCs. It is stated in somewhat greater generality for any square-integrable kernel, not just the AUC kernel (2). The proof actually works for any random variables M, N, such that $EM \neq 0, EN \neq 0$, $EM^{-2} < \infty, EN^{-2} < \infty$, i.e., M and N need not have the interpretation as the lengths of X, Y.

Theorem 5. *Let $\psi : V \times V \to \mathbb{R}$, $(X, Y, M, N) \sim P$ with $(X, Y) \in V \times V$, $\psi \in L^2(P)$, M and N counting numbers ≥ 0 with finite means. Then*

$$\sqrt{I}(\hat{\theta}_{12} - \theta_{12}, \hat{\theta}_{11} - \theta_{11}) \rightsquigarrow \mathcal{N}(0, \Sigma)$$

with

$$
\begin{aligned}
a\Sigma_{11} &= \lim_{I\to\infty} I \operatorname{Var}(\hat{\theta}_{12}) = E\left(\frac{E(\psi_{12} \mid W_1) + E(\psi_{21} \mid W_1)}{EMEN} - \theta_{12}\left(\frac{M_1}{EM} + \frac{N_1}{EN}\right)\right)^2 \\
\Sigma_{22} &= \lim_{I\to\infty} I \operatorname{Var}(\hat{\theta}_{11}) = \operatorname{Var}(\psi_{11}/(M_1N_1)) \\
\Sigma_{12} &= \lim_{I\to\infty} I \operatorname{Cov}(\hat{\theta}_{12}, \hat{\theta}_{11}) = \theta_{12} E\left(\frac{\psi_{11}}{M_1N_1}\left(\frac{\psi_{12}+\psi_{21}}{E\psi_{12}} - \frac{M_1}{EM} - \frac{N_1}{EN}\right)\right)
\end{aligned}
$$

Corollary 6. *Under the assumptions of Theorem 5, let* $(X_1, Y_1, M_1, N_1), \ldots, (X_I, Y_I, M_I, N_I)$, *be IID according to* P. *For* $1 \le i \le I$ *define*

$$
\begin{aligned}
\psi_{i\cdot} &= \sum_{j=1}^{I} \psi(X_i, Y_j) \\
\psi_{\cdot i} &= \sum_{j=1}^{I} \psi(X_j, Y_i) \\
\phi_i &= \frac{\psi(X_i, Y_i)}{M_i N_i}.
\end{aligned}
$$

The asymptotic covariance matrix Σ *of* $(\hat{\theta}_{12}, \hat{\theta}_{11})$ *may be consistently estimated by* $\hat{\Sigma}$ *given by*

$$
\begin{aligned}
\hat{\Sigma}_{11} &= \frac{1}{I-1}\sum_{i=1}^{I}\left(\frac{\psi_{i\cdot}+\psi_{\cdot i}}{M_{\cdot}N_{\cdot}} - \hat{\theta}_{12}\left(\frac{M_i}{M_{\cdot}} + \frac{N_i}{N_{\cdot}}\right)\right)^2 \\
\hat{\Sigma}_{22} &= \frac{1}{I-1}\sum_{i=1}^{I}(\phi_i - \phi_{\cdot})^2 \\
\hat{\Sigma}_{12} &= \frac{1}{I}\sum_{i=1}^{I}\left(\frac{\phi_i}{\phi_{\cdot}}\left(\frac{\psi_{i\cdot}+\psi_{\cdot i}}{\psi_{\cdot\cdot}} - \frac{M_i}{M_{\cdot}} - \frac{N_i}{N_{\cdot}}\right)\right)
\end{aligned}
$$

The estimator $\hat{\Sigma}_{11}$ of the asymptotic variance of $\hat{\theta}_{12}$ is the same as given by [6], derived by a different method.

The finite-sample performance of this estimator is examined in Section 3.

3 Simulation

[[give purpose of simulation section]]

A popular parametric model for the AUC is the binormal model, so called since the case and control observations are assumed to follow a normal distribution [1] . Following [6] we extend this model to accommodate clustered data by modeling the observations as multivariate normal vectors with an exchangeable correlation structure.

$$
(X, Y) \mid (M, N) \sim \mathcal{N}_{M+N}\left(\begin{pmatrix}\mu_X \mathbb{1}_M \\ \mu_Y \mathbb{1}_N\end{pmatrix}, \mathbb{1}_{M+N}\mathbb{1}_{M+N}^T + (1-\rho) Id_{M+N}\right) \tag{12}
$$

That is, the case and control observations of a given cluster all have unit variance and share the same pairwise correlation ρ, $-1/(M+N) \leq \rho \leq 1$, all the case observations have mean μ_Y, and all the control observaions mean μ_X. The binormal model is an example of the random effect model described in Section 2.2. As the effect of the random effect is only to change the intra-cluster correlation or mean, it is actually redundant to the usual multivariate normal parameters and omitted from (12). Moreover, further parameters such as for intra-cluster correlations $\mathrm{Corr}(X_{11}, X_{12})$, $\mathrm{Corr}(Y_{11}, Y_{12})$, or for non-unit variances $\mathrm{Var}(X_{11})$ and $\mathrm{Var}(Y_{11})$, are redundant for our purpose of modeling AUCs.

Let $\Delta = \mu_Y - \mu_X \geq 0$. Using Proposition 2,

$$\begin{aligned} \theta_{12}(P) &= \Phi\left(\frac{\Delta}{\sqrt{2}}\right) \\ \theta_{11}(P) &= \Phi\left(\frac{\Delta}{\sqrt{2(1-\rho)}}\right) \end{aligned} \tag{13}$$

The formulas (13) show that θ_{12} and θ_{11} are simultaneously $> 1/2, 1/2$, or $< 1/2$. We give two benefits. The first is that $(\theta_{12}, \theta_{11})$ can be restricted without loss of generality to $[1/2, 1] \times [1/2, 1]$, and may then serve as a parameterization of the binormal model (12). The second involves testing. Though AUCs are often compared by magnitude [[cite multi-reader papers]], e.g., $H_0 : AUC_1 - AUC_2 > 0$, one is usually interested in the informativity, i.e., $|AUC_1 - 1/2|$ versus $|AUC_2 - 1/2|$. The hypothesis $H_0 : AUC_1 - AUC_2 > 0$ indicates that AUC_1 is more informative when both are greater than $1/2$, but less informative if both are less than $1/2$. A further complication is that one may be greater than $1/2$ and the other less, which will not be solved by switching the class designations. These complications are avoided in the binormal model for the individual and population AUCs. A test of the order $\theta_{12} < \theta_{11}$ is also a test of informativity $|\theta_{12} - 1/2| < |\theta_{11} - 1/2|$.

To generate (M, N), first the number $M + N$ of combined case and control observations belonging in a sample is randomly selected from among $k \in \{[[...]]\}$. Next, to obtain the allocation to case and control observations, $M + N$ normal variables are first sampled with unit variance and common pairwise correlation $\rho_{MN} \in \{[[...]]\}$. The number M of control observations is taken to be those greater than 0, and the remainder the number N of case observations. The greater the correlation, the greater the imbalance between case and control observations within the clusters.

3.1 Coverage

The parameters Δ and ρ were set to correspond to population AUC of $\theta_{12} \in \{[[...]]\}$ and individual AUCs of $\theta_{11} \in \{[[...]]\}$. For each setting of $\rho_{MN}, \theta_{12}, \theta_{11}$, $[[...]]$ replicates of size $I = [[...]]$ were sampled and used to form a confidence ellipse for $(\theta_{12}, \theta_{11})$. Specifically, with $\hat{\theta}_{12}, \hat{\theta}_{11}$ computed as in Section 2.1 and Σ as in Theorem 5, under P,

$$\left| \Sigma^{-1/2} \left(\begin{pmatrix} \theta_{12} \\ \theta_{11} \end{pmatrix} - \begin{pmatrix} \hat{\theta}_{12} \\ \hat{\theta}_{11} \end{pmatrix} \right) \right|^2 \tag{14}$$

	theta.12	theta.11	D.corr	coverage	bias.theta.11	bias.theta.12	vcov.11	vcov.12	vcov.22
1	0.60	0.60	0.00	0.90	0.01	0.00	-0.04	-0.03	-0.02
2	0.60	0.60	0.40	0.97	0.00	0.00	-0.01	0.00	0.00
3	0.60	0.60	0.80	0.93	-0.00	-0.00	0.03	0.01	0.01
4	0.60	0.60	0.95	0.90	-0.01	0.00	-0.05	-0.02	-0.01
1.1	0.60	0.77	0.00	0.97	-0.00	-0.00	0.01	-0.00	0.00
2.1	0.60	0.77	0.40	1.00	0.00	0.00	0.00	0.00	0.03
3.1	0.60	0.77	0.80	0.97	0.00	-0.00	0.01	-0.00	0.05
4.1	0.60	0.77	0.95	0.90	0.01	0.01	-0.01	-0.02	-0.02
1.2	0.60	0.95	0.00	1.00	0.00	-0.00	0.00	0.00	-0.00
2.2	0.60	0.95	0.40	0.87	0.00	0.00	-0.00	-0.00	-0.03
3.2	0.60	0.95	0.80	1.00	0.00	0.00	0.00	0.00	-0.01
4.2	0.60	0.95	0.95	0.97	-0.00	0.01	0.00	-0.01	-0.01
1.3	0.80	0.80	0.00	0.90	0.00	0.00	0.00	0.00	-0.00
2.3	0.80	0.80	0.40	0.93	-0.00	0.00	0.01	-0.00	-0.00
3.3	0.80	0.80	0.80	0.90	0.00	0.00	-0.01	-0.00	0.00
4.3	0.80	0.80	0.95	0.93	-0.00	0.00	-0.02	0.00	0.00
1.4	0.80	0.88	0.00	0.97	-0.00	0.00	0.00	0.00	0.01
2.4	0.80	0.88	0.40	0.93	-0.00	0.01	0.00	0.00	0.00
3.4	0.80	0.88	0.80	0.87	0.01	0.01	0.01	0.01	-0.01
4.4	0.80	0.88	0.95	0.90	0.01	0.01	0.01	0.01	0.00
1.5	0.80	0.95	0.00	0.93	0.00	0.00	0.00	0.00	0.01
2.5	0.80	0.95	0.40	1.00	-0.00	-0.00	0.00	-0.00	0.02
3.5	0.80	0.95	0.80	0.90	-0.00	-0.00	0.00	0.01	0.00
4.5	0.80	0.95	0.95	1.00	-0.00	-0.00	0.00	-0.00	0.01

Table 1: The results of a simulation examining the coverage of a nominal 95% confidence ellipse obtained using the asymptotic estimator given in Section 2.5.[[preliminary–only 30 monte carlo reps. Need to pretty print table]].

has a chi-squared distribution with 2 degrees of freedom. If q is an upper α quantile of this distribution, then

$$\left\{ \begin{pmatrix} x \\ y \end{pmatrix} : \left| \Sigma^{-1/2} \left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \hat{\theta}_{12} \\ \hat{\theta}_{11} \end{pmatrix} \right) \right|^2 < q \right\}$$

is a level $1-\alpha$ confidence region for $(\theta_{12}, \theta_{11})$, which then covers $(\theta_{12}, \theta_{11})$ when (14) is $< q$. In the simulation, we substitute for Σ the asymptotic approximation $\hat{\Sigma}$ given in Corollary 6. This process was repeated [[...]] times. Results are presented in Table 1.

[[TODO: 1. similar simulation performed with truncated normal or other nonnormal distr. 2. check M,N imbalance affects precision as in obuchowski.]]

3.2 Power

We examine the power of testing the null hypothesis $H_0 : \theta_{12} = \theta_{11}$ using the proposed variance estimators under the binormal model (12). Restricting to $\rho > 0$ in (12), the alternatives

to $H_0 : \theta_{12} = \theta_{11}$ is the set $H_A : |\theta_{12} - 1/2| < |\theta_{11} - 1/2|$, i.e., where the individual AUC is more informative than the population AUC.

The data is generated under (12) using $(\theta_{12}, \theta_{11})$ selected from [[a grid or randomly sampled?]] $[\frac{1}{2}, 1] \times [\frac{1}{2}, 1]$. Estimates $\hat{\theta}_{12}, \hat{\theta}_{11}, \hat{\Sigma}$ were then obtained as described above. The test is carried out by testing the signficance of the z-statistic

$$(\hat{\theta}_{12} - \hat{\theta}_{11})/\sqrt{c^t \hat{\Sigma} c}$$

with the contrast c is $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$.

The results of the simulation are presented visually in Fig.3. [[give some summary of results maybe.]]

4 Data analysis

5 Discussion/Conclusion

Directions for future work:

1. multiple correlated AUCs–easy extension of the above results
2. longitudinal analysis
3. covariate-adjusted AUC

References

[1] Donald D Dorfman and Edward Alf Jr. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating-method data. *Journal of mathematical psychology*, 6(3):487–496, 1969.

[2] Birol Emir, Sam Wieand, Sin-Ho Jung, and Zhiliang Ying. Comparison of diagnostic markers with repeated measurements: a non-parametric roc curve approach. *Statistics in Medicine*, 19(4):511–523, 2000.

[3] A J Lee. *U-statistics: Theory and Practice*. Routledge, 2019.

[4] Dennis V Lindley and Melvin R Novick. The role of exchangeability in inference. *The annals of statistics*, pages 45–58, 1981.

[5] Haben Michael, Lu Tian, and Musie Ghebremichael. The roc curve for regularly measured longitudinal biomarkers. *Biostatistics*, 20(3):433–451, 2019.

[6] Nancy A Obuchowski. Nonparametric analysis of clustered roc curve data. *Biometrics*, pages 567–578, 1997.

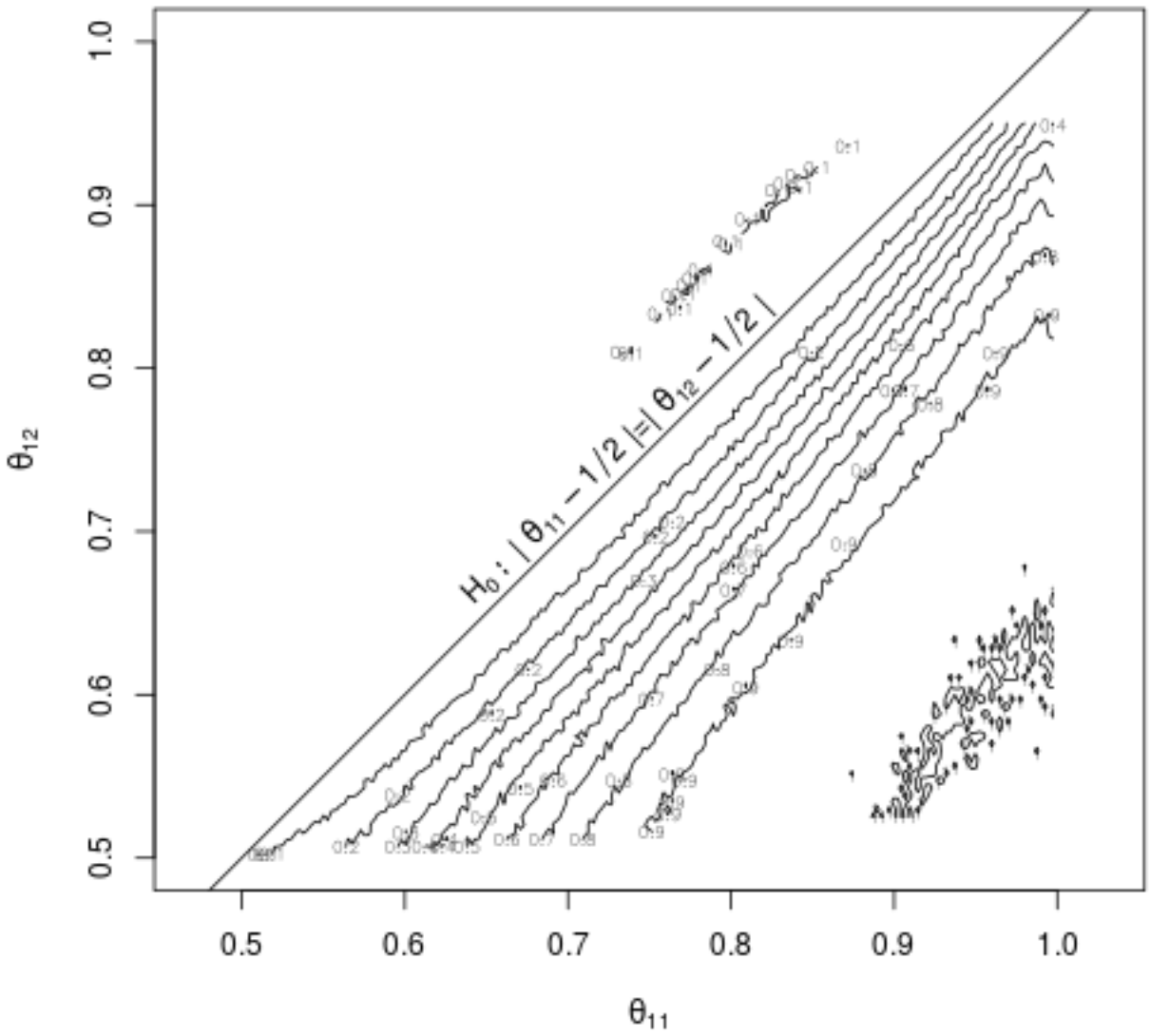

Figure 3: Power of the test of $H_0 : \theta_{12} = \theta_{11}$ using the asymptotic estimator given in Section 2.5. [[put null in usual format,get rid of artifacts, more monte carlo interations to smooth out]]

[7] Judea Pearl. Comment: understanding simpson's paradox. *The American Statistician*, 68(1):8–13, 2014.

[8] Yougui Wu and Xiaofei Wang. Optimal weight in estimating and comparing areas under the receiver operating characteristic curve using longitudinal data. *Biometrical journal*, 53(5):764–778, 2011.

Proof of Proposition 1. 1. By the LLN $1 \geq I^2/MN \to 1/(EMEN)$ almost surely and by Lemma 7 $\sum_{i,j} \psi_{ij}/I^2 \to E\psi_{12}$ almost surely. Conditioning on the sample,

$$\begin{aligned} E\psi(\xi_I, \eta_I) &= E(E(\psi(\xi_I, \eta_I) \mid (X_1, Y_1, M_1, N_1), \ldots, (X_I, Y_I, M_I, N_I)) \\ &= E\left(\frac{\sum_{1\leq i,j\leq I} \sum_{1\leq k\leq M_i, 1\leq l\leq N_j} \psi(X_{ik}, Y_{jl})}{\sum_{i=1}^I M_i \sum_{i=1}^I N_i} \right) \\ &= E\left(\frac{\sum_{1\leq i,j\leq I} \psi_{ij}}{\sum_{i=1}^I M_i \sum_{i=1}^I N_i} \right) \to \frac{E\psi_{12}}{EMEN} = \theta_{12}. \end{aligned}$$

The limit is justified by dominated convergence, given the boundedness of I^2/MN and moment condition on ψ.

2. The second part follows on showing that $(\xi_I, \eta_I) \to (\xi_\infty, \eta_\infty)$ in distribution. For $a, b \in \mathbb{R}$, by a similar argument as above,

$$\begin{aligned} P(\xi_I < a, \eta_I < b) &= E\left(\frac{\sum_{1\leq i,j\leq I} \sum_{1\leq k\leq M_i, 1\leq l\leq N_j} \{X_{ik} < a, Y_{jl} < b\}}{\sum_{i=1}^I M_i \sum_{i=1}^I N_i} \right) \\ &\to \frac{E\left(\sum_{k=1}^{M_1} \{X_{1k} < a\}\right)}{EM} \frac{E\left(\sum_{l=1}^{N_1} \{Y_{1l} < b\}\right)}{EN}. \end{aligned}$$

The probability of sampling an element from a cluster of size $M = m$ given an initial segment of I samples $(X_1, Y_1, M_1, N_1), \ldots, (X_I, Y_I, M_I, N_I)$, is $\frac{\sum_{i=1}^I \{M_i = m\}}{\sum_{i=1}^I M_i}$. Along almost any sequence of samples as $I \to \infty$ this relative frequency tends to $\frac{mP(M=m)}{EM}$[[, the probability of sampling an element from a clusters of size m from the sequence]]. Therefore

$$\begin{aligned} P(\xi_\infty < a) &= \sum_{m=1}^\infty P(\xi_\infty < a \mid \xi_\infty \text{ is sampled from a cluster of size } m) \cdot \\ &\qquad P(\xi_\infty \text{ is sampled from a cluster of size } m) \\ &= \sum_{m=1}^\infty \frac{1}{m} \sum_{k=1}^m P(X_{1k} < a \mid M = m) \frac{mP(M = m)}{EM} \\ &= \frac{1}{EM} \sum_{m=1}^\infty \sum_{k=1}^m P(X_{1k} < a \mid M = m) P(M = m) \\ &= \frac{1}{EM} E\left(\sum_{k=1}^M \{X_{1k} < a\} \right). \end{aligned}$$

Analogously,

$$P(\eta_\infty < a) = \frac{1}{EN} E\left(\sum_{l=1}^{N} \{X_{1l} < a\}\right).$$

The product is the distributional limit for $P(\xi_I < a, \eta_I < b)$ given above.

□

The following lemma gives convergence results for a two-sample U-statistic with kernel of degree $(1,1)$ where the data is paired. The relevant definitions and the result for independent samples is given in, e.g., [3]. [[Need to define V, the space where X, Y lie. These are vectors of variable length so V should be at least as big as c_{00}]]

Lemma 7. *Given a sample $(X_1, Y_1), \ldots, (X_I, Y_I)$ on $V \times V$ IID according to P and a function $\psi : V \times V \to \mathbb{R}$ in $L^2(P)$, define*

$$U_I = I^{-2} \sum_{1 \le i,j \le I} \psi(X_i, Y_j)$$

and

$$\hat{U}_I = I^{-1} \sum_{i=1}^{I} \left(E(\psi(X_i, Y_0) \mid X_i, Y_i) + E(\psi(X_0, Y_i) \mid X_i, Y_i)\right) - 2E\psi(X_1, Y_2).$$

Then

$$E(U_I - EU - \hat{U}_I)^2 = O(I^{-2}).$$

Proof of Lemma 7. a[[fill in]] □

Corollary 8. *With the same setup as Lemma 7, $\hat{U}_I \to EU$ a.s. and $\sqrt{I}(\hat{U}_I - EU)/\sqrt{\mathrm{Var}(U_I)} \to \mathcal{N}(0,1)$ in distribution.*

Proof of Corollary 8. By Lemma 7, $U_I \to \hat{U}_I$ a.s. and $\sqrt{I}(U_I - \hat{U}_I) \to 0$ in quadratic mean, and $\hat{U}_I$ is an IID sum subject to the usual LLN and CLT. □

Proof of Proposition 2.

$$\begin{aligned}
\theta_{11}(P) &= E\left(\frac{\sum_{k=1}^{M}\sum_{l=1}^{N} \psi(X_{1k}, Y_{1l})}{MN}\right) \\
&= E\left(\frac{1}{MN} E\left(\sum_{k=1}^{M}\sum_{l=1}^{N} \psi(X_{1k}, Y_{1l}) \mid M, N\right)\right) \\
&= E\left(\frac{1}{MN} MN E\psi(X_{11}, Y_{11})\right) = E\psi(X_{11}, Y_{11}).
\end{aligned}$$

Lemma 9 was used to get the third equality.

If $E\psi(X_{1k}, Y_{1l})$ does not depend on k, l, then neither does $E\psi(X_{1k}, Y_{2l})$. Similar to the above,

$$\begin{aligned}\theta_{12}(P) &= \frac{E\left(\sum_{k=1}^{M_1}\sum_{l=1}^{N_2}\psi(X_{1k}, Y_{2l})\right)}{E(M)E(N)} \\ &= \frac{E(M)E(N)E\psi(X_{11}, Y_{21})}{E(M)E(N)} = E\psi(X_{11}, Y_{21}).\end{aligned}$$

□

Lemma 9. *Given random variables $M, V, X_1, X_2, \ldots,$ such that $M \in \{1, 2, \ldots\}, E|M| < \infty$... [[other moment conditions]]*

$$E\left(\sum_{i=1}^{M} X_i \middle| M, V\right) = \sum_{i=1}^{M} E(X_i \mid M, V)$$

Proof of Lemma 9.

$$\begin{aligned}E\left(\sum_{i=1}^{M} X_i \middle| M, V\right) &= E\left(\sum_{m=1}^{\infty}\{M = m\}\sum_{i=1}^{m} X_i \middle| M, V\right) \\ &= \sum_{m=1}^{\infty} E\left(\{M = m\}\sum_{i=1}^{m} X_i \middle| M, V\right) \\ &= \sum_{m=1}^{\infty}\sum_{i=1}^{m}\{M = m\}E(X_i \mid M, V) \\ &= \sum_{i=1}^{M} E(X_i \mid M, V).\end{aligned}$$

[[justify interchange in 2nd equality]] □

Proof of Lemma 4. Define for $n \in \mathbb{N}$ approximations to θ_{11} and θ_{12} by

$$\begin{aligned}A_{ij}^{(n)} &= \left\{(x, y) : \frac{i}{2^n} \le x < \frac{i+1}{2^n}, \frac{j}{2^n} \le y < \frac{j+1}{2^n}\right\}, \quad -2^{2n} \le i, j < 2^{2n} - 1 \\ \theta_{11}^{(n)} &= \sum_{i=-2^{2n}}^{2^{2n}-1}\sum_{j=i+1}^{2^{2n}-1} P(A_{ij}^{(n)}) + \frac{1}{2}\sum_{i=-2^{2n}}^{2^{2n}-1} P(A_{ii}^{(n)}) \\ \theta_{12}^{(n)} &= \sum_{i=-2^{2n}}^{2^{2n}-1}\sum_{j=i+1}^{2^{2n}-1} P_{\perp\!\!\!\perp}(A_{ij}^{(n)}) + \frac{1}{2}\sum_{i=-2^{2n}}^{2^{2n}-1} P_{\perp\!\!\!\perp}(A_{ii}^{(n)}).\end{aligned}$$

Since $\bigcup_n \bigcup_i \bigcup_{j>i} A_{ij}^{(n)} = \{x < y\}$ and $\bigcap_n \bigcup_i A_{ii}^{(n)} = \{x = y\}$, by continuity of measure $\theta_{11}^{(n)} \to \theta_{11}$ and $\theta_{12}^{(n)} \to \theta_{12}$. Therefore, it is enough to establish the inequality (11) for $\theta_{11}^{(n)}$ and $\theta_{12}^{(n)}$.

Fixing n,

$$\sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{j=i+1}^{2^{2n}-1} P_{\perp\!\!\!\perp}(A_{ij}^{(n)}) = \sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{j=i+1}^{2^{2n}-1} P_{\perp\!\!\!\perp}(A_{ij}^{(n)})$$

$$= \sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{j=i+1}^{2^{2n}-1} P_{\perp\!\!\!\perp}(\frac{i}{2^n} \le x < \frac{i+1}{2^n}) P_{\perp\!\!\!\perp}(\frac{j}{2^n} \le y < \frac{j+1}{2^n})$$

$$\ge \sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{j=i+1}^{2^{2n}-1} (P(A_{ii}^{(n)}) + \sum_{k=i+1}^{2^{2n}-1} P(A_{ik}^{(n)}))(P(A_{jj}^{(n)}) + \sum_{l=-2^{2n}}^{j-1} P(A_{lj}^{(n)}))$$

$$= \sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{j=i+1}^{2^{2n}-1} \left(\sum_{k=i+1}^{2^{2n}-1} P(A_{ik}^{(n)}) \sum_{l=-2^{2n}}^{j-1} P(A_{lj}^{(n)}) + P(A_{ii}^{(n)}) \sum_{l=-2^{2n}}^{j-1} P(A_{lj}^{(n)}) + P(A_{jj}^{(n)}) \sum_{k=i+1}^{2^{2n}-1} P(A_{ik}^{(n)}) + P(A_{ii}^{(n)})P(A \right.$$

We lower bound the first three terms in parentheses.
First term:

$$\sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{j=i+1}^{2^{2n}-1}\sum_{k=i+1}^{2^{2n}-1} P(A_{ik}^{(n)}) \sum_{l=-2^{2n}}^{j-1} P(A_{lj}^{(n)})$$

$$= \sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{k=i+1}^{2^{2n}-1} P(A_{ik}^{(n)}) \sum_{j=i+1}^{2^{2n}-1}\sum_{l=-2^{2n}}^{j-1} P(A_{lj}^{(n)})$$

$$\ge \sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{k=i+1}^{2^{2n}-1} P(A_{ik}^{(n)}) \sum_{j=i+1}^{2^{2n}-1}\sum_{l=i}^{j-1} P(A_{lj}^{(n)})$$

$$= \sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{k=i+1}^{2^{2n}-1} P(A_{ik}^{(n)}) \sum_{l=i}^{2^{2n}-2}\sum_{j=l+1}^{2^{2n}-1} P(A_{lj}^{(n)})$$

$$= \sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{k=i+1}^{2^{2n}-1} P(A_{ik}^{(n)}) \sum_{j=i+1}^{2^{2n}-1} P(A_{ij}^{(n)}) + \sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{k=i+1}^{2^{2n}-1} P(A_{ik}^{(n)}) \sum_{l=i+1}^{2^{2n}-2}\sum_{j=l+1}^{2^{2n}-1} P(A_{lj}^{(n)})$$

$$\ge \sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{j=i+1}^{2^{2n}-1} P(A_{ij}^{(n)})^2 + \sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{k=i+1}^{2^{2n}-2}\sum_{j=k+1}^{2^{2n}-1} P(A_{ij}^{(n)})P(A_{ik}^{(n)}) + \sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{k=i+1}^{2^{2n}-1} P(A_{ik}^{(n)}) \sum_{l=i+1}^{2^{2n}-2}\sum_{j=l+1}^{2^{2n}-1} P(A_{lj}^{(n)})$$

$$= \sum\sum\sum\sum_{\substack{i\neq k \text{ or } j\neq l \\ j>i \text{ and } l>k}} P(A_{ij}^{(n)})P(A_{kl}^{(n)}) + \sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{j=i+1}^{2^{2n}-1} P(A_{ij}^{(n)})^2$$

$$= \frac{1}{2}\left(\sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{j=i+1}^{2^{2n}-1} P(A_{ij}^{(n)}) \right)^2 + \frac{1}{2}\sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{j=i+1}^{2^{2n}-1} P(A_{ij}^{(n)})^2.$$

Middle two terms:

$$
\begin{aligned}
&\sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{j=i+1}^{2^{2n}-1}\left(P(A_{ii}^{(n)})\sum_{l=-2^{2n}}^{j-1}P(A_{lj}^{(n)})+P(A_{jj}^{(n)})\sum_{k=i+1}^{2^{2n}-1}P(A_{ik}^{(n)})\right)\\
&=\sum_{i=-2^{2n}}^{2^{2n}-2}P(A_{ii}^{(n)})\sum_{l=i}^{2^{2n}-2}\sum_{j=l+1}^{2^{2n}-1}P(A_{lj}^{(n)})+\sum_{j=-2^{2n}+1}^{2^{2n}-1}P(A_{jj}^{(n)})\sum_{i=-2^{2n}}^{j-1}\sum_{k=i+1}^{2^{2n}-1}P(A_{ik}^{(n)})\\
&=\sum_{i=-2^{2n}}^{2^{2n}-2}P(A_{ii}^{(n)})\sum_{l=i}^{2^{2n}-2}\sum_{j=l+1}^{2^{2n}-1}P(A_{lj}^{(n)})+\sum_{i=-2^{2n}+1}^{2^{2n}-1}P(A_{ii}^{(n)})\sum_{l=-2^{2n}}^{i-1}\sum_{j=l+1}^{2^{2n}-1}P(A_{lj}^{(n)})\\
&=\left(\sum_{i=-2^{2n}}^{2^{2n}-1}P(A_{ii}^{(n)})\right)\left(\sum_{l=-2^{2n}}^{2^{2n}-2}\sum_{j=l+1}^{2^{2n}-1}P(A_{lj}^{(n)})\right).
\end{aligned}
$$

The second-to-last equality is just renaming indices.

With these lower bounds,

$$
\begin{aligned}
\theta_{12}^{(n)}&=\sum_{i=-2^{2n}}^{2^{2n}-1}\sum_{j=i+1}^{2^{2n}-1}P_{\perp\!\!\!\perp}(A_{ij}^{(n)})+\frac{1}{2}\sum_{i=-2^{2n}}^{2^{2n}-1}P_{\perp\!\!\!\perp}(A_{ii}^{(n)})\\
&\geq\frac{1}{2}\left(\sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{j=i+1}^{2^{2n}-1}P(A_{ij}^{(n)})\right)^2+\left(\sum_{i=-2^{2n}}^{2^{2n}-1}P(A_{ii}^{(n)})\right)\left(\sum_{l=-2^{2n}}^{2^{2n}-2}\sum_{j=l+1}^{2^{2n}-1}P(A_{lj}^{(n)})\right)+\\
&\sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{j=i+1}^{2^{2n}-1}P(A_{ii}^{(n)})P(A_{jj}^{(n)})+\frac{1}{2}\sum_{i=-2^{2n}}^{2^{2n}-1}P(A_{ii}^{(n)})^2\\
&=\frac{1}{2}\left(\sum_{i=-2^{2n}}^{2^{2n}-2}\sum_{j=i+1}^{2^{2n}-1}P(A_{ij}^{(n)})+\sum_{i=-2^{2n}}^{2^{2n}-1}P(A_{ii}^{(n)})\right)^2\\
&=\frac{1}{2}\left(\theta_{11}^{(n)}+\frac{1}{2}\sum_{i=-2^{2n}}^{2^{2n}-1}P(A_{ii}^{(n)})\right)^2.\\
&=\frac{1}{2}\left(\theta_{11}^{(n)}+\frac{1}{2}P(X=Y)\right)^2+o(1).
\end{aligned}
$$

The upper bound then follows by the same symmetry argument as given in Section 2.3.

□

Proof of Theorem 5. By Corollary 8

$$
\sqrt{I}\left(\frac{(I)_2^{-1}\sum_{i\neq j}\psi_{ij}-E\psi_{12}}{\mathrm{sd}(\sqrt{I}(I)_2^{-1}\sum_{i\neq j}\psi_{ij})},\frac{I^{-2}\sum_{i,j}M_iN_j-EMEN}{\mathrm{sd}(I^{-3/2}\sum_{i,j}M_iN_j)},\frac{I^{-1}\sum_i\psi_{ii}/(M_iN_i)-E(\psi_{11}/M_1N_1)}{\mathrm{sd}(\psi_{11}/M_1N_1)}\right)
$$

converges to

$$I^{-1/2}\sum_{i=1}^{I}\left(\frac{E(\psi_{i0}\mid W_i)+E(\psi_{0i}\mid W_i)-2E\psi_{12}}{\mathrm{sd}(E(\psi_{10}\mid W_1)+E(\psi_{01}\mid W_1))},\frac{M_iEN+N_iEM-2EMEN}{\mathrm{sd}(M_1EN+N_1EM)},\frac{\psi_{ii}/(M_iN_i)-E(\psi_{11}/M_1N_1)}{\mathrm{sd}(\psi_{11}/M_1N_1)}\right)^{\top}$$

in mean-square. The latter is an IID sum with finite covariance matrix and is asymptotically normal by the usual CLT. Applying the delta method with the function $(x,y,z)\mapsto(x/y,z)$, with derivative

$$\begin{pmatrix}1/y & -x/y^2 & 0\\ 0 & 0 & 1\end{pmatrix}$$

for $y\neq 0$, i.e., $E(M)\neq 0, E(N)\neq 0$, gives the asymptotic normality of $(\theta_{11},\theta_{12})$. The asymptotic covariance matrix is given by delta method. [[track down the 1/2 term–see notes P.3]]

□