Computational Journal of Mathematical and Statistical Sciences
3(2), 228–239
DOI:10.21608/CJMSS.2024.251550.1030
https://cjmss.journals.ekb.eg/

On the exchangeability property in causal models

Vincenzo Adamo

Italian Revenue Agency

* **Correspondence:** adamo.vincenzo@tiscali.it

Abstract: Exchangeability is one of the most important concepts in Bayesian probability theory [7], as well as in causal analysis, particularly within the theory based on the potential outcomes (see [18], [21], [23] and [15]).
In this paper, we propose a way to make explicit the link between the two concepts. We show they are almost coincident with the exchangeability property introduced by de Finetti [3], without making use of notions such as partial, conditional, or hierarchical exchangeability. To do this, we will start from the exchangeability property described in Greenland et al. [14], and assuming the use of a recursive linear Gaussian structural equation model, we will show how it is possible to exploit the properties of de Finetti's representation theorem, without performing any computation, to obtain an estimate of the average causal effect by calibrating a simple linear regression.
This is achieved by showing the role of a specific subset of the latent variables in the data-generating process for the variable Y|X=x, linking the exchangeability property required for the identification of the causal coefficient, with the non-correlation between regressors and error term in linear regression, needed to obtain an unbiased coefficient estimation. The results here proposed are not restricted to the Gaussian family of random variables distributions.

Keywords: exchangeability, Bayesian analysis, causal analysis, linear regression.
Mathematics Subject Classification: 68U02; 62F02.
Received: 30 December 2023; Revised: 23 January 2024; Accepted: 28 January 2024; Online: 25 April 2024.

1. Introduction

Exchangeability is one of the building blocks for assessing causal inference results, and every book concerning causal analysis explains the implication of this necessary condition, which allows us to deal with counterfactuals. Here, we want to restate the link between the property of exchangeability used in the context of causal analysis and the classical notion of exchangeability well known in the Bayesian field and due to de Finetti [3].

In the context of causal analysis, this property is related to the identifiability of the causal effect and is also known as the *strong ignorability* of Rubin and Rosenbaum [23], which refers to the condition $Y^{X=x} \perp\!\!\!\perp X$, or its conditional form $Y^{X=x} \perp\!\!\!\perp X|Z$ (where Y is the outcome, X the treatment and Z a covariate) in the context of potential outcomes. It has been further detailed in the literature in the same theoretical context (see, for example, [15] Technical Point 2.1), or it has been used as a synonym for a *no-confounding* condition (Greenland and Robins [12, 13], Greenland et al. [14], Pearl [19], Hernán and Robins [15]).

Exchangeability à la de Finetti appears explicitly in the pre-treatment exchangeability conditions, with features similar to the exchangeability described in [14], and the post−treatment condition both proposed by Dawid [9], as well as in the work of Lindley and Novick [17], Saarela et al. [24], Greenland and Robins [12, 13]and Greenland et al. [14], among others.

It is from the description given by Greenland et al. [14] that we will derive some steps, given below, to make more explicit the link between the two concepts of exchangeability, that is, to show that exchangeability in the causal domain is essentially exchangeability à la de Finetti.

This paper also mentions the concept of a hierarchical model (Bernardo and Smith [1]) related to the exchangeability of characteristic parameters of different sequences of partially exchangeable variables (de Finetti [6], Diaconis [10]) to describe a property of group exchangeability (of subjects selected to be treated or untreated). Recently, a conditional exchangeability property was also proposed by Saarela et al. [24], and within a randomized experiment, it involves sharing the same P parameter (in the representation theorem) among the output variables, covariates and unobservable variables (Y,X,U), as distinct from the P_Z parameter characterizing the treatment variable (Z). This structure could be a constraint on the contexts where the proposed model can be applied, although the main goal of such a model is to use exclusively statistical and Bayesian analysis concepts for the purpose of causal analysis, as in the work of Dawid [9].

In what follows, to link the classical concept of Bayesian exchangeability with the one used in causal analysis, it is shown how exchangeability between groups (treated and untreated) is essentially achieved through the use of Bayesian exchangeability applied to the probability functions of the two types of treatment assignment (together with the exchangeability of the latent data-generating process), without using partial, hierarchical or conditional exchangeability models.

In the context of recursive linear Structural equation models with Gaussian variables [2, 19], we achieve the goal of using linear regression to estimate the average causal effect (ACE), as reported in the literature [15, 19], without performing the computation contained in de Finetti's representation theorem formula but using only the properties of the variables described therein. The structural equation related to Y in the linear SEM is used to derive a linear regression that is estimable. This is accomplished by showing the properties of a subset of the latent variables which can entail both the exchangeability property for causal identification and the non-correlation between the regressor and the error term needed in the linear regression for an unbiased causal estimation.

The rest of the paper is summarized as follows: in the second paragraph, we will formalize a description of exchangeability and the representation theorem in Bayesian theory, also reporting insights from de Finetti on the topics of probability and causality. In the third paragraph, we will introduce some basic concepts of causal analysis and report the result of applying de Finetti's exchangeability in this context, relying on the characteristics of a subset **U**' of the latent variables. In the fourth paragraph, a couple of examples are reported where the role of a subset **U**' of the latent variables has as a consequence both the exchangeability property and the causal effect estimation via linear regression. Some brief final remarks will end the paper.

2. Exchangeability in Bayesian theory

The link between de Finetti and causality is closer than one might imagine. In *Probabilismo* [4], one of his earliest writings, he describes the existence, in everyone, of a probabilistic mental model of the reality around us, where cause-effect relationships are one of the components, having a probabilistic nature. This is an evolution of the thought of Hume, Poincaré, and Tilgher, among others, with Bayes' theorem playing a fundamental role in the human learning process.

At the end of the first paragraph, we read:

"[...] no science will permit us to say: this fact will come about, it will be thus and so because it follows from a certain law, and that law is an absolute truth. Still less will it lead us to conclude sceptically: the absolute truth does not exist, and so this fact might or might not come about, it may go like this or in a totally different way, I know nothing about it. What we can say is this: *I foresee* that such a fact will come about and that it will happen in such a way because past experience and its scientific elaboration by human thought make this forecast seem reasonable to me. Here, the essential difference lies in what the "why" applies to: I do not look for *why* THE FACT that I foresee *will come about* but *why* I DO *foresee* that the fact will come about. It is no longer the facts that need causes; it is our thought that finds it convenient to imagine causal relations to explain, connect and foresee the facts.".

Later, in paragraph 9, he writes:

"[...] But let us examine our conscience and see when it is that we admit that a circumstance can influence a certain event. Isn't it precisely when knowledge of it influences our probability judgment? [...] the concept of cause is only subjective, and it depends essentially on the concept of probability".

Today, these kinds of thoughts may seem almost part of the commonsense, for many people, but in 1929, when the text was written, the subjectivist theory of probability, the probabilistic but convergent nature of experience in prediction, and the understanding of a mental reality (model) as an instance of an external reality (and not the other way around) were by no means part of the mainstream scientific culture.

And again [5], on the fundamental concept of cause:

"Repetition and association give rise in us to the idea of cause, the anticipation of further repetition."

On the following definition of exchangeability, de Finetti himself said [8]:

"[...] exchangeability is a necessary condition because those conditions that are usually described in terms of independent and constant but unknown probabilities hold. However, the reciprocal statement is also true [...]".

Definition (*exchangeability* – de Finetti)

Let $\{X_i\}_{1..}$ be an infinite sequence of Boolean (Bernoulli) random variables. Suppose that the probability $Pr(x_1,..,x_n)$ of having observed $\mathbf{x} = (x_1,..,x_n)$ with n any finite number does not depend on the specific order, i.e., that $Pr(X_1{=}x_1,..,X_n{=}x_n) = Pr(X_1{=}x_{k(1)},..,X_n{=}x_{k(n)})$ for any permutation of the observations k. Then, the sequence of random variables X_i, is said to be *exchangeable*.

This property is related to the *homogeneity* or *similarity* of the X_i and is fundamental for inductive reasoning, where we learn from the past to predict the future, in contrast to the case of independence of the same variables (here independence is conditional). Under this light see equations (1) and (2) below, the former for estimating a subsequent realization $X_{n+1}=x_{n+1}$ and the latter for estimating the parameter underlying the conditional probability of the X_i.

If the X_i are i.i.d. $\Rightarrow$ the X_i are exchangeable, but the converse is not true, as the example of Polya's urn shows (distribution of exchangeable urns, identical, but not independent). Moreover, independence alone is not sufficient, as it is possible to have $Pr(A{=}1,B{=}0) = Pr(A{=}1)Pr(B{=}0) \neq Pr(A{=}0)Pr(B{=}1) = Pr(A{=}0,B{=}1)$, so independence does not imply exchangeability. The following relationships of set inclusion apply: i.d. $\supset$ exchangeability $\supset$ i.i.d. For Gaussian variables $Y_i \sim N(\mu,\sigma)$, the properties of i.i.d. and exchangeability are the same.

Representation Theorem (de Finetti).

Let $\{X_i\}_{1..}$ be an infinite sequence of Boolean (Bernoulli) random variables, exchangeable for each subsequence $\{X_i\}_{1..n}$ with $n>1$, and let $S_n = X_1{=}x_1+\ldots+X_n{=}x_n$ be the number of 1's in the subsequence of n elements. We will then have that there exists a random variable Θ and *a posteriori* distribution $F(\theta)$, with $dF(\theta) = \pi(\theta)d\theta = Pr(\theta)d\theta$:

1. $Pr\left(\lim_{n\to\infty} \bar{x}_n = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1..n} x_i = \Theta\right) = 1$, with $\Theta \sim F(\theta)$
2. $Pr(x_1,..,x_n) = \int_0^1 \theta^{S_n}(1-\theta)^{(n-S_n)}\, dF(\theta) = \int_0^1 \left(\prod_{i=1..n} \theta^{x_i}(1-\theta)^{(1-x_i)}\right) \pi(\theta)d\theta =^{*} \int_{\theta\in\Theta} Pr(x_1,...,x_n|\theta)Pr(\theta)d\theta = \int_{\theta\in\Theta} \left(\prod_{i=1..n} Pr(x_i|\theta)\right) Pr(\theta)d\theta$

where in 1) the random variable Θ does not necessarily converge to a constant value, and the product of probabilities in 2) is justified by the fact that $X_i \perp\!\!\!\perp X_j | \Theta{=}\theta$ with $i{\neq}j$. The theorem was generalized by Hewitt and Savage [16].

Before observing realizations X_i, any assumption about the density $\pi(\theta) = Pr(\theta)$ is admissible and will be the *a priori* density of Θ. After observing $x_1,..,x_n$ realizations, the (best) estimate of a future realization x_{n+1} is obtained using the *a posteriori* estimate of the probability of Θ (probability conditional on past observations $x_1,..,x_n$) and, in the case of Bernoulli variables, is given by the *a posteriori* mean of the variable Θ:

Pr(

Exchangeability in the causal analysis is one of the necessary conditions that allow the ACE of a generic treatment X on a response variable Y can be identified. A parameter of a probabilistic model is identifiable if it can be uniquely determined from the model's probability distribution, and if an estimator is used, it must be unbiased or at least consistent.

The other components required for the identification of the ACE are positivity, consistency, and non-interference among the treated variables (SUTVA). The reader can refer to [19] or [15] for further discussion on this topic.

* Written in its most general form (Bernardo et al., 1994).

In the following, we report some classical notations and results from causal analysis theory that will be used in demonstrating proposition 1, denoting X as the treatment variable and Y as the outcome variable.

1. For the sake of simplicity, we suppose that the joint distribution of variables in our model is Gaussian and described by a recursive linear Gaussian SEM (see [2]) with variables (Y, X, **W, K**, **U**), where **W** is a vector of observed variables not belonging to causal paths (X,Y), **K** is a vector of observed variables belonging to at least a causal path (X,Y), and **U** is a vector of unobservable and mutually independent Gaussian variables. Among the others, the reader can refer to Pearl's book [19] for an in-depth discussion of this linear causal structure, underlying the joint Gaussian distribution of the observed variables, and for the related direct acyclic graph (DAG) model useful for several of key concepts, like open/blocked path, $d-$separation, *backdoor* path, *backdoor* criterion, and so forth.
2. In detail, let

$\mathrm{Y} = f(\mathbf{W}_Y, \mathbf{K}_Y, \mathrm{U}_Y)$ (3)
be a linear function defining the variable Y in the recursive linear Gaussian SEM, with $\mathbf{K}_Y$ the generic vector of parents of Y belonging to at least one causal path (X,Y) and thus depending by the value X has been conditioned on. We denote by $\mathbf{K}_Y(\mathrm{X}=x)$ the $\mathbf{K}_Y$ vector related to X=x, pointing out that its realizations depend on X.

1. In the literature, we have several equivalent notations for the expression $Pr_{\mathrm{X}=x}(\mathrm{Y} = y)$ used in Greenland et al. (1999), and among them, we find that of Neyman-Rubin's potential outcomes $Pr(\mathrm{Y}^x = y)$ and that of Pearl's *do*(.) operator $Pr\,(\mathrm{Y} = y|do(\mathrm{X} = x))$.
2. Identification without *confounding* (Pearl 2009): $Pr\,(\mathrm{Y} = y|do(\mathrm{X} = x)) = Pr\,(\mathrm{Y} = y|\mathrm{X} = x)$.
3. Identification with *confounding* (*deconfounding* through the *adjustment formula*): $Pr\,(\mathrm{Y} = y|do(\mathrm{X} = x)) = \sum_z Pr\,(\mathrm{Y} = y|\mathrm{X} = x, \mathrm{Z} = z)\,Pr\,(\mathrm{Z} = z)$, with Z a sufficient variable (or set of variables) able to block all the *backdoor* paths from X to Y; see Pearl (2009).
4. Identification with *confounding* (*deconfounding* through IPTW): the *Inverse Probability of Treatment Weighted* (IPTW) method is used to obtain a new pseudo-population wherein the number of treated subjects is multiplied by $\frac{1}{PS}$ and the number of untreated subjects is multiplied by $\frac{1}{1-PS}$, where PS stands for the so-called *propensity score*. For further discussion of the method and the mathematical derivation of the PS, see, among others, [15], [19], and [25].
5. Although several methods are described in the literature to remove *confounding* between a pair of variables (X,Y) in an observational study, see the adjustment formula for example, in the following we will refer to the IPTW method. Also, it is supposed that no biases are present in the model (selection or others).
6. The use of the *randomized experiment* has historically been the golden method for eliminating (or reducing as much as possible) *confounding* and identifying the ACE even in a Bayesian context [21, 22]. We will assume irrelevance of any difference (due to chance) between the observable characteristics of two randomly drawn populations, so sufficiently large, resulting in a partition of the starting set of observations.
7. To align the conditions related to the *randomized experiment* with those of an *observational study* where the IPTW method is applied (absence of *confounding* between X and Y), we assume that all variables necessary to correctly characterize the study are observable.

From (3), we derive the following properties of the random variable Y conditional on X (Y|X).

1. The population represented by the available (potentially infinite) realizations of the variable Y($\mathbf{W}_Y$,$\mathbf{K}_Y$(X=x),U_Y) does not depend on the specific value X=x (see [20] Theorem 4.3.1), and we can obtain it as a realization of U_Y and the U_i∈**U** connected to Y by an open path, i.e., the U_i∈**U** able to influence one of its parents ($\mathbf{W}_Y$,$\mathbf{K}_Y$) joined U_Y. We denote this subset by **U**'⊆**U**. The U_i∈**U**' thus identified, whether they belong to causal or *backdoor* paths or some other type of open path to Y, turn out to be independent of X (**U**'⫫ X) after the IPTW procedure or randomization. Furthermore, the vector **U**' has its Gaussian distribution N(**0**,$\Sigma_{U'}$),with $\Sigma_{U'}$ a diagonal matrix, and any infinite random sample of realizations {**u**'$_i$}$_{1..}$ is exchangeable.
2. From a single realization of **u**'∈**U**', we can obtain several distinct realizations of (Y|X). For example, one conditional on X=x and one conditional on X=x' with x≠x'. Conversely, assigning a value to X, the value of (Y|X) is determined by at least one realization **u**'∈**U**'. In the following, this allows us to associate the same population, corresponding to an unbounded set of exchangeable realizations {**u**'$_i$}$_{1..}$, for at least a couple of realizations: one for (Y|X=0) and one for (Y|X=1). Moreover, to any exchange in a (unobservable) sequence {**u**'$_i$}$_{1..n}$ there is an admissible corresponding exchange in the related (observable) realizations {y_i|X=1}$_{1..n}$ and {y_i|X=0}$_{1..n}$.
3. Any randomly sampled sequence of realizations of the conditional variable (Y|X=x) are exchangeable because realizations of the same random variable (Y|X=x) ∼ N(μ_Y,σ_Y|X=x) = N(μ_Y+$\sigma_{YX/}\sigma_{XX}$(x–μ_X),σ_{YY}–$\sigma^2{}_{YX/}\sigma_{XX}$), where μ_Y and μ_X are the average values of the Y and X, while σ_{YX} = Cov(Y,X) and σ_{XX} = Var(X); see [26] page 189.
4. Given an unlimited (potentially infinite) sequence of realizations of (Y|X=x), we can partition that sequence into two unlimited sub-sequences that will have the same representativeness for the variable N(μ_Y,σ_Y|X=x) as the original one.
5. The probability $Pr\,(\mathrm{Y} = y|\mathrm{X} = x)$ is related to a Gaussian random variable, for which the mean depends on the specific value X=x, but the variance does not (see (c)).
6. In the following, we narrow the admissible values for the treatment variable X, focusing only on binary variables: X=1 for the treated and X=0 for the untreated, as in [14].

The deductions given in (a) and (b) are useful in defining a single infinite population of realizations of **U**' to be used in describing the variables (Y|X=0) and (Y|X=1) and their probabilities, while that in (d) will be used to partition the initial population into two sets of equal cardinalities (A and B).
The exchangeability described in [14], referring to two populations A and B, is formulated for two generic parameters of the probability distribution, μ_A and μ_B, and results in the following equation $Pr_{\mathrm{X}=x}\,(\mu_\mathrm{A},\mu_\mathrm{B}) = Pr_{\mathrm{X}=x}\,(\mu_\mathrm{B},\mu_\mathrm{A})$ where x∈{0,1}. Here, we will refer to the formulation of exchangeability with two vectors of the output variable of equal size, $\mathbf{y}_A$ and $\mathbf{y}_B$, for example, used for the computation of μ_A and μ_B, and belonging to two sets A and B that here form an equipartition of the randomly sampled population of realizations. This is a particular form of exchangeability, directly derivable from de Finetti's exchangeability [1], and implies that, relative to probability $Pr_{\mathrm{X}=x}$(.), the elements y_i, associated to {A∪B} and derived from **u**'$_i$, are i.i.d. conditional on a vector of parameters θ unknown to the analyst. Concerning (a), (b) and (e) above, we can refer to the same population {A∪B} induced by {**u**'$_i$}$_{1..}$ for the realizations of both $Pr_{\mathrm{X}=1}$(.) and $Pr_{\mathrm{X}=0}$(.). Moreover, about point (e) above, the Gaussian probability distributions $Pr_{\mathrm{X}=1}$(.)and$Pr_{\mathrm{X}=0}$(.) share the same standard deviation σ.

Proposition 1

Let us have a recursive linear Gaussian SEM with (3) describing the outcome variable Y, an infinite population of realizations of the vector **U**' defined as in (a), and an infinite sequence of realizations $\{Y_i|X{=}0\}_{1..}$ and $\{Y_i|X{=}1\}_{1..}$ both derived from $\{\mathbf{u}'_i\}_{1..}$. Let us randomly extract two sub-sequences $\{y_i|X{=}0\}_{1..\,2n}$ and $\{y_i|X{=}1\}_{1..2n}$, and randomly draw two sets, A and B, of cardinality n. It is assumed that the treatment performed on one subject does not influence the outcome of another subject (treated or untreated) and that each value of the treatment variable is associated with a positive probability. In the case of a *randomized experiment* or an *observational* setting where we have removed the *confounding* between X and Y by applying the IPTW method, the ACE of X on the outcome variable Y can be identified, and estimated by calibrating a simple linear regression of the form $\alpha + \beta x + \epsilon$.

Proof

We know that we are in the conditions under which the property $Pr\,(\mathrm{Y} = y|do(\mathrm{X} = x)) = Pr\,(\mathrm{Y} = y|\mathrm{X} = x)$ applies. Under the hypotheses described, *randomized experiment* or IPTW, de Finetti's exchangeability holds for random sampled observations $\{y_i|X{=}x\}_{1..2n}$ of $(Y_i|X{=}x) \sim N(\mu_x,\sigma)$, with mean μ_x and variance σ unknown to us, as independent realizations of variables with the same probability distribution (i.i.d.). We can imagine that we know the indices with the population observations belonging to A and B so that we write the specific property of exchangeability for $Pr_{\mathrm{X}=1}(.)$, $Pr_{\mathrm{X}=1}\,(y_{\mathrm{A}}, y_{\mathrm{B}}) = Pr_{\mathrm{X}=1}\,(y_{\mathrm{B}}, y_{\mathrm{A}})$, as any permutation π of the indices that exchanges the elements of A with those of B, for example, such that $\pi(\mathrm{A}i) = \mathrm{B}i$ and $\pi(\mathrm{B}i) = \mathrm{A}i$ (with a corresponding exchange in the set $\{\mathbf{u}'_i\}_{1..2n}$), and finally $Pr\,(y_{\mathrm{A}}, y_{\mathrm{B}}|\mathrm{X} = 1)$=$Pr\,(y_{\mathrm{B}}, y_{\mathrm{A}}|\mathrm{X} = 1)$. This is true for $Pr_{\mathrm{X}=0}(.)$ as well, with $(y_{\mathrm{A}}, y_{\mathrm{B}}|\mathrm{X} = 0)$ derived from the same $\{\mathbf{u}'_i\}_{1..2n}$ generating $(y_{\mathrm{A}}, y_{\mathrm{B}}|\mathrm{X} = 1)$.

Assuming we also know the three real parameters, $\mu_1 = \mu_{Y|X=1}$, $\mu_0 = \mu_{Y|X=0}$ and σ, which characterize the two realizations of $(Y_i|X{=}1)$ and $(Y_i|X{=}0)$, we can write the exchangeability as $Pr\,(y_{\mathrm{A}}, y_{\mathrm{B}}|\mathrm{X} = 1, \mu_1, \sigma)$=$Pr\,(y_{\mathrm{B}}, y_{\mathrm{A}}|\mathrm{X} = 1, \mu_1, \sigma)$. Additionally, from the i.i.d. of the individual conditional variables, we can write $Pr\,(y_{\mathrm{A}}, y_{\mathrm{B}}|\mu_1, \sigma)$ as $\prod_{i=1..n} Pr\,(y_{\mathrm{A}i}|\mu_1, \sigma) \prod_{i=1..n} Pr\,(y_{\mathrm{B}i}|\mu_1, \sigma)$: μ_1 and σ are sufficient statistics, so we omit X=1. Exchangeability and de Finetti's central limit theorem (Theorem 1, 1)) guarantee us that as the number of observations tends to infinity, the parameters vector θ of de Finetti's representation theorem will tend to the real parameters, and we can obtain $\lim_{n\to\infty} \theta_1 = (\mu_1, \sigma)$ for $Pr_{\mathrm{X}=1}(.)$ and $\lim_{n\to\infty} \theta_0 = (\mu_0, \sigma)$ for $Pr_{\mathrm{X}=0}(.)$. We have the equality $\lim_{n\to\infty} \prod_{i=1..n} Pr\,(y_{\mathrm{A}i}|\mu_1, \sigma) = \lim_{n\to\infty} \prod_{i=1..n} Pr\,(y_{\mathrm{B}i}|\mu_1, \sigma)$, because $\mathbf{y}_A$ and $\mathbf{y}_B$ both describe the same probability shape when the number of observations indefinitely increases. The same properties can be derived for $Pr_{\mathrm{X}=0}(.)$.

We are allowed to use the same unlimited sets A and B for both $Pr_{\mathrm{X}=0}(.)$ and $Pr_{\mathrm{X}=1}(.)$ because realizations deterministically induced by the same exchangeable set $\{\mathbf{u}'_i\}_{1..2n}$ of **U**' (see (b) described above) which is stochastically independent of X. This lets us identify the average causal effect through the following identity [14]:

$\mathrm{E}[\mathrm{Y}|do(\mathrm{X}{=}1)] - \mathrm{E}[\mathrm{Y}|do(\mathrm{X}{=}0)] = \mathrm{E}_A[\mathrm{Y}|do(\mathrm{X}{=}1)] - \mathrm{E}_A[\mathrm{Y}|do(\mathrm{X}{=}0)] = \sum_{y_{\mathrm{A}i}} y_{\mathrm{A}i} Pr_{\mathrm{X}=1}\,(y_{\mathrm{A}i}) - \sum_{y_{\mathrm{A}i}} y_{\mathrm{A}i} Pr_{\mathrm{X}=0}\,(y_{\mathrm{A}i}) = \sum_{y_{\mathrm{A}i}} y_{\mathrm{A}i} Pr_{\mathrm{X}=1}\,(y_{\mathrm{A}i}) - \sum_{y_{\mathrm{B}i}} y_{\mathrm{B}i} Pr_{\mathrm{X}=0}\,(y_{\mathrm{B}i}) = \sum_{y_{\mathrm{B}i}} y_{\mathrm{B}i} Pr_{\mathrm{X}=1}\,(y_{\mathrm{B}i}) - \sum_{y_{\mathrm{A}i}} y_{\mathrm{A}i} Pr_{\mathrm{X}=0}\,(y_{\mathrm{A}i}) = \sum_{y_{\mathrm{B}i}} y_{\mathrm{B}i} Pr\,(y_{\mathrm{B}i}|\mu_1, \sigma) - \sum_{y_{\mathrm{A}i}} y_{\mathrm{A}i} Pr\,(y_{\mathrm{A}i}|\mu_0, \sigma) = \mu_1 - \mu_0$.

Considering that we do not know the two distributions of the Gaussian response variable Y (we do not know the actual parameters μ_1, μ_0 and σ), however, we do know that the *unconfoundness* of (X,Y) is sufficient for the identifiability of E[Y|*do*(X=x)] and thus the computation of the average causal effect.

Moreover, E[Y|*do*(X=x)] has a linear form in X (see [15], [19]): E[Y|*do*(X=x)] = E[Y|X=x] = $\alpha + \beta x$. Note that the value of E[Y|X=x] before and after IPTW could be different.
Therefore, we can calibrate, using the OLS method, a simple linear regression of the form $\alpha + \beta x + \epsilon$ ([14] [25]) on the observed values of the populations A and B, using the values (y_{Ai}|X = 1) and (y_{Bi}|X = 0), for example, to estimate the average causal effect of X on Y: E[Y|X=1]−E[Y|X=0] = β = Cov(Y,X)/Var(X), as first shown by S. Wright [27], where Cov(Y,X) refers to the joint distribution of (Y,X).
Under the assumptions made of *no−confounding*, whether in the context of a *randomized experiment* or *observational study* with IPTW, the OLS method applied to the linear form with intercept, $\alpha + \beta x$, provides an unbiased and consistent estimate of E[Y|X=x]:
$E[Y|X{=}x] = \mu_Y + \frac{\sigma_{YX}}{\sigma_{XX}}(x - \mu_X) \approx \bar{y} + \hat{\beta}(x - \bar{x}) = (\bar{y} - \hat{\beta}\bar{x}) + \hat{\beta}x = \hat{\alpha} + \hat{\beta}x$ (4)
where $\bar{y}$ and $\bar{x}$ are the arithmetic mean of the y_i and x_i observations in the population {A∪B}, and σ_{YX}, σ_{XX} are scalar values (see among others [26] chapter 3). The OLS estimate is unbiased and the condition of the Gauss−Markov's theorem, referring to the lack of correlation between x and ϵ, is met (ϵ is the stochastic error of the linear regression $\alpha + \beta x + \epsilon$).
Actually, X and ϵ are uncorrelated (independent) because the variability of Y is inducted by the vector **U**', which is represented in the linear regression through the error term ϵ (ϵ is a function of **U**'), and is stochastically independent from X. Also, we see that Cov(Y,X) depends only on the causal paths between X and Y ([27], [19]), and the coefficient α does not depend on the values of Y conditional on X or even on the specific value X=x but on the X's mean value μ_X and the Y's mean value μ_Y. α deals with not-centred variables and also guarantee that E[ϵ] = 0. ◊

The previous result is a substantial rewriting, with a few modifications and the addition of some detailed steps, of what was reported in Greenland and Robins [12] and Greenland et al. [14], and allows, by using the exchangeability property, *no-confounding* and the assumptions about the probability distribution law of the outcome variables, to obtain an operational method for estimating the ACE: it is not necessary to apply de Finetti's representation theorem, but it is sufficient to use its peculiarities.
The main elements used to achieve the results described in proposition 1 are as follows:

1. the causal nature of the Cov(X,Y) covariance, i.e., obtained exclusively through causal paths, for a random experiment or an observational study with IPTW (*no-confounding*);
2. we identified the vector **U**' with the exchangeable sequence $\{\mathbf{u}'_i\}_{1..}$, as the core of the data-generating process for both the variable Y|X=1 and Y|X=0;
3. the property **U**'⫫ X let unchanged the Cov(X,Y), and Var(X) as well, when we use the same $\{\mathbf{u}'_i\}$ to switch from Y|X=1 to Y|X=0, leaving unchanged the causal effect in both scenarios. What does change (and we expect to be so) is the mean value from Y|X=1 to Y|X=0, for the identification and the estimation of the ACE;
4. as Pearl says [19], in an observational context the property **U**'⫫ X cannot be assured without a causal analysis of (what we know about) the data-generating process. It appears not to be only a statistical/probabilistic matter. In randomized experiments, with sufficiently large samples, this is assured by design and randomization imply exchangeability of the $\mathbf{u}'_i$.

It should be stressed that most of the results shown before are linked to an unlimited number of observations. When we deal with a finite number of observations, also in the case of randomized experiments, estimation of causal effects can be problematic because the influence of chance is no longer negligible and can bias (reducing or amplifying) the causal effect [12, 13, 21, 22] and flaws the population representativeness of the two sets A and B.

Finally, we want to add just one more thought regarding partial exchangeability and hierarchical exchangeable models. Suppose now that we have two parameters X_A and X_B that are subject to two sets of observations of Gaussian variables with the same standard deviation σ,$(y_{A1}, ..., y_{An})$ and $(y_{B1}, ..., y_{Bn})$, partially exchangeable, for which there is also a joint distribution $Pr(y) = Pr(y_{A1}, ..., y_{An}, y_{B1}, ..., y_{Bn})$. This time let us assume that the following joint probability form holds for the pair (X_A, X_B), $Pr(X_A{=}1, X_B{=}0) = Pr(X_A{=}0, X_B{=}1) = 1/2$, $Pr(X_A{=}0, X_B{=}0) = Pr(X_A{=}1, X_B{=}1) = 0$, and that 1 and 0 represent *a priori* values for the mean values μ_1, μ_0 associated with treated and untreated subjects. It is easily verified that the pair (X_A, X_B) is exchangeable, although it does not admit a representation theorem [10,11]. We thus obtain a hierarchical system of exchangeable values (y_i and x_i); see [1], although not hierarchically representable à la de Finetti.

If the exchangeability of (X_A, X_B) reflected an exchangeability at the level of a single pair of observations, $Pr(X_{Ai}{=}1, X_{Bi}{=}0) = Pr(X_{Ai}{=}0, X_{Bi}{=}1)$, this would represent precisely the equal probability of the Y_i observations to be able to flow into one or the other set (treated and untreated) and thus make them exchangeable, in the manner described in [12], as randomly drawn (uniform probability 1/2 is associated with the two treatment values used, with three values and three sets we would have probability 1/3, and so on).

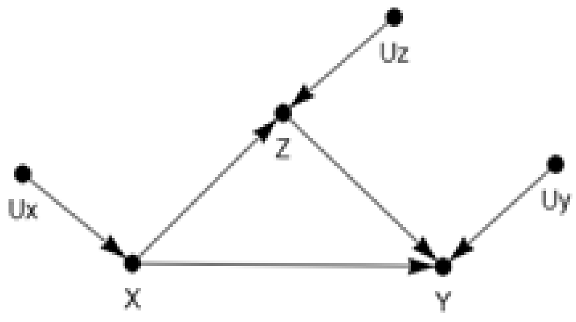
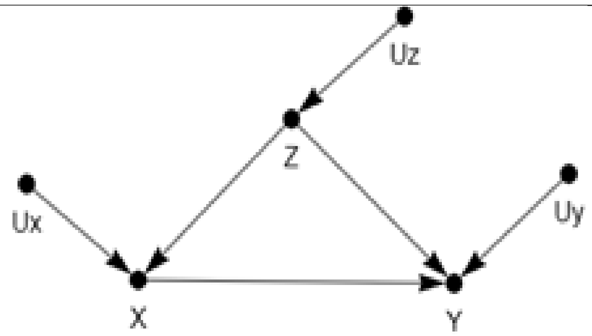
3. Examples

In this last section, we propose a couple of examples (see Table 1) showing the role of the subset of latent variables **U**' in the data-generating process for the variable Y|X=x, without confounding between X and Y, and the relationship with the error variable ϵ (additive noise) in the linear regression, used to estimate the beta value of the ACE of treatment X. The ACE we obtain could be the total effect (direct + indirect) or only the direct effect: it depends on the characteristics of the probabilistic model encoded within the DAG associated with the SEM.

The characteristics of this subset **U**' are useful both for deriving the exchangeability property à la de Finetti and for assessing the zero−correlation (independence) between regressors and error term needed to obtain an unbiased estimate of the beta coefficient in the linear regression (through the Gauss−Markov theorem).

In example A the covariance between X and Y depends only on the variance of X and the structural coefficients of the DAG. Note that the variable X is stochastically independent of both Uz and Uy, and from the linear regression $Y \sim \alpha + \beta x$ the total causal effect can be derived with coefficients α and β unbiased. Conditioning also on Z allows us to obtain only the direct effect. In example B, the covariance between X and Y depends on the variance of Z. It is therefore necessary to condition the latter variable to obtain the total causal effect (which is indeed the direct effect), and the coefficients of $Y \sim \alpha + \beta x + \delta z$ will be unbiased (both Z and X are stochastically independent of Uy). Note that in example B, if we used regression $Y \sim \alpha + \beta x + \epsilon$, the coefficient $\hat{\alpha}$ and $\hat{\beta}$ would be biased because the Gauss−Markov's conditions are violated: the variable Uz that contributes to the Y's dispersion and which is contained in the error term ϵ, is not independent of X and the two variables are correlated.

Table 1. Mediator and confounding graphs.

Example A: ***Mediator***	**Example B:** ***Confounding***
 	
SEM equations: Y = bX+cZ+Uy Z = aX+Uz X = Ux Ux~N(0,σ_{XX}), Uz~N(0,σ_{ZZ}), Uy ~N(0,σ_{YY}) mutually stochastically independents	SEM equations: Y = bX+cZ+Uy Z = Uz X = aZ+Ux Ux~N(0,σ_{XX}), Uz~N(0,σ_{ZZ}), Uy ~N(0,σ_{YY}) mutually stochastically independents
Cov(X,Y) = Cov(X, bX+cZ+Uy) = Cov(X, bX) + Cov(X, caX+cUz) + Cov(X, Uy) = bVar(X) + acVar(X) + cCov(X, Uz) = bVar(X) + acVar(X) = acVar(X)+bVar(X)	Cov(X,Y) = Cov(X, bX+cZ+Uy) = Cov(X, bX) + Cov(X, cZ) + Cov(X, Uy) = bVar(X) + cCov(X, Z) = bVar(X) + cCov(aZ, Z) + cCov(Ux, Z) = bVar(X) + acVar(Z) = acVar(Z)+bVar(X)
SEM equation of Y wrt X: Y $\sim$ bX+c(aX +Uz)+Uy = (b+ac)X+cUz+Uy = βX+ϵ where β=b+ac, ϵ=cUz+Uy and X$\perp\!\!\!\perp$(Uz,Uy).	SEM equation of Y wrt X: Y $\sim$ bX+cUz+Uy = βX+ϵ where β=b, ϵ=cUz+Uy and X$\perp\!\!\!\perp$Uy, but *not*(X$\perp\!\!\!\perp$Uz) imply confounding on (X,Y).
Causal effect identified by: Y $\sim \alpha + \beta x + \epsilon$ ACE (total) τ_{YX}: $\hat{\beta}$ = ac+b Variability of ϵ being induced by {Uz,Uy}=**U**' $\hat{\beta}$not biased: E[$\hat{\beta}$] = b+ac. α deals with not-centred variables [26].	Causal effect identified by controlling also on Z $\Rightarrow$ Uz$\notin$**U**': Y $\sim \alpha + \beta x + \delta z + \epsilon$ ACE (direct) τ_{YX}: $\hat{\beta}$ = b Variability of ϵ being induced by {Uy}=**U**' $\hat{\beta}$not biased: E[$\hat{\beta}$] = b. α deals with not-centred variables [26].

4. Conclusions

In this paper, we have proposed a rewriting of the exchangeability property used in the context of causal analysis to make more explicit its relationship with the exchangeability property introduced by de Finetti [3] and of fundamental importance in the context of Bayesian probability theory.

No notions such as partial, conditional, or hierarchical exchangeability were used. We started from the exchangeability property described by Greenland et al. [14] and show how it is possible to exploit the properties of de Finetti's representation theorem, without performing any computation, to obtain exchangeability between groups (treated and untreated) from the classical Bayesian exchangeability property.

Using a recursive linear Gaussian structural equation model, we also related the exchangeability result to the classical way of estimating the average causal effect through calibration of a simple linear regression, although these results can be extended to other kinds of probability distributions. This has been achieved using the characteristics of a subset of the latent variables useful both for deriving the exchangeability property à la de Finetti and for assessing the independence between regressors and error term in the linear regression. A further step will be generalizing these results to a broader class of probability distributions.

References

1. Bernardo, J.M. (1996). *The concept of exchangeability and its applications*. Far East Journal of Mathematical Sciences, Spec. II, 111–121.
2. Bollen, K.A. (1989). *Structural Equations with Latent Variables*. Wiley.
3. de Finetti, B. (1930). *Funzione caratteristica di un fenomeno aleatorio*. Memorie della Regia Accademia Nazionale dei Lincei VI, 86:1930.
4. de Finetti, B. (1931). *Probabilismo*, Napoli, Logos14, 163–219.
5. de Finetti, B. (2006). "L'invenzione della verità" Raffaello Cortina Editore. Original text dated Nov. 16, 1934.
6. de Finetti, B. (1938). *Sur la condition d'équivalence partielle*. Actual. Sci. Ind. 739.
7. de Finetti, B. (1970). *Teoria delle Probabilità*. Torino: Einaudi. English translation as Theory of Probability in 1974, Chichester: Wiley.
8. de Finetti, B. (1995) *Filosofia della probabilità*, ed. by Alberto Mura, Milano: Il Saggiatore, 1995.
9. Dawid, A.P. (2021) *Decision-theoretic foundations for statistical causality*. J Causal Infer. 9:39–77.
10. Diaconis, P. (1988). *Recent progress on de Finetti's notions of exchangeability*. In Bayesian Statistics. Oxford Sci. Publ. 111–125. Oxford Univ. Press, New York.
11. Diaconis, P. (1977). *Finite forms of de Finetti's theorem on exchangeability*. Synthese 36, 271–281.
12. Greenland, S. and Robins, J.M. (1986). *Identifiability, exchangeability, and epidemiological confounding*. Int. J. Epidemiol. 15 413–419.
13. Greenland, S. and Robins, J.M. (2009). *Identifiability, exchangeability and confounding revisited*. Epidemiol Perspect Innov. Sep 4;6(4).

14. Greenland, S., Robins, J.M. and Pearl, J. (1999). *Confounding and collapsibility in causal inference*. Statistical Science, Vol. 14, No. 1, pp. 29-46.

15. Hernán, M.A., Robins, J.M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

16. Hewitt, E. and Savage, L.J. (1955). *Symmetric measures on Cartesian products*. Transactions of the American Mathematical Society, 80:470–501.

17. Lindley, D.V. and Novick, M.R. (1981). *The role of exchangeability in inference*. Ann. Statist. 9 45–58.

18. Neyman, J. (1923). *Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes.* Master's Thesis. Excerpts reprinted in English, Statistical Science, Vol. 5, pp. 463–472. (D. M. Dabrowska, and T. P. Speed, Translators.).

19. Pearl, J. (2009). *Causality: models, reasoning, and inference*, 2nd edn. Cambridge University Press, Cambridge

20. Pearl, J., Glymour, M. and Jewell, N.P. (2016). *Causal inference in statistics: a primer*. Wiley, Chichester.

21. Rubin, D.B. (1974) *Estimating causal effects of treatments in randomized and nonrandomized stidies*. Journal of Educational Psycology. Vol. 66, No. 5, 688−701.

22. Rubin, D.B. (1978) *Bayesian inference for causal effects: the role of randomization*. The Annals of statistics. Vol. 6, No. 1, 34−58.

23. Rubin, D.B., Rosenbaum, P.R. (1983). *The Central Role of the Propensity Score in Observational Studies for Causal Effects*. Biometrika. 70 (1): 41–55.

24. Saarela, O., Stephens, D.A., Moodie, E.E.M. (2023). *The Role of Exchangeability in Causal Inference*. Statist. Sci. Advance Publication 1 - 17.

25. Shiba, K., Kawahara, T. (2021). *Using Propensity Scores for Causal Inference: Pitfalls and Tips*. J Epidemiol. Aug 5;31(8):457-463.

26. Theil, H. (1971). *Principles of Econometrics*. John Wiley & Sons.

27. Wright, S. (1921). *Correlation and causation*. Journal of Agricultural Research. 20: 557-585.

SASAR
Open Access Journal