Bias in Statistics and Causal Reasoning

Fabrizia Mealli

Department of Statistics, Computer Science, Applications
Florence Center for Data Science
University of Florence
fabrizia.mealli@unifi.it

Interdisciplinary Approaches for Bias Elicitation, NoBIAS, Pisa, May 4 2022

Introduction

- Bias in statistics
- Apparent statistical paradises created by Big Data
- More data, less uncertainty
- Is this really true?
- Large sample asymptotics (LLN, CLM)
- Sample size that matters n, not population size N
- Which one should I trust more: a 1% survey with 60% response rate or a self-reported administrative dataset covering 80% of the population?
- It depends on the goal and what we mean with "trust"
- Because statistics is a principled thinking and methodology development for dealing with uncertainty, we should be able to formally answer that question

Typical goals in statistical inference

- Descriptive vs causal quantities
- Finite population vs infinite (super) population
- Point estimation
- Set (interval) estimation
- Hypothesis testing (discovery)
- Prediction
- Typically we would like to infer, derive, statements that are valid "in general"

A simple example

- Population mean or population proportion $\mu = \bar{X}_N$
- (Finite) population inference (from big data)
- Random sampling allows to quantify uncertainty of sample average for fixed n, $\bar{x}_n$, namely $\sigma_X/\sqrt{n}$, with σ_X the standard deviation of X.
- Bias $E(\bar{x}_n - \bar{X}_N) = 0$
- Estimation error $\bar{x}_n - \bar{X}_N$
- Mean squared error MSE$= E(\bar{x}_n - \bar{X}_N)^2 = \sigma_X^2/n$
- Those concepts are useful also for Bayesian inference to assess the operating characteristics of some posterior summaries

Goal of ML tools

- Big data and machine learning opened a new field
- Having machines, algorithms, self extracting information from data, discovering the structure themeselves
- Tasks of supervised learning
- Tasks of unsupervised learning
- Tipically lead to prediction or classification not associated with measure of uncertainty
- Uncertainty depends on a lot of things: quality, quantity and characteristics of the data
- For some decisions not useful, but for others crucial
- In statistics we know that, and we also know that the further away I am from the support of observed features distribution the more uncertainty we have
  But it is not just that

Estimation error with nonrandom sampling - 1

The difference between estimate and estimand (estimation error) depends on

- a data quality measure, $\rho(R, X)$, the correlation between X and the response/recording indicator R
- a data quantity measure, $\sqrt{(N-n)/n} = \sqrt{(1-f)/f}$ with $f = n/N$;
- a problem difficulty measure, σ_X

(Meng, 2018)

Estimation error with nonrandom sampling - 2

- Probabilistic sampling ensures high data quality by controlling $\rho(R, X)$
- When we lose this control, the impact of N is no longer cancelled out by $\rho(R, X)$, leading to a Law of Large Populations (LLP), that is, our estimation error, relative to the benchmarking rate $n^{-1/2}$, increases with $N^{1/2}$
- The *bigness* of such Big Data (for population inferences) should be measured by the relative size $f = n/N$, not the absolute size n.

(Meng, 2018)

Estimation error with nonrandom sampling - 3

- When combining data sources for population inferences, those relatively tiny but higher quality ones should be given far more weights than suggested by their sizes.
- Estimates obtained from the Cooperative Congressional Election Study (CCES) of the 2016 US presidential election suggest a $\rho(R, X) \approx -0.005$ for self-reporting to vote for Donald Trump. Because of LLP, this seemingly minuscule data defect correlation implies that the simple sample proportion of the self-reported voting preference for Trump from 1% of the US eligible voters, that is, $n \approx 2,300,000$, has the *MSE* as the corresponding sample proportion from a genuine simple random sample of size $n \approx 400$, that is a 99.98% reduction of sample size.
- On average, the larger the state's voter populations, the further away the actual Trump vote shares from the usual 95% confidence intervals based on the sample proportions.
- This should remind us that, without taking data quality into account, population inferences with Big Data are subject to a Big Data Paradox: the more the data, the surer we fool ourselves.

Some formulas

- Let R_j be the recording indicator. The letter R represents the R-mechanism which may not be probabilistic
- Write $\bar{x}_n = \frac{\sum_1^N X_j R_J}{\sum_1^N R_J}$
- For Random sampling, $R = \{R_1, \cdots, R_N\}$ has a well-specified joint distribution
- Using the fact that the variance of the binary R_J is $V_J = f(1-f)$, we have

$$\bar{x}_n - \bar{X}_N = \rho(R, X) * \sqrt{(1-f)/f} * \sigma_X$$

that is Data Quality*Data Quantity*Problem Difficulty

- Compensating for quality with quantity is a doomed game

- This identity implies that once we lose control of probabilistic sampling, then the driving force behind the estimation error is no longer the sample size n, but rather the population size N.
- $$\frac{\bar{x}_n - \bar{X}_N}{\sqrt{V_{SRS}}} = \sqrt{N-1}\rho(R,X)$$
- A butterfly effect: The return of the long-forgotten monster N. To deliver how much damage a seemingly small $\rho(R,X)$ can cause we can observe that any routinely used confidence intervals of the form

  $$\bar{x}_n - M\hat{\sigma_X}/\sqrt{n}; \bar{x}_n + M\hat{\sigma_X}/\sqrt{n}$$

  will almost surely miss $\bar{X}_n$ for any conventional choice of the multiplier M unless we adopt an estimate of the standard deviation that overestimates σ_X by orders of magnitude to compensate for the colossal loss of the sample size.
- Worse, since the interval width shrinks with the apparent size n, our false confidence may increase with n, despite the fact that the interval has little chance to cover the truth because it is so precisely centered at a wrong location!!!

- A big data paradox? We statisticians certainly are responsible for the widely held belief that the population size N is not relevant for inference concerning population means and alike, as long as N is sufficiently large.
- But apparently we have been much less successful in communicating the "warning label" that this assertion is valid only if one has strict control of the sampling scheme (via probabilistic schemes).

A paradox? Not really

- Big Data Paradox: The bigger the data, the surer we fool ourselves.
- The Big Data Paradox is in the same spirit as Simpson Paradox
- These kinds of statistical phenomena are not paradoxes in mathematical or philosophical senses
- But they appear to be paradoxical because of our mis-formed or mis-informed intuitions. Here the phrase Big Data refers to those big datasets with an uncontrolled (or unknown) R-mechanism.
- If our big datasets possess the same high quality as those from well designed and executed probabilistic surveys in terms of $\rho(R, X)$, then we are indeed in paradise!
- In terms of information gathering-nothing beats high quality big data!

(Meng, 2018)

Missing Data and Missing Data Mechanisms

- The R-mechanism is the process that creates observed and missing data
- Rubin (1976) and Mealli and Rubin (2015) formalize the assumptions on such mechanisms
  - Missing Complete at Random (MCAR)
  - Missing at Random (MAR)
  - Missing Not at Random (MNAR)
  - Ignorability
- These assumptions allow to know when complete data analysis, single imputation, multiple imputation, likelihood analysis, Bayesian analysis lead to biased or unbiased estimation of quantities of interests

What about causality and causal inference?

- Research questions that motivate most studies in statistics-based sciences are causal in nature
- What can statistics say about causation?
- The usual motto is "correlation is not causation"
- Dominant methodology has excluded causal vocabulary both from its mathematical language and from its educational programs
- Yet, statisticians invented randomized experiments, universally recognized as a powerful aid in investigating causal relationships

- Statistics has a great deal to say about certain problems of causal inference
- Statistical models used to draw **causal inferences** are different from those commonly used to draw **associational inferences**
- Variety of questions under causality heading
  - ✓ the philosophical meaningfulness of the notion of causation
  - ✓ deducing the causes of a given effect
  - ✓ understanding the details of a causal mechanism
- I will focus on measuring the effects of causes because this seems to be a place where statistics, which is concerned with measurement, has major contributions to make

- The purpose is to present a model that is complex enough to allow us to formalize basic intuitions concerning causes and effects, to define causal effects and to make assumptions allowing estimation of such effects clear and explicit
- A statistical framework for causal inference is the one based on potential outcomes.
  - ✓ It is rooted in the statistical work on randomized experiments by Fisher (1918, 1925) and Neyman (1923), as extended by Rubin (1974, 1976, 1977, 1978, 1990a,b) and subsequently by others to apply to nonrandomized studies and other forms of inference
  - ✓ See Imbens and Rubin (2015) for a textbook discussion
- This perspective was called "Rubin's Causal Model" by Holland (1986) because it viewed causal inference as a problem of missing data, with explicit mathematical modeling of the assignment mechanism as a process for revealing the observed data (Ding and Li, 2018).

Associational Inference vs Causal Inference

- Standard statistical models for associational inference relate two (or more) variables in a population
- The two variables, say Y and A, are defined for each and all units in the population and are logically on equal footing
- Joint distribution of Y and A
- Associational parameters are determined by this joint distribution: for example, the conditional distribution of Y given A describes how the distribution of Y changes as A varies
- A typical associational parameter is the regression of Y on A, that is, the conditional expectation $E(Y|A)$
- Associational inference is simply descriptive
- Role of time

Associational Inference vs Causal Inference

- Causal inference is different
- Use of language of experiments
- Model for causal inference starts with a population of units (persons, places, or things at a particular point in time) upon which a cause or a treatment may operate or act
- A single person, place, or thing at two different times comprises two different units
- The terms **cause** and **treatment** will be used interchangeably
- The effect of a cause is almost always relative to another cause: "A causes B" means relative to some other condition that may include "not A"
- The language of experiments: "treatment" vs "control"
- The key notion in causal inference is that each unit is potentially exposable to any one of the causes.
  - ✓ "She did well in the math test because she received good teaching"
  - ✓ "She did well in the math test because she is a girl"

Introducing Model and Notation

- Let W be the variable that indicates the treatment, 0 or 1, to which each unit is exposed
- The critical feature of the notion of a cause is that the value of W for each unit **could have been different**
- W must be a variable that is, at least in principle, manipulable
- Role of time: the fact that a unit is exposed to a cause or treatment must occur at a specific time
- Pre-exposure or pre-treatment variables, sometimes labelled covariates, X, whose values are determined prior to exposure to the cause
- Post-exposure or response variables, Y, on which to measure the effect of the cause

Introducing Model and Notation

- To represent the notion of causation, we postulate the existence of two variables, $Y(1)$ and $Y(0)$ for each unit, which represent the potential responses or potential outcomes associated with the two treatments
- These are the values of a unit's measurement of interest after (a) application of the treatment and (b) non-application of the treatment (i.e., under control)
- A causal effect is, for each unit, the comparison of the potential outcome under treatment and the potential outcome under control
- For example, we can say that treatment 1 (relative to treatment 0) causes the effect $Y_i(1) - Y_i(0)$ for unit i

The Science

Units	Covariates	Potential Outcomes		Unit-level	Summary
	X	$Y(1)$	$Y(0)$	Causal Effects	Causal Effects
1	X_1	$Y_1(1)$	$Y_1(0)$	$Y_1(1)$ vs $Y_1(0)$	Comparison of
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$Y_i(1)$ vs $Y_i(0)$
i	X_i	$Y_i(1)$	$Y_i(0)$	$Y_i(1)$ vs $Y_i(0)$	for a common
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	set of units
N	X_N	$Y_N(1)$	$Y_N(0)$	$Y_N(1)$ vs $Y_N(0)$	

- "The fundamental problem of causal inference": each potential outcome is observable but we can never observe all of them
- Summary causal effects: the critical requirement is that for a comparison to be causal it must be a comparison of $Y_i(1)$ and $Y_i(0)$ on a common set of units

What we are able to observe

Units	Covariates X	Treatment W	Potential Outcomes $Y(1)$	$Y(0)$	Unit-level Causal Effects
1	X_1	1	$Y_1(1)$	?	?
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	X_i	0	?	$Y_i(0)$	?
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	X_N	1	$Y_N(1)$	?	?

SUTVA

- The table for the Science requires the Stable Unit Treatment Value Assumption (SUTVA, Rubin, 1990b) to be adequate
- No interference between units, that is, neither $Y_i(1)$ nor $Y_i(0)$ is affected by what action any other units received
- No hidden version of treatments: no matter how unit i received treatment 1, the outcome that would be observed would be $Y_i(1)$
- Also implicit in the representation is that the Science is not affected by how or whether we try to learn about it, whether by randomized block designs, observational studies or other methods

SUTVA and Other Assumptions

- Without these assumptions causal inference using potential outcomes is not impossible, but it is far more complicated
- SUTVA is commonly made, or studies are designed to make SUTVA plausible
- Nothing is wrong with making assumptions and causal inference is impossible without making assumptions; assumptions are the strand that links statistics to science
- It is the scientific quality of the assumptions, not their existence, that is critical
- In causal inference assumptions are always needed, and they typically do not generate testable implications, so it is imperative that they are explicated and justified

Scientific and Statistical Solutions

- Because at least half of the potential outcomes are always missing, as such, the fundamental problem of causal inference is not solved by observing more units
- The notation explicitly representing both potential outcomes is an exceptional contribution to causal inference
- Despite its apparent simplicity it did not arise until 1923 with the work of Neyman and only in the context of completely randomized experiments
- We had to wait until the seventies with the work of Rubin to use the notation of potential outcomes to describe causal effects in any setting, including observational studies
- Despite the fundamental problem of causal inference, there are some solutions to the fundamental problem

Scientific and Statistical Solutions

- The scientific solution exploits various homogeneity or invariance assumptions
  - ✓ $Y_i^{t-1}(0) = Y_i^t(0)$
  - ✓ Then, expose units to 1 and measure $Y_i(1)$
  - ✓ The scientist has made an untestable homogeneity assumption
- Science has made enormous progress using this approach, and it is the approach that we informally use often in our lives
- The statistical solution uses the observed values of W and $Y(W)$, together with assumptions about the way units where exposed to either $W = 1$ or $W = 0$ to address the problem

The Role of the Assignment Mechanism

- The key in Rubin's work is to see randomization as just one way to create missing and observed data in the potential outcomes
- There are many other processes for creating missing data and those were called **assignment mechanisms** (Rubin, 1978)
- The assignment mechanism gives the probability of each vector of assignments, W, given the Science:

$$Pr(W \mid X, Y(1), Y(0))$$

- Before Rubin (1975), there were written descriptions of assignment mechanisms, but no formal mathematical statement or notation showing the possible dependence of treatment assignments on BOTH potential outcomes
- Y_{obs}: the collection of observed potential outcomes, with $Y_{obs,i} = W_i Y_i(1) + (1 - W_i) Y_i(0)$
- Y_{mis}: the collection of missing or unobserved potential outcomes, with $Y_{mis,i} = (1 - W_i) Y_i(1) + W_i Y_i(0)$

The Role of the Assignment Mechanism

- The definition of the assignment mechanism states that probability of something that we *do now*, W, can depend, not only on things that we observe now, X, or even Y_{obs} in sequential experiments, but moreover on other things that will never even be realized, Y_{mis}. Yet, as a formal probability statement, it is mathematically coherent
- Understanding the assignment mechanism's possible dependence on values of the potential outcomes: think of unobserved - to the analyst of the data - covariates U that are associated with the future potential outcomes and are used by the assigner of treatments, hypothetical or real, in addition to X
- $Pr(W \mid X, Y(1), Y(0), U) = Pr(W \mid X, U)$
- When this expression is averaged over the values of U for fixed values of X, $Y(1)$, $Y(0)$ to calculate the assignment mechanism, the result yields dependence on $Y(1)$, $Y(0)$

Types of Assignment Mechanism

- The assignment mechanism is unconfounded (with the potential outcomes, Rubin, 1990b) if:

$$Pr(W \mid X, Y(1), Y(0)) = Pr(W \mid X)$$

- An unconfounded assignment mechanism is probabilistic if all the unit-level probabilities, the propensity scores (Rosenbaum and Rubin, 1983), are strictly between zero and one:

$$0 < e_i = Pr(W_i \mid X) < 1$$

- An unconfounded probabilistic assignment mechanism is called strongly ignorable
- Classical randomized experiments are special cases of strongly ignorable assignment mechanisms
- In observational studies the assignment mechanism is not known and we need to make assumptions in order to be able to draw inference on causal effects
- Design stage of observational studies
- Big data, and machine learning are not a substitute of a thoughtful study design, nor can overcome issues regarding data quality, missing confounders, interference, and extrapolation (Bargagnli-Stoffi, Dominici and Mealli, 2021)

Confounding bias

- Confounding (or common cause) is the main complication/hurdle between association and causation
- Examples of Confounding
  - ✓ Education and income. Confounder: SES of family
  - ✓ Medical treatment and patient outcome. Confounders: age, sex, other complications
- An extreme example of confounding is Simpson's paradox: confounder reverses the sign of the correlation between treatment and outcome
  - ✓ Simpson's paradox or Yule-Simpson effect: a trend appears in different groups of data but disappears or reverses when these groups are combined

(Pearson et al., 1899; Yule, 1903; Simpson, 1951)

Simpson's paradox: Hypothetical Example

- Target Population: Unemployed workers looking for a job
- Treatment: Participation ($W_i = 1$) versus no participation ($W_i = 0$) in a training program
- Outcome: Employment status one year after the end of the training program
- Proportion of employed subjects one year after end of the training program by training program participation for the all sample and two sub-samples with and without high school degree

High school degree	$W_i = 0$	$W_i = 1$
No	0.87 (335/386)	0.93 (116/125)
Yes	0.70 (80/114)	0.73 (275/375)
All	0.83 (415/500)	0.78 (391/500)

- The paradox arises because subjects with no high school degree, before treatment assignment, are less likely to participate in the training program

Unconfoundedness

- Unconfoundedness: $\{Y(1), Y(0)\} \perp W|X$.
- Unconfoundedness is an assumption on unmeasured data, and hence inherently untestable: the data (no matter how big!) are uninformative about the distribution of $Y(0)$ for treated units and $Y(1)$ for control units
- Unconfoundedness implies, within strata of observed covariates, potential outcomes corresponding to both treatment conditions would be balanced between groups
- Re-think balance: randomization balance both covariates and potential outcomes
- What we really want to balance in observational studies: potential outcomes between groups
- Specifically, we want to balance:
  $\Pr(Y(0)|W = 1)$ vs. $\Pr(Y(0)|W = 0)$, and
  $\Pr(Y(1)|W = 1)$ vs. $\Pr(Y(1)|W = 0)$
- In practice, we use balance in covariates as a proxy to balance in potential outcomes

Distinguishing between the Science and the Assignment Mechanism

- Using the potential outcomes notation maintains the critical distinction between **what we are trying to estimate**, the Science, and **what we do to learn about it**, the assignment mechanism
- This distinction was in the work of Neyman or Fisher, so that extensions to observational studies of classical methods of inference in randomized experiments, due to Fisher (1925) and Neyman (1923), are natural within the RCM framework
- We cannot formally state the benefit of randomized experiments using the observed outcome notation Y_{obs}, which mixes up the Science with how we learn about the Science, the assignment mechanism
- Yet the reduction to the observed outcome notation is exactly what regression approaches, path analyses, directed acyclic graphs (DAGs), etc. essentially compel us to do (Rubin, 2005)

Modes of Inference: Causal Inference Based Solely on the Assignment Mechanism

- Both Fisher and Neyman proposed methods of causal inference based solely on the randomization distribution of statistics induced by classical randomized assignment mechanisms
- **Fisher's Exact p-values for Sharp Null Hypotheses**
- Fisher's method was essentially a stochastic proof by contradiction
- He wanted to prove that $H_0 = Y_i(1) = Y_i(0) \forall i$ is wrong using the randomization distribution under H_0

Modes of Inference: Causal Inference Based Solely on the Assignment Mechanism

- **Neyman's Randomization-Based Estimates and Confidence Intervals**
- Neyman (1923) showed that, in a completely randomized experiment, $\bar{y}_1 - \bar{y}_0$ is unbiased (averaging over all randomizations) for the average causal effect and propose a large-sample interval estimate for the average causal effect, which became the standard one in much of statistics and applied fields
- Neyman's approach has advantages over Fisher's in that it can deal with random sampling of units from a population; much of the theory behind propensity score methods is generalization of Neyman's approach
- Fisher's approach has the obvious advantage in not requiring large samples for the exactness of its probabilistic statements
- Fisher's and Neyman's approaches rarely addressed the real reasons we conduct studies: to learn about which interventions should be applied to future units
- **The third leg of the RCM is critical: pose a model on the Science and derives the Bayesian posterior predictive distribution of the missing potential outcomes**

Elements of the RCM

- The first leg is using potential outcomes to define causal effects no matter how we try to learn about them: *First define the Science*
- The second leg is to describe the process by which some potential outcomes will be revealed: *Second, posit an assignment mechanism*
- The third leg is placing a probability distribution on the Science to allow formal probability statements about the causal effects: *Third, incorporate scientific understanding in a model for the Science.*
- The Bayesian approach directs us to condition on all observed quantities and predicts, in a stochastic way, the missing potential outcomes of all units, past and future, and thereby makes informed decisions

Bayesian Model-Based Imputation

- The benefits of modeling the science in causal inference include the ability to deal with more complex situations and to summarize results more logically
- We directly confront the fact that at least half of the potential outcomes are missing and create a posterior predictive distribution for them
- From a model on the science, $Pr(X, Y(1), Y(0))$, and the model for the assignment mechanism, we can find the posterior predictive distribution of Y_{mis}, given the observed values of W, X, and Y_{obs}

$$Pr(Y_{mis}|X, Y_{obs}, W) \propto Pr(X, Y(1), Y(0))Pr(W|X, Y(1), Y(0))$$

- We can calculate the posterior distribution of any causal estimand by multiply imputing Y_{mis}: draw a value of Y_{mis}, impute it, calculate the causal estimand, redraw Y_{mis}, and so on

Bayesian Model-Based Imputation

- Two critical facts simplify this approach

$$Pr(X, Y(0), Y(1)) = \int \prod f(X_i, Y_i(0), Y_i(1)|\theta)p(\theta)d\theta,$$

where $f(\cdot|\theta)$ is an iid model for each unit's science given a hypothetical parameter θ with prior (or marginal) distribution $p(\theta)$

- This modelling task if far more flexible than specifying a regression model
- If the treatment assignment mechanism is ignorable then when the expression for the assignment mechanism is evaluated at the observed data, it is free of dependence on Y_{mis}.
- So the explicit conditioning on W can be ignored (hence the term ignorable assignment mechanism):

$$Pr(Y_{mis}|X, Y_{obs}, W) \propto Pr(Y_{mis}|X, Y_{obs})$$
$$Pr(Y_{mis}|X, Y_{obs}) = \int Pr(Y_{mis}|X, Y_{obs}, \theta)Pr(\theta, X, Y_{obs})d\theta$$

where $Pr(\theta|X, Y_{obs})$ is the posterior distribution of θ, equal to the prior distribution $p(\theta)$ times the likelihood of θ

Bayesian Model-Based Imputation

- Thus by supplementing the assignment mechanism with a model on the science, we can adopt, a Bayesian framework to inference for causal effects
- The Bayesian perspective is extremely flexible and is especially convenient for summarizing the current state of knowledge about the science in complex situations
- Assuming this summary of the current state of knowledge is accurate, this can be combined with various assessment of costs and benefits of various decisions to choose which decision to make (Dehejia, 2003)

Extensions

- The potential outcome framework combined with Bayesian inference allowed us to make enormous progress in formalizing and solving problems in both randomized and observational studies
- The framework allowed to understand the meaning of IV estimation developed in Econometrics, by bridging randomized experiments with noncompliance with IV settings (Angrist, Imbens and Rubin, 1996)
- It provided insights into understanding causal mechanisms through *principal stratification* (Frangakis and Rubin, 2002), an approach to handling intermediate variables within the RCM

References

- Angrist J. D., Imbens G. W., and Rubin D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American statistical Association*, 91(434), 444-455.
- Bargagli Stoffi F., Dominici F. and **Mealli F.** (2021). From Controlled to Undisciplined data: Estimating Causal Effects in the Era of Data Science Using a Potential Outcome Framework. Harvard Data Science Review, 3(3).
- Dehejia R. H. (2003). Was there a Riverside miracle? A hierarchical framework for evaluating programs with grouped data. *Journal of Business & Economic Statistics*, 21(1), 1-11.
- Ding P. and Li F. (2018). Causal Inference: A Missing Data Perspective. *Statistical Science*, 33(2), 214-237.
- Fisher R. A. (1918). The Causes of Human Variability. Eugenics Review, Vol. 10: 213-220.
- Fisher R. A. (1925). *Statistical Methods for Research Workers*, 1st ed, Oliver and Boyd.
- Frangakis C. E., and Rubin D. B. (2002). Principal Stratification in Causal Inference. *Biometrics*, 58(1), 21-29.
- Holland P. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81, 945-970.

- Imbens G.W., Rubin D.B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.
- **Mealli F.**, Rubin D.B. (2015) Clarifying Missing at Random and Related Definitions, and Implications when Coupled with Exchangeability, *Biometrika*, 102 (4): 995-1000. (Correction in *Biometrika* (2016) 103 (2): 491).
- Meng X. L. (2018). Statistical Paradises and Paradoxes in Big Data (I) Law of Large Population, Big Data Paradox, and the 2016 US Presidential Election. The Annals of Applied Statistics, 12(2), 685-726.
- Neyman J. (1923). On the Application of Probability Theory to Agricultural Experiments. *Essay on Principles*, Section 9.
- Pearson K., Lee A., and Bramley-Moore L. (1899). VI. Mathematical Contributions to the Theory of Evolution.âVI. Genetic (Reproductive) Selection: Inheritance of Fertility in Man, and of Fecundity in Thoroughbred Racehorses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 192, 257-330.
- Rosenbaum P. and Rubin DB. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41â55.
- Rubin D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies. *Journal of Educational Psychology*, V66, 688-701.
- Rubin, D. B. (1975). Bayesian Inference for Causality: The Importance of Randomization. *Proceedings of the Social Statistics Section of the American Statistical Association*, 233-239.

- Rubin D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3), 581-592.
- Rubin D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics*, 2(1), 1-26.
- Rubin D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *Annals of Statistics*, 6, 34- 58.
- Rubin D. B. (1990a). Formal Modes of Statistical Inference for Causal Effects. *Journal of Statistical Planning and Inference*, 25, 279-292.
- Rubin D. B. (1990b). [On the Application of Probability Theory to Agricultural Experiments. Essay on principles. Section 9.] Comment: Neyman (1923) and Causal Inference in Experiments and Observational studies. *Statistical Science*, 5(4), 472-480.
- Rubin D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469), 322-331.
- Simpson E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society: Series B*, 13(2), 238-241.
- Yule, G. U. (1903). Notes on the Theory of Association of Attributes in Statistics. *Biometrika*, 2(2), 121-134.