Design and Causality

(A quick tour over the main ideas)

R. Bellio & N. Torelli

Spring 2018

University of Udine & University of Trieste

Experimental data & the two big ideas of design

Observational data & the problem of causality

Experimental data & the two big ideas of design

Aim of a statistical analysis (from DAAG book)

Statistical data analysis can greatly help in answering scientific questions, but does not stand alone, and it must be interpreted against a background of subject area knowledge.

The data available are not suitable for any possible question! Ideally, they should be collected (or, even, generated) after the aims of the analysis have been planned.

Usually, the aims include **scientific understanding** of some critical points (*is a certain treatment effective for a certain disease?*), or prediction of some key variables (*which is the price that house purchasers may be willing to pay for a certain area and house size?*)

Experimental vs observational data

A critical distinction is between what can be reached by an **experiment**, where data are obtained under controlled conditions, and what can be reached by an **observational study**, where there is no control over the data-generation process.

Well designed experiments give highly reliable results, whereas observational studies require a lot of care, and can even be misleading.

Statistics is traditionally more concerned with experimental data, but there are plenty of methods developed for observational data as well. It could be argued that the latest developments of statistics are mainly focused on observational data.

Both settings ought to be considered in the study of statistical methods.

Two big ideas for experimental data

As reported in the CS book

> *Statistical design theory is concerned with the design of surveys and experiments so as to obtain data that will be best able to answer the statistical questions of interest.*

The theory is very extensive, and here we just focus on the two main ideas: **randomization** and **design of experiments (DOE)**.

They are two cornerstones of practical usage of statistical methods.

Randomization

The first important idea is well explained by an example, taken from the CS book.

- Consider an experiment to test whether a new drug is effective at reducing blood pressure, compared to a standard treatment.
- Subjects in the study differ for several other variables, that may affect the blood pressure (such as *age*, *weight*, *sex*, ...). They are not of direct interest, so that they are called *confounding variables*.
- To know what the effect of the drug is, we must allow for the presence of the confounding variables, otherwise we may mismeasure the effect of the drug.
- One possibility is to include the confounding variables in a statistical model. This is useful, but perhaps we do not know *all* the possible confounders! (We shall return on this later).
- **Randomization** provides an answer: patients are **randomly allocated to drug type**, thus breaking any association between the treatment and the confounding variables.

Randomization: why it works

Randomization of experimental units (patients in the example) to experimental treatments (drug treatment) turns the effects of unmeasured confounder variables into *effects that can be modelled as random noise*, which is always allowed for in a statistical model.

In the example, the effect of the new drug can be simply assessed by means of a suitable two-sample test for independent samples.

Design of Experiments (DOE)

The theory of Design of Experiments takes the randomization concept, and raises it to another level

'Design of experiments' means something specific in the statistical literature, which is different from its more general use in science. The key notion is that there is an *intervention* applied to a number of *experimental units*; these interventions are conventionally called treatments. The treatments are usually assigned to experimental units using a randomization scheme, and randomization is taken to be a key element in the concept in the study of design of experiments. The goal is then to measure one or more responses of the units, usually with the goal of comparing the responses under the various treatments. Because the intervention is under the control of the experimenter, a designed experiment generally provides a stronger basis for making conclusions on how the treatment affects the response than an observational study.

Figure 1: From Nancy Reid (2008), *Some Aspects of Design of Experiments*, in Proceedings of PHYSTAT Workshop On Statistical Issues for LHC Physics

Classical applications of DOE

Experiment design has been traditionally employed in Agriculture, Industrial and Bio-Medical research, but in later years also for *computer experiments*

The original area of application was agriculture, and the main ideas behind design of experiments, including the very important notion of randomization, were developed by Fisher at the Rothamsted Agricultural Station, in the early years of the twentieth century. A typical agricultural example has as experimental units some plots of land, as treatments some type of intervention, such as amount of or type of fertilizer, and as primary response yield of a certain crop. The theory of design of experiments is widely used in industrial and technological settings, where the experimental units may be, for example, manufactured objects of some type, such as silicon wafers, the treatments would be various manufacturing settings, such as temperature of an oven, concentration of an etching acid, and so on, and the response would be some measure of the quality of the resulting object. In so-called *computer experiments*, the experimental units are simulation runs, of, for example, a very complex system such as used for climate modelling or epidemic modelling; the 'treatments' are settings for various systematic or forcing parameters, and the response is the output of the climate model or epidemic model. Principles of experimental design are also widely used in clinical trials, where the experimental units are often patients, the treatments are medical interventions, and the response is some measure of efficacy of the treatment.

Figure 2: From Nancy Reid (2008), *Some Aspects of Design of Experiments* in Proceedings of PHYSTAT Workshop On Statistical Issues for LHC Physics

A glimpse on a traditional technique

Assume that a certain response is influenced by 4 experimental factors. A 2^4 **factorial design** would assign the treatments in the following manner

run	A	B	C	D	response
1	-1	-1	-1	-1	$y_{(1)}$
2	-1	-1	-1	$+1$	y_d
3	-1	-1	$+1$	-1	y_c
4	-1	-1	$+1$	$+1$	y_{cd}
5	-1	$+1$	-1	-1	y_b
6	-1	$+1$	-1	$+1$	y_{bd}
7	-1	$+1$	$+1$	-1	y_{bc}
8	-1	$+1$	$+1$	$+1$	y_{bcd}
9	$+1$	-1	-1	-1	y_a
10	$+1$	-1	-1	$+1$	y_{ad}
11	$+1$	-1	$+1$	-1	y_{ac}
12	$+1$	-1	$+1$	$+1$	y_{acd}
13	$+1$	$+1$	-1	-1	y_{ab}
14	$+1$	$+1$	-1	$+1$	y_{abd}
15	$+1$	$+1$	$+1$	-1	y_{abc}
16	$+1$	$+1$	$+1$	$+1$	y_{abcd}

Statistical methods for DOE

The design of the previous slide is an example of a *full factorial design*, perhaps the simplest possible instance.

There are countless other types of design, some of which are very sophisticated, each suited to a different sort of investigation.

The statistical analysis of data obtained from experimental designs is usually carried out by **linear models** or **generalized linear models**, covered later on in this course.

Observational data & the problem of causality

The issue of observational data

The main issue of observational data is that they are not experimental !

We cannot randomize treatments to observations, hence we cannot rule out the effect of confounders.

We end up observing *correlations* between variables, without being able to giving them a *causal interpretation*. Indeed, *correlation is not causation* is a well-known fact in statistics, valid for observational data.

Correlation is not causation

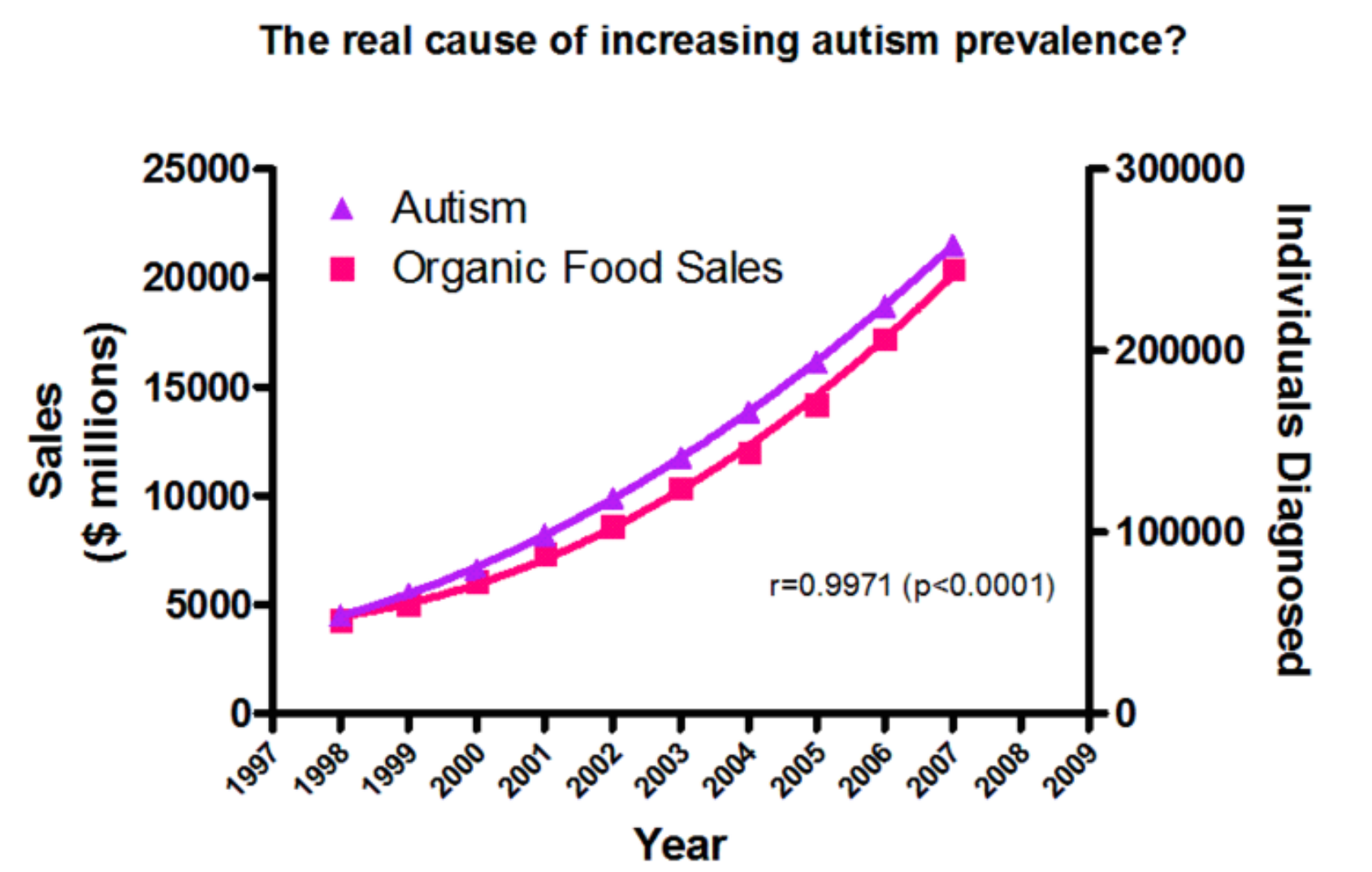

Figure 3: http://www.abc.net.au/science/articles/2013/04/29/3740590.htm?site=science_dev&; see also http://www.tylervigen.com/spurious-correlations

The issue of sample selection

The usage of non-random samples causes the so-called **sample selection bias**.

Indeed, when we say that a sample *does not represent the target population*, we typically mean that the sample was not selected randomly from the population of interest.

Lack of random selection may lead to inferential results
(in the form of parameter estimates or hypothesis testing) that may be misleading.

Sample selection bias may also arise with experimental data, whenever the experimental protocol is not properly followed.

We illustrate the issue by means of three examples.

Example 1: The Lanarkshire Milk Experiment

It took place in 1930, and it is an infamous example on *how not to run an experiment*. "Student" (William S. Gosset) readily spotted the pitfalls

THE LANARKSHIRE MILK EXPERIMENT.

By "STUDENT."

In the spring of 1930 * a nutritional experiment on a very large scale was carried out in the schools of Lanarkshire.

For four months 10,000 school children received $\frac{3}{4}$ pint of milk per day, 5000 of these got raw milk and 5,000 pasteurised milk, in both cases Grade A (Tuberculin tested); another 10,000 children were selected as controls and the whole 20,000 children were weighed and their height was measured at the beginning and end of the experiment.

It need hardly be said that to carry out an experiment of this magnitude successfully requires organisation of no mean order and the whole business of distribution of milk and of measurement of growth reflects great credit on all those concerned.

It may therefore seem ungracious to be wise after the event and to suggest that had the arrangement of the experiment been slightly different the results would have carried greater weight, but what follows is written not so much in criticism of what was done in 1930 as in the hope that in any further work full advantage may be taken of the light which may be thrown on the best methods of arrangement by the defects as well as by the merits of the Lanarkshire experiment.

Example 1: The Lanarkshire Milk Experiment

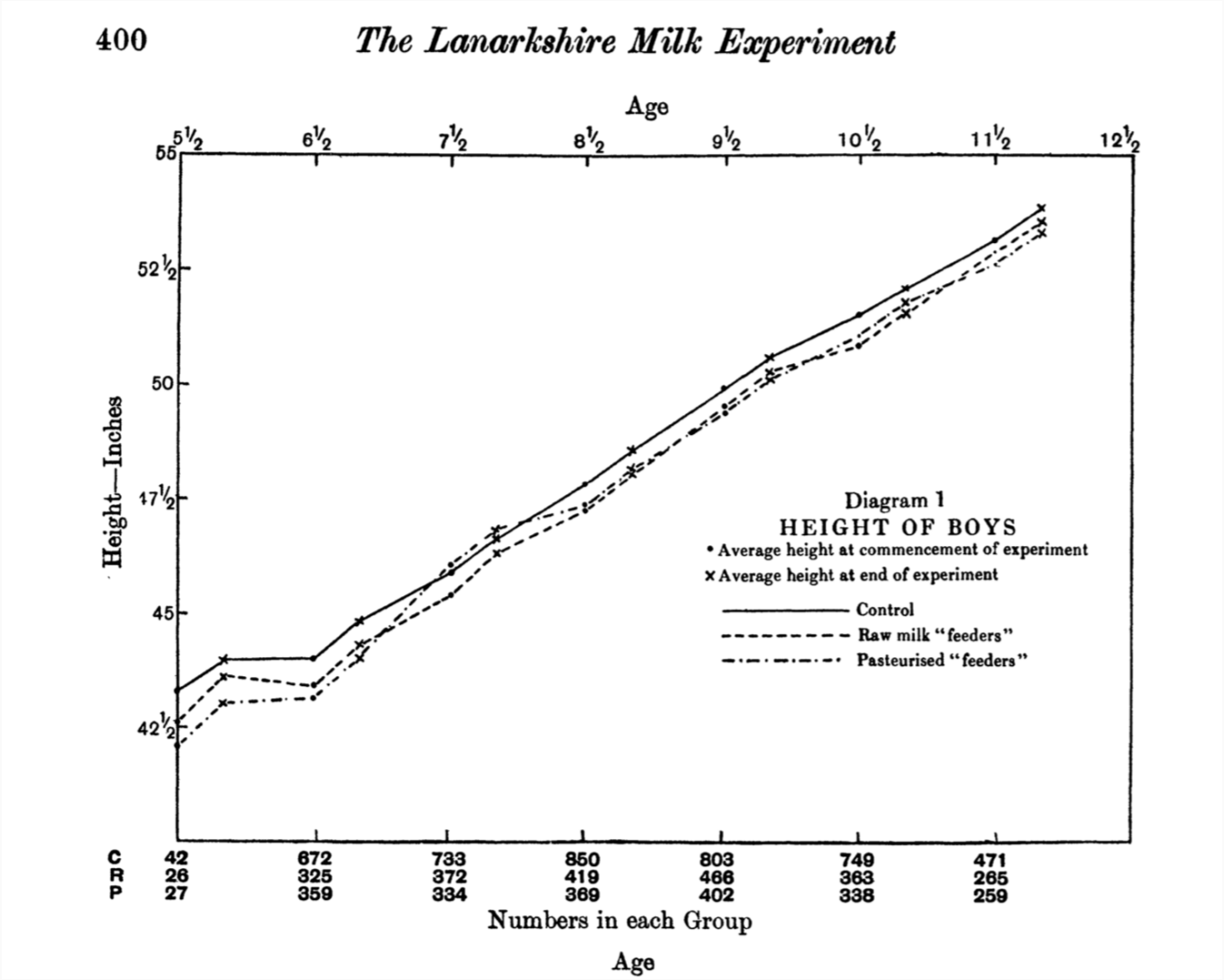

Figure 5: From Student (1931), Biometrika 23, pp.398-406

Example 1: The Lanarkshire Milk Experiment

Presumably this discrimination in height and weight was not made deliberately, but it would seem probable that the teachers, swayed by the very human feeling that the poorer children needed the milk more than the comparatively well to do, must have unconsciously made too large a substitution of the ill-nourished among the "feeders" and too few among the "controls" and that this unconscious selection affected, secondarily, both measurements.

Figure 6: From Student (1931), Biometrika 23, pp.398-406

Example 2: Bombers in WWII

During WWII, statistician Abraham Wald was asked to help the British decide where to add armor to their bombers. After analyzing the records, he recommended adding more armor to the places where there was no damage! (https://www.johndcook.com/blog/2008/01/21/selection-bias-and-bombers/)

(Btw, Abraham Wald is the same statistician we met for the Wald pivot)

Example 3: A casual glance at some recent news...

«E gli studenti italiani sono ancora bamboccioni» titola laRepubblica.it: «*Due ragazzi su tre durante il percorso universitario continuano a vivere con mamma e papà. In Europa uno su tre*», «*Studiano comodamente a casa, quasi sempre accuditi dai genitori, e raramente lavorano per pagarsi gli studi*». Che sia stato un caso o sia stata la valanga di tweet polemici, il giorno dopo il titolo è già cambiato e non contiene più l'epiteto denigrativo. Al di là delle polemiche, facciamo un esercizio di *fact checking* e vediamo cosa dice l'indagine AlmaLaurea: «*Le esperienze di lavoro hanno caratterizzato il 65% dei laureati triennali, il 58% dei magistrali a ciclo unico e il 67% dei magistrali biennali*». Insomma, gli studenti universitari che lavorano sono ben più di un terzo (anche se in calo a causa della crisi economica e della riduzione di studenti in età adulta). Ma i numeri di Repubblica da dove provenivano? Provengono dall'indagine Eurostudent. Se si controlla a pagina 23 del rapporto si vede che l'Italia è l'unica nazione che ha fatto ricorso esclusivamente a interviste telefoniche. Nel rapporto italiano (pag. 19) si legge che il campione è di 5.043 intervistati con un numero di contatti falliti pari a 3.958 e un numero di interviste non realizzate a causa di rifiuto pari a 2.629. L'indagine AlmaLaurea coinvolge il 90% di tutti i laureati degli atenei italiani, con un tasso di risposta dell'82% tra i laureati ad un anno. Sorge il dubbio che il dato del rapporto Eurostudent sia stato ripreso senza troppe verifiche proprio perché rinforzava il clichè dei bamboccioni "*comodamente a casa, quasi sempre accuditi dai genitori*". Ed è forse per la stessa ragione che non fa notizia AlmaLaurea quando scrive che «*Il 65% dei laureati del 2016 ha svolto un'esperienza di lavoro nel corso degli studi*». Se la realtà non conferma il cliché, allora non è reale.

Figure 7: From https://www.roars.it/online/secondo-la-repubblica-gli-studenti-italiani-sono-ancora-bamboccioni/

Causal inference for observational studies

Despite the lack of randomization may challenge the discovery of causal mechanisms, this does not mean that there are no solutions whatsoever.

Before mentioning some possible workarounds, we need to make our notation a bit more precise.

It is worth introducing the **fundamental problem of causal inference**, as introduced in the book by Gelman and Hill (2007, http://www.stat.columbia.edu/~gelman/arm/)

Back to a typical medical example

We aim at estimating the effect of a binary treatment T on an outcome y. For a given unit i, we define the *potential outcomes*, that would have been observed under control and treatment conditions:

- For $T_i = 0$ (control) we would get the outcome y_i^0
- For $T_i = 1$ (treatment) we would get the outcome y_i^1

Then the *treatment effect* for unit i is simply

$$\text{treatment effect for unit } i = y_i^1 - y_i^0$$

Averaging over the sample would then give the estimated effect.

Ideal experiments

In ideal experiments both outcomes are simultaneously observed, and then a simple **test for paired data** would readily allow for some causal conclusion.

The table below reports an instance

(Hypothetical) complete data:

Unit, i	Pre-treatment inputs X_i			Treatment indicator T_i	Potential outcomes y_i^0	y_i^1	Treatment effect $y_i^1 - y_i^0$
1	2	1	50	0	**69**	*75*	*6*
2	3	1	98	0	**111**	*108*	*−3*
3	2	2	80	1	*92*	**102**	*10*
4	3	1	98	1	*112*	**111**	*−1*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	4	1	104	1	*111*	**114**	*3*

Figure 8: From Gelman and Hill (2007), p. 171

Actual experiments

The problem is that only one of the two outcomes is available:

Observed data:

Unit, i	Pre-treatment inputs X_i			Treatment indicator T_i	Potential outcomes y_i^0	y_i^1	Treatment effect $y_i^1 - y_i^0$
1	2	1	50	0	69	?	?
2	3	1	98	0	111	?	?
3	2	2	80	1	?	102	?
4	3	1	98	1	?	111	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	4	1	104	1	?	114	?

Figure 9: From Gelman and Hill (2007), p. 171

The missingness of one of the two outcomes is the fundamental problem of causal inference.

Ways to get around the problem

There are essentially three ways to solve the problem:

1. *Close substitutes of ideal outcomes*. Sometimes we can get a way to split the unit in two parts (like for the `pair65` data) or to assign both treatments to the same unit at different times, ruling out any interference. Then we could obtain causal conclusion by means of a **test for paired data**.
2. *Randomization*. As already discussed, since we cannot compare treatment and control outcomes for the same units, we compare them on *similar units*. Similarity is attained by using randomization to decide which units are assigned to the treatment group and which units are assigned to the control group. We then obtain the causal conclusion by means of a **test for independent samples**.
3. *Statistical adjustments*. When neither 1. nor 2. are possible, some ingenuity is called for.

Statistical adjustments for causal inference with observationl studies

There are several methods to carry out these adjustments. Broadly speaking, and somewhat over-simplifying a large body of scientific literature, we can mention three ways.

1. We could identify all the possible confounders, or at least the most important, and include them as predictors (covariates) in a **regression model** jointly with the treatment effect. This is by far the most used methodology, common to all the disciplines that endorse statistical methods.
2. We could identify the most important confounders, and **stratify** the units according to their values in the sample, then average the estimated treatment effect over the different strata. This is usually done in epidemiological studies. It is a strongest version of 1., and requires both a precise theory and a large sample size. A variation of this methodology employs stratification by *propensity score*.

Make your theories elaborate

3. The third class of methods employs more sophisticated modelling of the relation between the observed outcome and the selection of the unit in the samples. They are more common in economics and the social sciences, but used also elsewhere.

Although they may seem a bit over-ambitious, these methods perfectly conform to a famous suggestion of Sir Ronald Fisher (perhaps the greatest statistician ever, the inventor of maximum likelihood estimation and among the first to introduce randomization):

> *When asked what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: "Make your theories elaborate".*

When constructing a causal hypothesis with observational studies, a very careful analysis of the problem is called for.