Difference-in-differences

Denis Fougère, Nicolas Jacquemet

SciencesPo
LABORATORY FOR INTERDISCIPLINARY EVALUATION OF PUBLIC POLICIES

LIEPP METHODS BRIEF n°10

2023

This brief is part of a set of methods briefs published by LIEPP. As such, it is supported by the ANR and the State under the "Investissements d'avenir" programme within the framework of the IdEx Université Paris Cité (ANR-18-IDEX-0001).

Difference-in-differences

Denis FOUGÈRE (Sciences Po, CRIS, CNRS, LIEPP, CEPR and IZA)
denis.fougere@sciencespo.fr

Nicolas JACQUEMET (Université Paris 1, CES, PSE)
nicolas.jacquemet@univ-paris1.fr

How to cite this publication:
FOUGERE, Denis, and JACQUEMET Nicolas, **Difference-in-differences**, *LIEPP Methods Brief* n°10, 2023-05

Publication initialement rédigée en français :
FOUGERE, Denis, et JACQUEMET Nicolas, **Méthode des doubles différences (difference-in-differences)**, *LIEPP Fiche méthodologique* n°9, 2023-05

IN A NUTSHELL:

The difference-in-differences method is a quantitative, quasi-experimental method to assess the impact of an intervention by setting up comparison groups and measuring the change in an outcome between a pre- and a post-intervention period when only one of the two groups has access to the intervention. This method is very useful for *ex-post* impact evaluation.

Keywords: Quantitative methods, quasi-experimental methods, difference-in-differences, difference-in-difference-in-differences, longitudinal data, parallel trends, entropy balancing, synthetic/artificial control group

I. How is this method useful for policy evaluation?

Although the evaluation of public policies covers a very broad set of issues and tools, which goes well beyond the mere quantification of their effects, the question of the effectiveness of policies implemented in the past is obviously of primary importance, as it constitutes a useful guide for considering their continuation, evolution, generalisation or even abandonment. Such an evaluation requires a clear definition of the objectives pursued. For example, the effect of raising the minimum retirement age on retirement frequency, the impact of setting up a university grant system on transitions to higher education, or the consequences of introducing financial aid to facilitate access to healthcare on the use of the healthcare system. A natural reflex, which appears (too) often in the public debate, is to compare the situation of people who have benefited from the interventions implemented with that of others who have not. In order to assess the effectiveness of an unemployment insurance reform offering personalised job search assistance, one could thus compare those who benefited from this assistance with those who did not. As the study by Fougère, Kamionka and Prieto (2010, see Figure 3) illustrates, such a comparison shows unambiguously that job search assistance programs lead to a much slower return to employment for those who have benefited from them. Does this mean that services offered are detrimental to the probability of finding a job for unemployed workers?

Of course not. An alternative interpretation is that people who are offered job search assistance are precisely those who have the greatest difficulty in finding a job. When comparing their situation to that of unemployed people who have not received assistance, the implicit assumption is that the return to employment observed in this category can serve as a reference (i.e., a *counterfactual*) to the situation that would have been experienced by the beneficiaries in the absence of the assistance program. However, the beneficiaries are precisely those whose situation would have been particularly difficult in the absence of the assistance program. To avoid such confusions, the difference-in-differences method consists of defining the comparison group in such a way that the observed difference provides a more convincing estimate of the intervention effect.

II. What does this method consist of?

Suppose we observe changes between two dates in an outcome variable (also called a response variable or dependent variable) in two distinct groups. The first of these groups, called the *treatment* group, benefits from a given intervention or policy (referred to as the treatment); the second, called the *control* group, does not. The policy is implemented between the two dates. The measurement of the intervention effect is based exclusively on the variation of the outcome variable between these two dates. This variation differs in the two groups, generally from the moment the treatment comes into effect. It is this inflection in the difference between the two groups that is interpreted here as the average effect of the treatment on the outcome variable.

Why is this procedure called the *difference-in-differences* method? The first difference is the difference between the average value of the outcome variable in the treatment group at the second date (after implementation of the policy to be evaluated) and the average value of the same variable in the same group at the initial date (before implementation of the policy to be evaluated). From this first difference, we then subtract the analogous difference for the control group. The difference-in-differences method therefore exploits the longitudinal dimension of the data (or pseudo-longitudinal, as the individuals belonging to each of the groups may not remain the same over time) in order to provide an *ex-post* evaluation of the public policy that has been implemented.

The ability of this method to measure the average effect of the intervention is not based on the hypothesis that the non-beneficiaries can serve as a reference group for the beneficiaries in the absence of the intervention, but only on the fact that in the absence of the intervention, the average evolution of the outcome variable for the individuals in the treated group would have been the same as that observed in the control group (this is called the *parallel trends* assumption). The validity of this assumption, which cannot be verified, can be supported by the fact that before the policy was implemented, the outcome variable evolved in the same way in both groups (this is called the *common pre-trend* assumption). In contrast to the previous assumption, this second assumption can be tested using data observed prior to the implementation of the intervention, provided that the pre-intervention observation period is long enough - for example, at least five observations in both groups prior to the implementation of the policy being evaluated (these observations are called *leads* in the academic literature). The parallel trends assumption is equivalent to assuming that the pre-existing gap between the two groups, which may be explained by the various factors leading to different levels of the outcome variable within these groups, would have remained the same in the absence of the intervention, so that the observed change in this gap can be interpreted as the average effect of the intervention.

This approach is therefore only valid if the intervention leaves the outcome variable in the control group unchanged (this is the so-called *SUTVA,* i.e., *Stable Unit Treatment Value Assumption*). Indeed, any indirect effect of the intervention on this group (if, for example, the difficulty of finding a job increases because the acceleration of the return to work in the treatment group increases the tension in the labour market) calls into question the parallel trends assumption. Similarly, the parallel trends assumption could be challenged if the treatment group anticipates a positive effect of the intervention, and subsequently reduces job search intensity -- a violation known as the *Ashenfelter gap*.

Given the many factors that can affect the validity of the approach, recent developments of the-difference-in-differences method aim in particular to refine the constitution of the groups in order to increase their comparability (see Roth *et al*., 2022, for a detailed description). It is for instance possible to use *matching* methods (see the brief dedicated to matching methods) which, based on a statistical criterion, associate each person benefiting from the intervention with the person or persons in the control group whose observable characteristics are close - so that the comparison is carried out between statistical *nearest neighbors* - or the entropy balancing method which permits to equalise the first moments (mean, variance, skewness, etc.) of the distributions of the covariates. A similar approach can be applied to the outcome variable rather than to the distribution of observable characteristics. This is the goal of the synthetic control method, which consists of creating an artificial control group from the observed control group by means of an appropriate system of weights. This synthetic control group is constructed in such a way that the past evolution of the outcome variable within this synthetic group is identical to that of the same variable in the treatment group. For that purpose, we minimise, by reweighting the observations in the control group, the distance between the outcome variable in the treatment group and this variable in the synthetic control group before the intervention. When the number of treated units is very large, it is possible that the synthetic control of a treated unit is not unique. Several recent contributions have proposed solutions

to this difficulty. Among these, some suggest the use of matrix completion techniques, others propose sampling-based inferential methods.

One of the most popular extensions which accounts for the existence of unobservable interactions between group and time characteristics that the difference-in-differences method might omit is the *difference in difference-in-differences* method. This method relies on the observation of two additional groups, a *fake* treatment group or a *fake* control group. For example, let us consider a health policy that is implemented in region *A* to people over 65. In order to evaluate the effects of this policy on the use of health care and on the health status of the persons concerned, it is possible to consider the persons aged 65 to 69 in region *A* as the treatment group, and to use the status of those aged 60 to 64 in this same region as the control group. A first difference-in-differences applied to these two groups should in principle produce an estimate of the average effect of the intervention on health care use and on the health status of people over 65 in region *A*. However, this approach can be criticised since it compares populations that are not quite the same in terms of their health status: people aged 68 or 69 are probably in poorer health than those aged 60 or 61, and therefore exposed to higher risks of health deterioration over time. To address this criticism, it is possible to consider the same age groups in a second region, say region *B*, where the same policy is not implemented, and then calculate a second difference-in-differences (DiD hereafter) estimate in region *B*. This second DiD estimate in region *B* can then be subtracted from that calculated in region *A*. The second DiD estimate applied to the two groups in region *B* eliminates the differences in health between age groups that naturally prevail in the population as a whole (the assumption of parallel trends is therefore weakened, and here concerns the relative difference between the two categories of population in each of the two regions).

In addition to the quality of the comparison between groups, a second limitation of the difference-in-differences method is that the effect of the intervention is not always identical within different subgroups of beneficiaries, or over time: then the effect of the intervention is said to be *heterogeneous*. By relying on the evolution of the gap between two groups only, this method only measures an average effect, which is only compatible with very large variations in the intervention effect between different subgroups. In order to study variations in the effect over time, it is useful to have observations of the outcome variable in both groups well beyond the date following the implementation of the intervention (such observations are sometimes called *lags*). This ensures that the policy being evaluated has significant effects in the medium term, or even in the long term if the statistical follow-up is long enough.

Such heterogeneity in the effects of the intervention also raises important difficulties when its diffusion in the treatment group is gradual. The usual method, which consists of integrating observations into the group of beneficiaries as they become eligible for the intervention, leads to unfounded conclusions (which can go so far as to conclude that an intervention with positive effects for all beneficiaries is ineffective). Recent studies recommend focusing only on observations that correspond to changes in treatment status, which implies to combine multiple difference-in-differences estimates calculated at all the dates at which the set of beneficiaries changes (see de Chaisemartin and d'Haultfoeuille, 2022, for a complete presentation).

III. An example of the use of this method in the field of employment

Like most labour market policies, the introduction of a minimum wage and the setting of its level is a delicate trade-off. When employers have a high bargaining power and can squeeze wages, the minimum wage provides some protection for employees and allows the benefits of production to be distributed more fairly. But the existence of a minimum wage also implies that all jobs that are less profitable than the minimum wage will not be offered on the market because they do not create enough value to cover the

cost of wages. The challenge is therefore to set a minimum wage that rebalances wage bargaining without excessively damaging economic efficiency.

One of the most famous studies of the implementation of the difference-in-differences method is Card and Krueger's (1994) paper on the New Jersey minimum wage increase in April 1992. In this study, Card and Krueger compare the level of employment in the fast-food industry (which is very intensive in low-skilled jobs that are usually paid at the minimum wage level) in New Jersey and Pennsylvania in February 1992 and November 1992. These dates frame an increase in the minimum hourly wage from US$4.25 to US$5.05 in April 1992 in New Jersey, while at the same time the minimum hourly wage remained constant at US$4.25 in Pennsylvania. Observing a change in employment in New Jersey between February and November 1992 by means of a first difference does not allow us to attribute this change to the increase in the minimum wage in that state alone, particularly because other concomitant factors, such as weather or macroeconomic conditions, could also explain this change. Furthermore, the difference in employment levels between the two states after the minimum wage was raised reflects not only the effect of the minimum wage policy but also the overall differences in the way the industry operates between New Jersey and Pennsylvania. By including both New Jersey (the treatment group) and Pennsylvania (the control group) fast food restaurants, located on both sides of the state border, Card and Krueger can limit the effects of these two types of factors by a second difference. Under the assumption of parallel trends, the change in fast-food employment in Pennsylvania can be interpreted as the change in fast-food employment in New Jersey that would have occurred if the minimum hourly wage had not increased in that state. Card and Krueger's estimates suggest that the minimum wage increase was not accompanied by a decrease in employment in New Jersey. Specifically, Card and Krueger estimate that the $0.80 increase in the hourly minimum wage in New Jersey resulted in (caused) an increase of 2.75 full-time jobs on average in each fast-food restaurant in this state.

IV. What are the criteria for judging the quality of the mobilisation of this method?

The estimator obtained will be more informative (and the hypothesis of parallel trends more credible) if the control group is similar to the treatment group in terms of observable explanatory characteristics (avoiding over-interpretation of such comparisons, since unobservable heterogeneity may vary considerably between groups without this being detectable). Unless the constitution of the groups follows a procedure that imposes such a condition, it is appropriate to ensure this by comparing the distribution of observable characteristics across subgroups (e.g., in a sample of employees, proportions of women, of different age groups, or the distribution of the levels of education) and then carrying out a set of statistical tests of the absence of significant differences between these subgroups (the procedure is known as a *balancing test*). A good practice is to condition the statistical analysis on any observable characteristic whose distribution varies across subgroups in order to take into account possible interactions between this characteristic and variations over time.

In order to check the robustness of the results, it is possible to use so-called *placebo* groups to replicate the analysis on a group of observations that has not been exposed to the intervention being evaluated. A first way to do this is to use a *fake* treatment group, which can be the same treatment group but observed at least two dates prior to the implementation of the public policy being evaluated, or a third group that is assumed to be unaffected by the policy being implemented. The robustness of the analysis is strengthened if this procedure leads to the conclusion that there is no effect. A second practice is to use another control group, whose observable characteristics are similar to those of the control group. In this case, the estimate of the average treatment effect should be approximately equal to that obtained with the original control group.

While the use of longitudinal data improves the quality of the comparisons that are made, it leads to working with observations that are potentially correlated over time. This drawback has long been neglected in the application of the difference-in-differences method, leading generally to an overestimation of the statistical significance of the estimated treatment effect. It is therefore crucial to take into account the correlation structure of the data in the statistical analysis (see Bertrand *et alii*, 2004).

V. What are the strengths and limitations of this method compared to others?

The difference-in-differences method is a quasi-experimental method, in the sense that it is primarily used to study changes that occur exogenously and in ways that are not directly related to the evaluation goals, but which produce observations that approximate an experimental situation. Like all quasi-experimental methods, the effects estimated with this method correspond to the effects of the policy on the sub-population that has benefited from this policy (in terms of the causal evaluation model of public policy, it measures the average treatment effect on the treated, hereafter ATT). Since the intervention has been deliberately targeted at particular categories of the population (who are particularly concerned by the intervention implemented, or who are particularly in need of it), this approach does not allow us to measure the average treatment effect (hereafter, ATE), i.e., the effect that it would produce if it were generalised to the whole population, or even the variations in the effect across different treated individuals. Athey and Imbens (2006) propose an alternative approach to the difference-in-differences method which provides an estimate of the entire counterfactual distribution of the outcome variable, and which produces a more refined measurement of variations in the effect of the intervention across different groups of beneficiaries.

Nevertheless, this method measures an average effect in a larger sub-population than most existing quasi-experimental methods. As such, it differs in particular from the regression discontinuity design (see the dedicated separate sheet) and the local average treatment effect (LATE) approach, which both only allow to estimate average treatment effects for some specific sub-populations. This is the case for the subgroup of people (known as compliers) whose access to treatment is solely due to their proximity to an exogenously fixed threshold (e.g., an age or income threshold) in the first case, and those who benefit from it because of an instrumental variable in the second case.

Some bibliographical references to go further

Athey, Susan, and Imbens, Guido. W. 2006. “Identification and Inference in Nonlinear Difference-in-Differences Models”, *Econometrica*, 74(2), 431–97. https://doi.org/10.1111/j.1468-0262.2006.00668.x

Bertrand, Marianne, and Duflo, Esther, and Mullainathan Sendhil. 2004. “How Much Should We Trust Differences-In-Differences Estimates?”, *Quarterly Journal of Economics*, 119(1), 249–275. https://doi.org/10.1162/003355304772839588

Card, David, and Krueger, Alan B. 1994. “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania”, *American Economic Review*, 84(4), 772-793. https://www.jstor.org/stable/2118030

De Chaisemartin, Clément, and D’Haultfoeuille, Xavier. 2022. “Difference-in-Differences Estimators of Intertemporal Treatment Effects”, NBER Working Paper No. 29873. https://doi.org/10.3386/w29873

Fougère, Denis, and Kamionka, Thierry, and Prieto, Ana. 2010. « L’efficacité des mesures d’accompagnement sur le retour à l’emploi », *Revue Economique*, 61(3), 599–612. http://dx.doi.org/10.3917/reco.613.0599

Roth, Jonathan, and Sant'Anna, Pedo H. C., and Bilinski Alyssa, and Poe John. 2022. “What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature”, arXiv:2201.01194, https://doi.org/10.48550/arXiv.2201.01194

Resources to implement this method with Stata and R softwares

Cunningham, Scott. 2021. *Causal Inference: The Mixtape.* Yale University Press: New Haven and London. Available in free access on the website https://mixtape.scunning.com/index.html

Huntington-Klein, Nick. 2022. *The Effect: An Introduction to Research Design and Causality*, Chapter 18. Chapman and Hall/CRC Press: Boca Raton, Florida. Available in free access on the website https://theeffectbook.net/ch-DifferenceinDifference.html