Experimental economics

Lecture 8: Experiment from a statistical perspective

Matej Lorko
matej.lorko@euba.sk
Materials: www.lorko.sk/lectures

References:

- Weimann, J., & Brosig-Koch, J. (2019). Methods in experimental economics. Springer International Publishing. Chicago
- Jacquemet, N., & l'Haridon, O. (2018). Experimental economics. Cambridge University Press.

The Experiment from a Statistical Perspective

- If a research question is to be answered experimentally and with the aid of statistical methods, the experiment must be designed in such a way that it answers this question as well as possible. “As well as possible” in means that the choice of method has been made in such a way that the formal method of analysis is appropriate to the statistical nature of the data generated so that they are compatible.

- Before we send subjects to the laboratory to generate raw data for us and before we commit ourselves to some statistical method of analysis, the research question, design of the experiment, the resulting raw data and the statistical analysis must be precisely matched with each other.

- A poorly designed experiment leads to a weak scientific result – even the most sophisticated method of analysis cannot change that. On the other hand, a well-founded analytical method can derive an even more significant scientific insight from a well-designed experiment.

The Experiment from a Statistical Perspective

- From a statistical point of view, the course of an experimental study should be divided into a design phase and an execution phase. The design phase, which is to be carried out first, consists of the following tasks and typical issues:

- Operationalizing the research question: What are the central constructs for which data must be collected during the experiment in order to answer the research question? Can these constructs be measured as variables? How should these variables be measured? Which of them is the dependent variable? Which of them are independent variables?

- Structuring the statistical design: Which variables are to be manipulated in which way by the experimenter (choice of treatments)? Which variables can I control and how can an undesired variation of the dependent variable be minimized? What is the observational unit and what is the experimental unit? How should a sample of subjects be selected? How many subjects do I need to show correctly that a certain effect “exists” with a given probability? What groups of subjects should be formed and what method should be used to form these groups? Will variables be measured on several levels (e.g. within-subject and between-subject)? How frequently and when should each subject’s variable be measured? Which are the qualitative variables and which are the quantitative variables?

- Translating the research question into a statistical hypothesis or a statistical model: What formal relationship could exist between the observed variation of the dependent variable and the variation of the independent variables? Which are the fixed-effect variables and which are the random-effect variables?

Choosing suitable statistical methods of analysis

- What is the purpose of my statistical analysis: To provide a descriptive presentation of the data and the treatment effects? To make a statistical conclusion concerning the population from which the sample is drawn (inference)? To make a prediction based on an estimated model? What are the main statistical characteristics of the experimental design or the resulting data (answers from previous questions)? What analytical methods can be used in view of the main statistical characteristics?

- Computer-assisted processing of the data: Are there missing values? Multiple measurements: long format vs. wide format; Conversion of the data into the format of the statistics software; Are there outliers? Are there subjects who have obviously made arbitrary decisions? What are short yet understandable variable names?

- Creating new variables from (a combination of) already collected variables (e.g. group averages); Creating a list of variables with descriptions.

- Computer-assisted analysis of the data: Describing the data using key indicators; Graphical representation of the data; Fitting the statistical model to the data by estimating the model parameters; Model diagnostics; Making inferences; Predictions using the estimated model.

- Conclusions: Can the treatment effects be verified statistically? Can the model explain the observed data well? Are further experimental treatments necessary?

Types of Variables

- In order to test a research idea experimentally, it is necessary to generate different types of variables. For example, suppose that the research hypothesis is that “the amounts offered in the ultimatum game are lower if the first mover is playing against a computer instead of a human being (and he or she knows this)”. In this case, the dependent variable is the amount offered by the first mover.

- An independent variable is expected by the experimenter to have an influence on the dependent variable, but not vice versa. In accordance with our research hypothesis, we expect the binary variable “computer opponent” (yes/no) to have an impact on the amounts offered. In a controlled experiment, the values of these independent variables are set systematically rather than simply being observed by the experimenter. In the above example, the experimenter measures the dependent variable “amounts offered” once under the value “yes” and once under the value “no” so that a comparison of both conditions is possible and the research hypothesis can be tested. In this case, the independent variable is also called a treatment variable, because its values represent the “treatments” or comparison conditions of the experiment under which the dependent variable is observed.

- Some further points need to be considered if the study is to draw a causal conclusion about the dependent and independent variables (and this is the main purpose of controlled experiments). If we observe a difference in the amounts offered once under each of the conditions “computer opponent – yes” and “computer opponent – no”, we must be able to rule out that this difference was caused by other influences - counfounding variables. Confounding variables blur the causality between dependent and independent variables because they have a “hidden” influence on the dependent variable that is not explicitly part of the experiment.

- Unfortunately, there are also confounding variables that cannot be controlled for. These are mainly such factors that make up the individual personality of a subject. Examples are intelligence quotient, income of parents, allergies, education, political sentiments, spatial ability, physical fitness and many more. Of course, not all possible uncontrollable variables are relevant to our own experiment, since many have no connection whatsoever to our dependent variable. Nonetheless, we would be well advised to carefully consider what, on the one hand, has a high probability of influencing our dependent variable and, on the other hand, can vary from subject to subject while at the same time remaining beyond our control.

Control, Randomization and Sample Size

- Regardless of whether or not an uncontrolled confounding variable is measurable, its impact on our dependent variable should be removed from the experiment as far as possible; otherwise a clear causal conclusion with respect to our treatment variable is no longer possible. A 100% control of such variables is hardly possible since many of them are not only not measurable, but also unknown and their influence is therefore “hidden”.

- Nevertheless, there is a simple statistical trick that can mitigate their impact. The basic idea is to form two groups of subjects across which the possible confounding factors are distributed as evenly as possible. This is done by randomly assigning each subject to one of the groups (randomization). In the process, it should be ensured that the groups consist of a sufficiently large number of independent subjects.

- All in all, in a laboratory experiment, the central variable is the dependent variable. Changes in this variable are due to the influence of explanatory variables and various confounding factors. If the observed change in the dependent variable is to be attributed to a change in the explanatory variable induced by the experimenter, the three most important concepts to be considered are:

  - Control (all the unwanted influences that can be kept constant should be kept constant);

  - Randomization (create comparison groups that are homogeneous on average by leaving it to chance which subject is placed into which group);

  - Sample size (or replication) - ensure a sufficient number of independent observations in a treatment, i.e. sufficiently large groups of subjects who do not systematically exhibit the same behavior.

Random Variables and Their Distribution

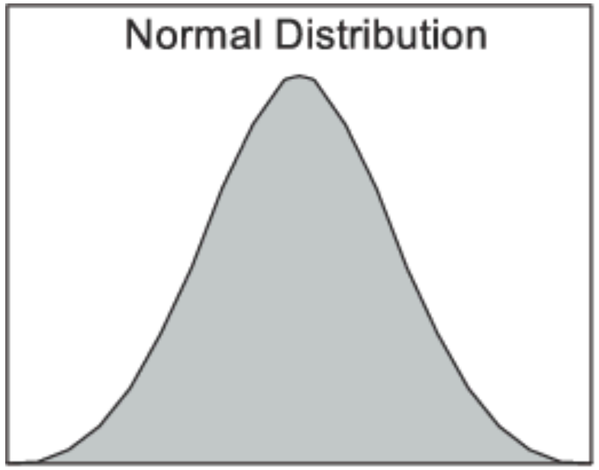
- In the statistical modeling of the relationship between variables, the dependent variable is interpreted as a random variable. Which values of a random variable are most likely and which are less likely is determined by their distribution. The so-called density function of a discrete random variable indicates the probability with which a certain value occurs. The outcomes of rolling a dice, for example, are evenly distributed, each with a respective probability of 1/6.

- In the case of a continuous random variable, such as the time it took the subject to make his decision, the probability of an individual value cannot be specified. If an infinite number of values exist, the probability of a single value must be infinitely close to zero. For this reason, with continuous variables, it is only possible to indicate specific probabilities for ranges of values, with the total area below the density function always being 1. The cumulative (continuous) distribution function is, mathematically speaking, the integral of the continuous density function. The value of the function at a point x thus indicates the probability with which the random variable assumes a value less than or equal to x.

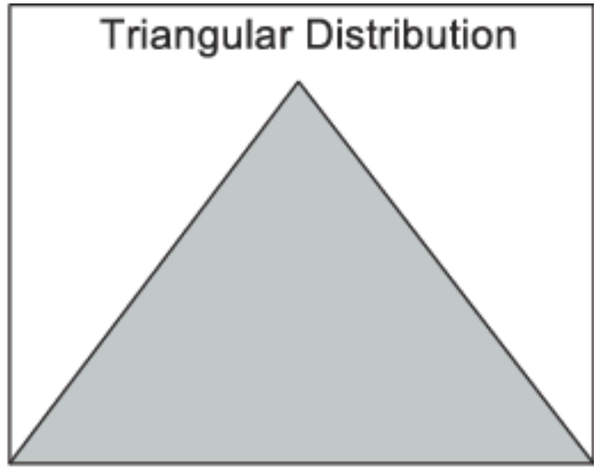
- Most statistical distributions have certain parameters which, depending on the value they have been set to, determine the shape of the density function. The three most important parameters are expected value, variance and degree of freedom. The expected value is the average of all the values drawn, if we (theoretically) draw a random sample infinitely often under the given distribution. For example, since there is an equal probability of rolling each number on a (normal) dice, the expected value is $1/6 \cdot (1 + 2 + 3 + 4 + 5 + 6) = 3.5$. The expected value of a distribution is a location parameter that provides information about where the theoretical mean value is located on the number line. The variance is the mean square deviation of all the realizations of the expected value and thus represents information about the dispersion of the random variable. The greater the variance, the wider and flatter the density function.

- The mother of all distributions is the normal distribution. Its parameters are the expected value μ and variance $\sigma 2$. The probability density is bell-shaped and symmetrical around μ, where it has the highest density function value. Other important distributions are not parameterized directly using expected value and variance, but indirectly using what is termed degrees of freedom, which influence the expected value and/or variance. The (Student's) t-distribution, for example, has such degrees of freedom, with the shape of its density function more and more closely approximating that of the density function of the standard normal distribution with increasing degrees of freedom.

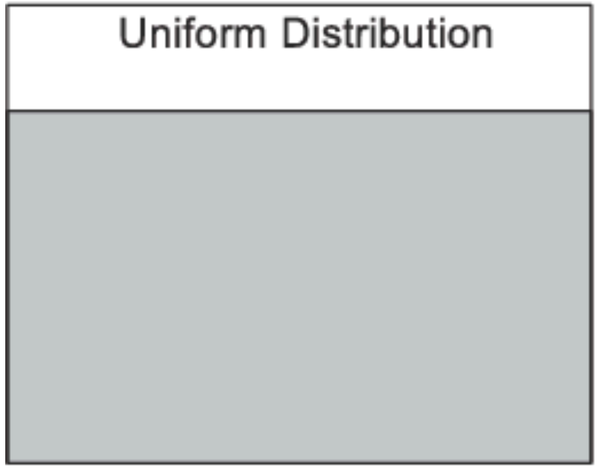
Random Variables and Their Distribution



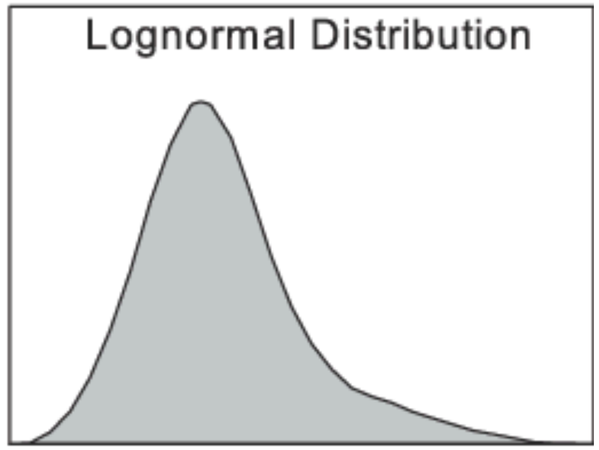
Normal distribution



Triangular distribution



Uniform distribution



Log-normal distribution

Creating the Statistical Design

- Compiling the Observation Units

- Selecting a certain number of subjects from a total population is referred to as sampling in statistics. Some thought needs to be given to the sample size, i.e. the question “How many subjects do I draw from the specified population?” Unfortunately, in experimental practice this question is often answered solely on the basis of the budget, true to the motto: “We simply take as many subjects as we can pay for, regardless of whether this number is large or small enough”. In the neurosciences, for instance, laboratory times are extremely expensive, so that sample sizes are (often have to be) in the single-digit range.

- However, such small samples are problematic, especially from the point of view of inferential statistics. The probability that a statistical hypothesis test correctly identifies an actual effect as present (this is called the power of a test) decreases drastically with smaller samples. In other words, even if in reality there is a relatively strong and scientifically relevant effect in the population, it will at best be recognizable as a “random artifact” and not as a statistically significant effect. On the other hand, there is also a “too large” in terms of sample size, since having samples that are too large can make statistical hypothesis tests too sensitive. This means that even the smallest, possibly scientifically insignificant effects become statistically significant.

- It is thus already clear that statistical significance should not be confused with scientific significance. Depending on the sample size, both can be completely different. This is because statistical significance is strongly influenced by the sample size, whereas the true effect to be detected in a population is not.

- If it is clear that a (sufficiently large) random sample is not affordable and a representative sample is still required, then stratified sampling is a good possibility. The population is first divided into subpopulations (strata), with the subjects within each subpopulation having at least one common characteristic that distinguishes them from the subjects of the other subpopulations. A random sample is then drawn from each stratum. Each of these samples must make up the same proportion of the total of all samples as each stratum in the total population.

Creating the Statistical Design

- How Do Experimental Treatments Differ?

- It is possible to classify experimental treatments according to the number of factor variables and their type as well as the number of possible values. In a single factorial design, only a single variable is changed. If this is a binary variable with just two values, or levels, we speak of a 1 × 2 factorial design. 1 × 2 factorial designs can be evaluated particularly easily since only the mean values of the dependent variables are usually compared under these two treatment conditions. Ideally, this difference is due to the treatment itself and is therefore called the (simple) treatment effect. The quantitative difference between the two values is called the size of the treatment effect or the (unstandardized) effect size. If, on the other hand, the factor variable has more than two levels, the treatment is called multilevel factorial design. In this case, the mean values of the dependent variable can be compared pairwise for every two levels or simultaneously for all levels.

- A design with two factors is considerably more complex than a single factorial design. For example, if we want to experimentally investigate how the factors “games against the computer” (Comp: no/ yes or 0/1) and “the experimenter knows who I am” (Anon: no/yes or 0/1) affect the giving behavior in a dictator game, then this hypothetical 2 × 2 factorial design.

- In the repeated measures design, each subject undergoes several measurements, either in one and the same treatment at different times (longitudinal design) or in different treatments, naturally also at different times (cross-over design). The sequence of treatments a subject goes through is again randomized. In each case, multiple measurements generate a within-subject structure with several observations for each subject.

- The main statistical problem with multiple measurements is the interdependence of the observations. In a 1 × 2 factorial design with multiple measurements, we get a control group (measured at level 1) and a treatment group (measured at level 2), which are related. Thus, the effect measured using the dependent variable can no longer be clearly attributed to the treatment, since it could just as easily be a time or sequence effect (e.g. learning, familiarization, fatigue). Counterbalancing the order (balancing) often comes to our aid in this case, i.e. two homogeneous groups are formed and one group is measured in the order level 1, then level 2 and the other in the order level 2, then level 1.

- The advantages of repeated measurements are lower costs due to fewer subjects, lower error spread, thus resulting in higher statistical power than comparable between-subject designs, and the possibility of measuring treatments over time (dynamics). The disadvantages of such a design are that it involves considerably more complex methods of analysis due to the dependency of the observations and weaker causalities owing to sequence, time and carry-over effects.

Describing your data

- The best way to learn about writing a data section is to read several data sections in the literature on your topic and pay attention to the kinds of information they contain. Your data section should do at least the following.

- Identify the data source. This means a sentence that explicitly says where your data come from.

- Describe the data source. You should tell your readers such things as the number of observations, the population groups sampled, the time period during which the data were collected, the method of data collection, etc.

- State the strengths and weaknesses of the data source. How do your data compare with other data sources used in the literature? Does yours provide more observations, and/or more recent observations, than other sources? Was the data collected in a more reliable manner? Why is the data source particularly suited (or not) to your study? Note any features of the data that may affect your results. Were certain populations overrepresented or underrepresented? Is there attrition bias or selection bias? Did the method of data collection change?

- Explain any computations or adjustments you made. Sometimes, a data source does not give you something directly; you perhaps had to add/subtract/multiply/divide two given pieces of data to get a third. Describe how you constructed your sample. Did you have to eliminate certain kinds of observations, for instance?

Descriptive statistics

- Data sections often contain a table of descriptive statistics, statistics of relevance about the sample. These statistics usually include the mean (e.g., mean income, mean age, mean years of schooling, etc.) and standard deviation. For categorical data (like race), however, you do not report a mean; instead, you report the percentage of the observations in each group.

- Expected value - The mean or average value of a sample statistic based on repeated samples from a population.

- Standard errors - The standard deviation or measure of variability or dispersion of a sampling distribution. The larger the sample, the smaller the standard error.

- Sampling distributions - A theoretical (non-observed) distribution of sample statistics calculated on samples of size N that, if known, permits the calculation of confidence intervals and the test of statistical hypotheses.

- NOTE: The mean and standard deviation work well for normal (bell curve shaped) distribution. If dealing with other distributions, it may be more useful to use median or mode to describe central tendency (expected value).

Plotting your data

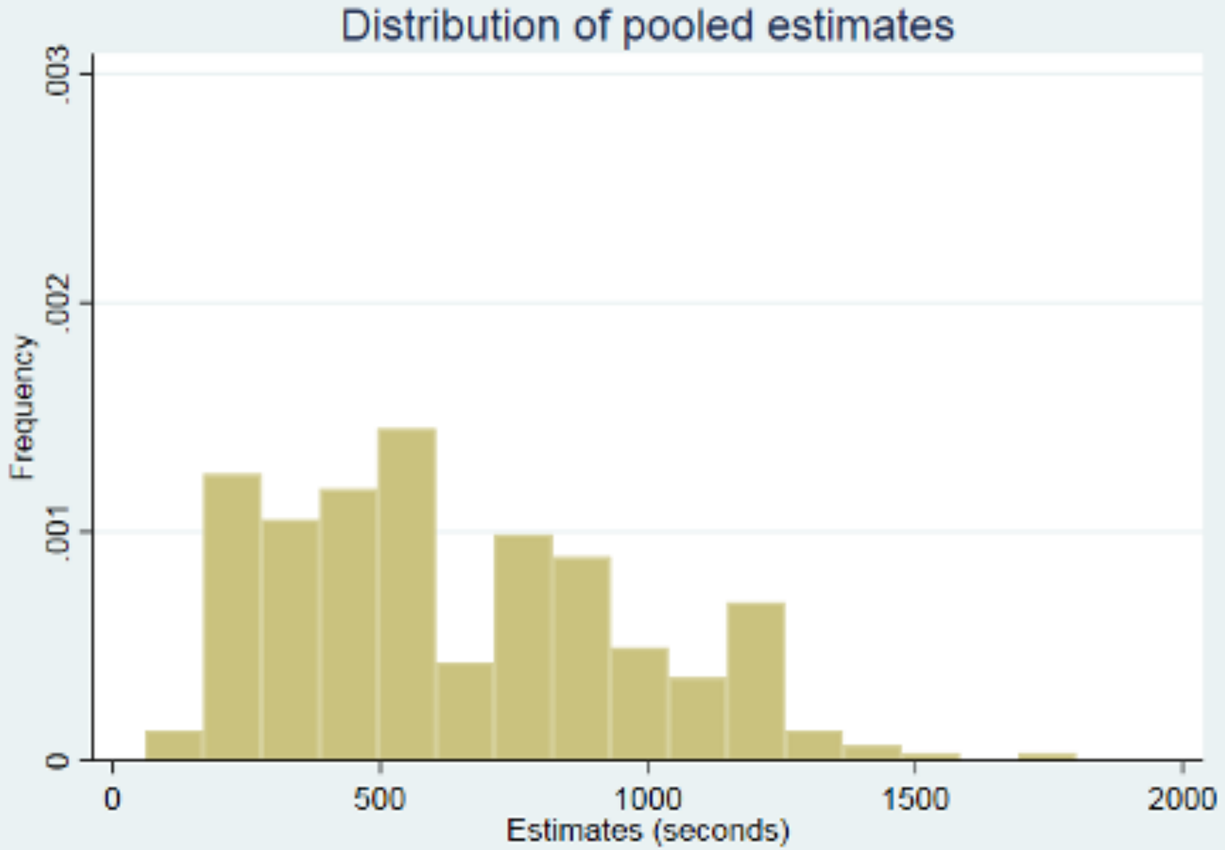
- A well-constructed graph can answer several questions at one time:
- Central tendency: Where does the center of the distribution lie?
- Dispersion or variation: How spread out or bunched up are the observations?
- The shape of the distribution: Does it have a single peak (one concentration of observations within a relatively narrow range of values) or more than one?
- Tails: Approximately what proportion of observations is in the ends of the distribution or in its tails?
- Symmetry or asymmetry (also called skewness): Do observations tend to pile up at one end of the measurement scale, with relatively few observations at the other end? Or does each end have roughly the same number of observations?
- Outliers: Are there values that,compared with most, seem very large or very small?
- Comparison: How does one distribution compare to another in terms of shape, spread, and central tendency?
- Relationships: Do values of one variable seem related to those of another?

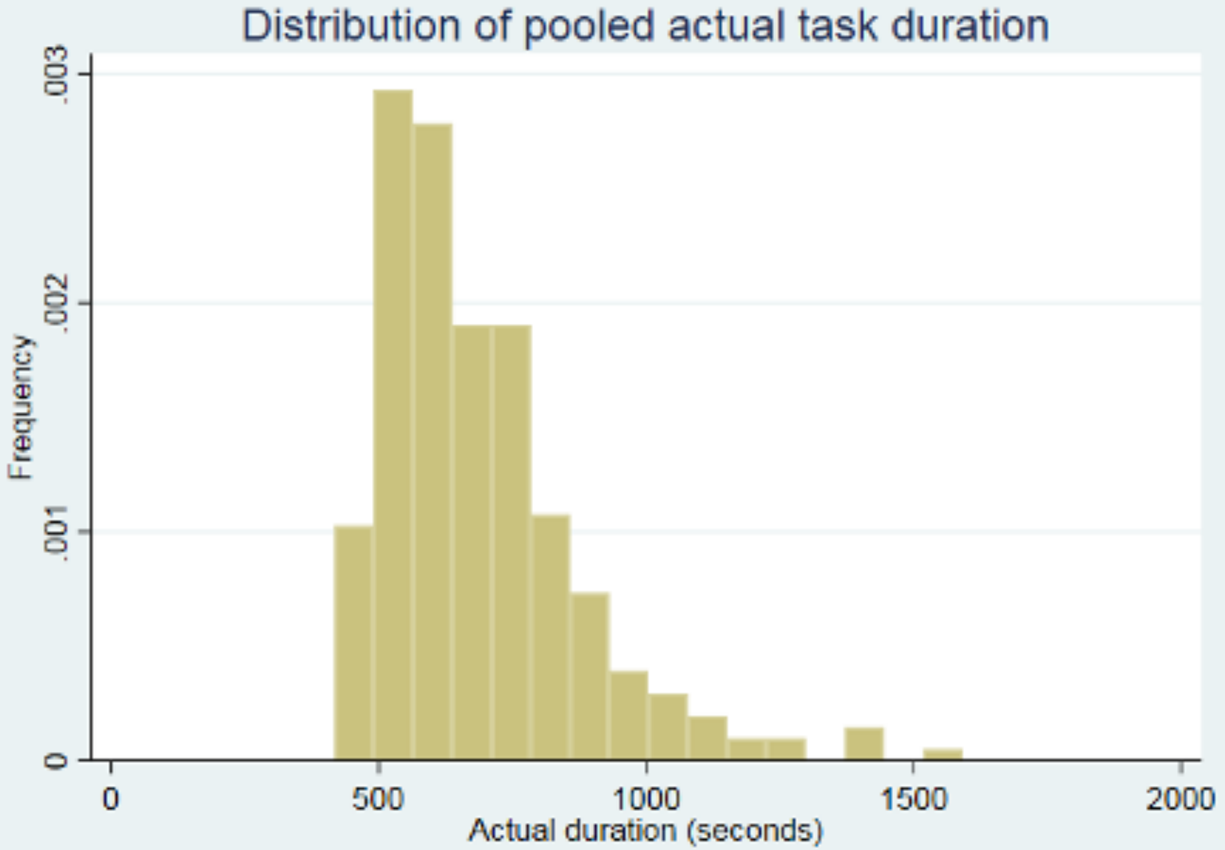
Choosing the right chart

TABLE 11-10 Typical Presentation and Exploratory Graphs

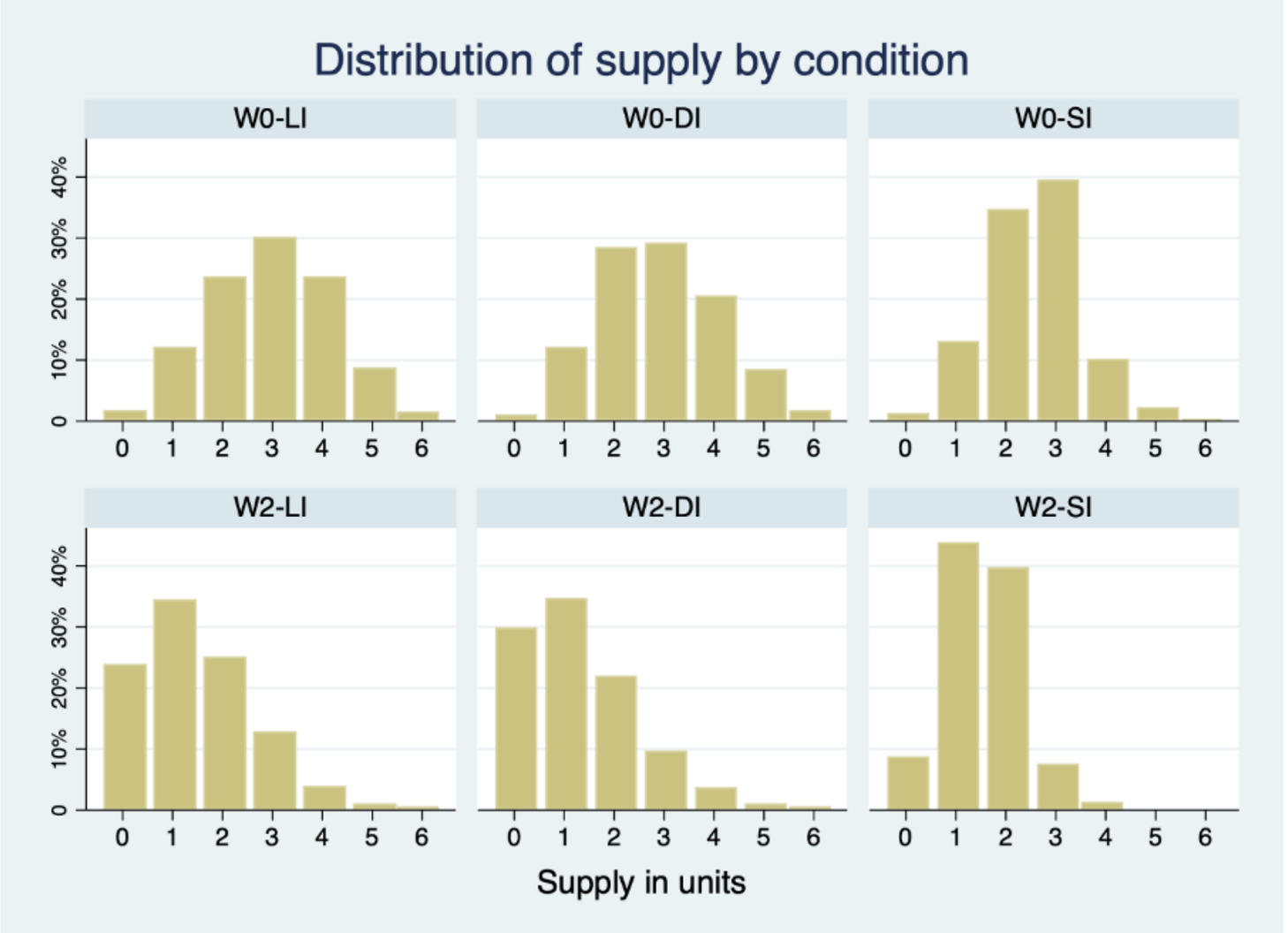
Type of Graph	What Is Displayed	Most Appropriate Level of Measurement	Number of Cases	Comments
Bar chart	Relative frequencies (percentages, proportions)	Categorical (nominal, ordinal)	3–10 categories	Common presentation graphic
Dot chart	Frequencies, distribution shape, outliers	Quantitative (interval, ratio)	*Less than* 50 cases	Displays actual data values
Histogram	Distribution shape	Quantitative	$N > 50$ cases	Essential exploratory graph for interval or ratio variables with a large number of cases
Boxplot	Distribution shape, summary statistics, outliers	Quantitative	$N > 50$ cases	Can display several distributions; actual data points, an essential exploratory tool
Time series plot	Trends	Quantitative (percentages, rates)	$10 < N < 100$	Common in presentation and exploratory graphics

Histogram

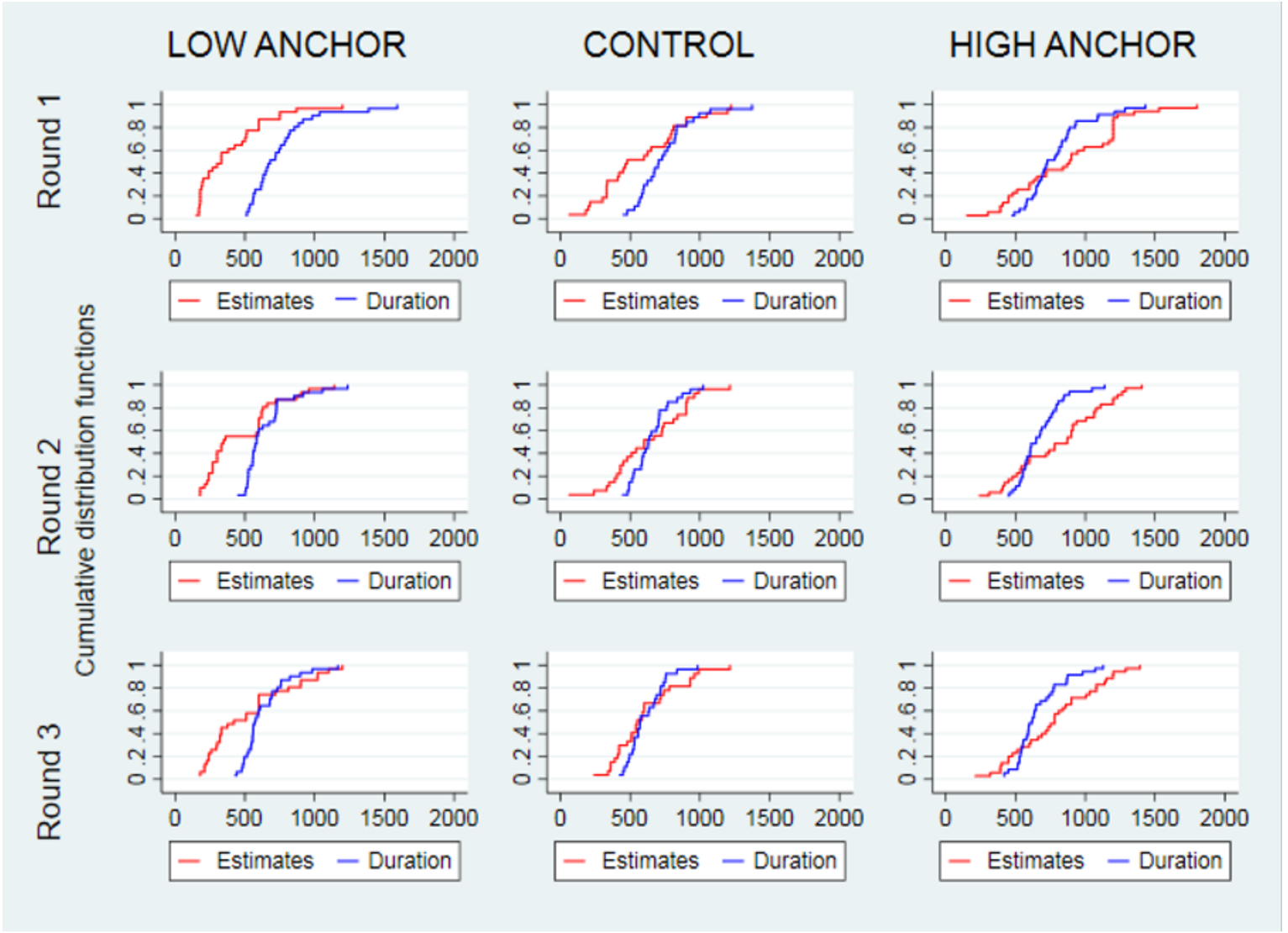

Distribution of pooled estimates
Frequency
0
.001
.002
.003
0
500
1000
1500
2000
Estimates (seconds)


Distribution of pooled actual task duration
Frequency
0
.001
.002
.003
0
500
1000
1500
2000
Actual duration (seconds)

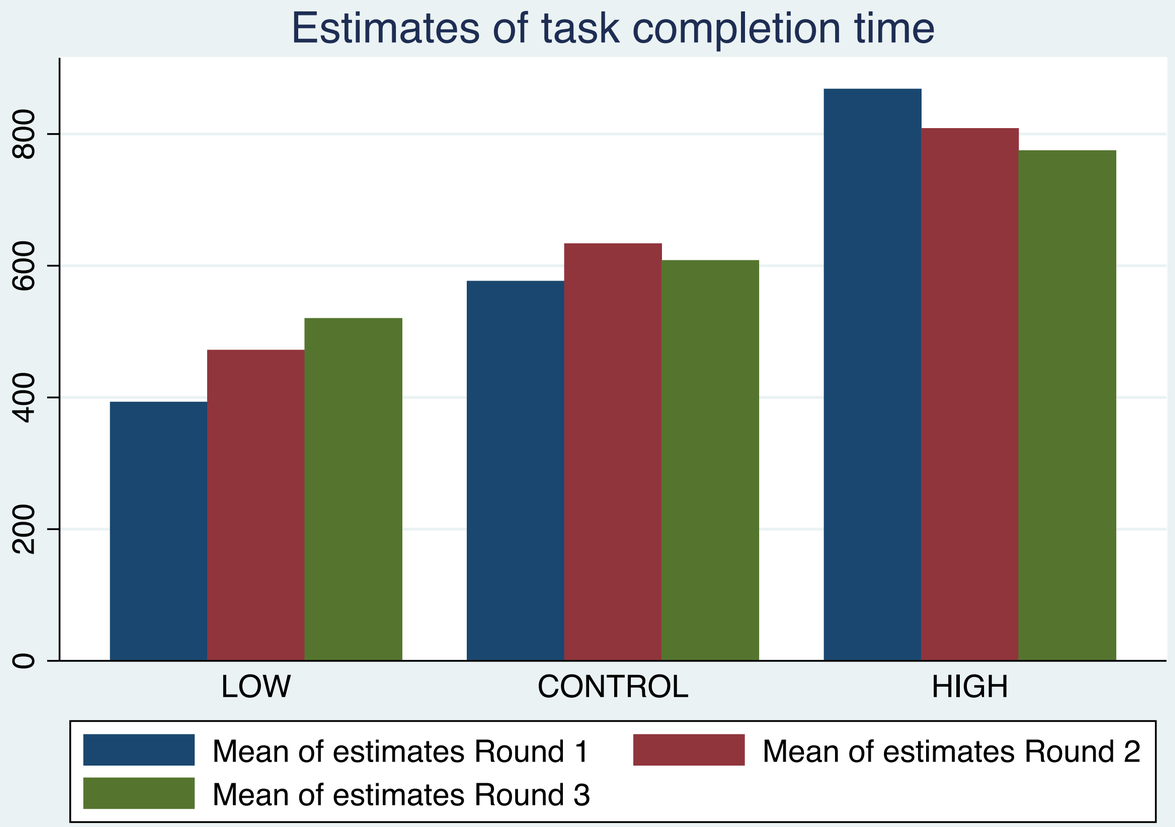
Histogram

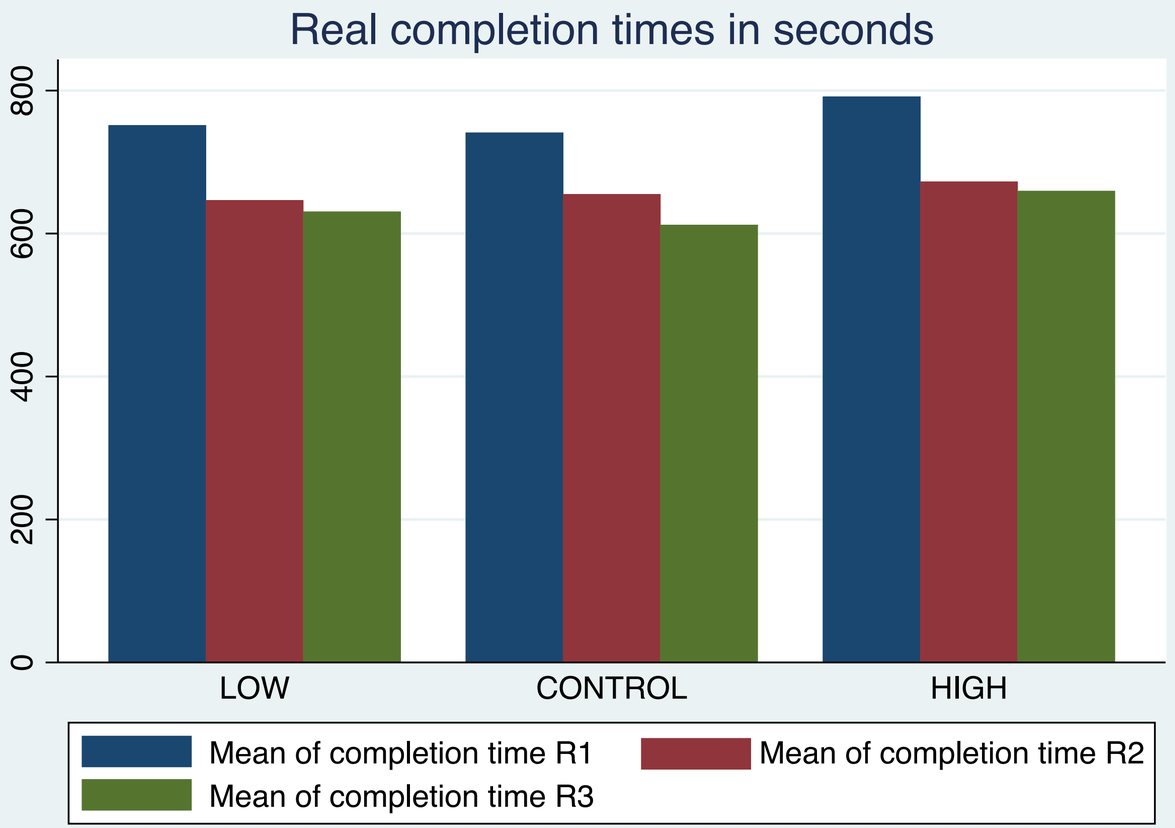

Distribution of supply by condition
W0-LI
W0-DI
W0-SI
W2-LI
W2-DI
W2-SI
0
10%
20%
30%
40%
0 1 2 3 4 5 6
Supply in units

Cummulative distribution

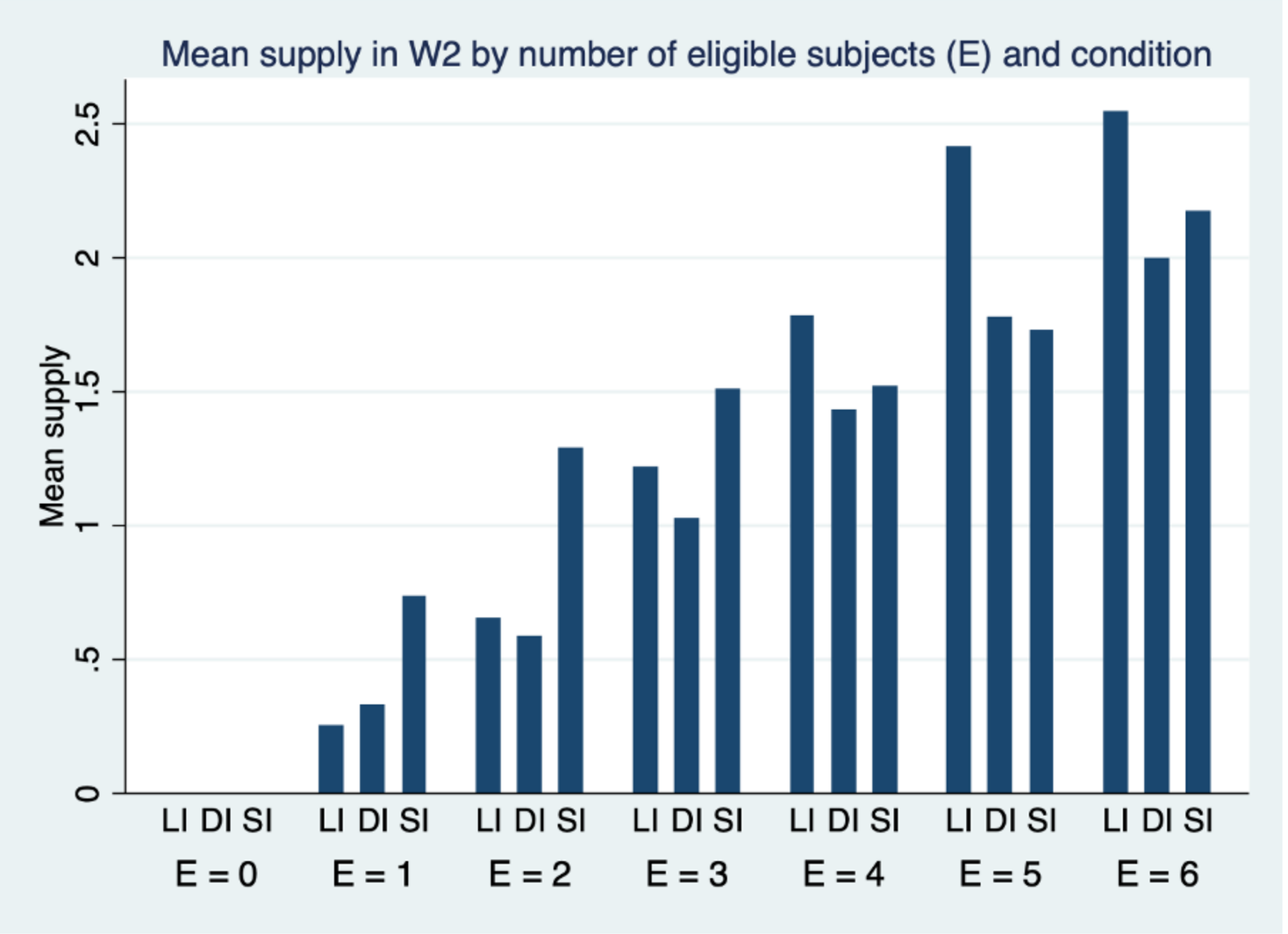


Bar chart

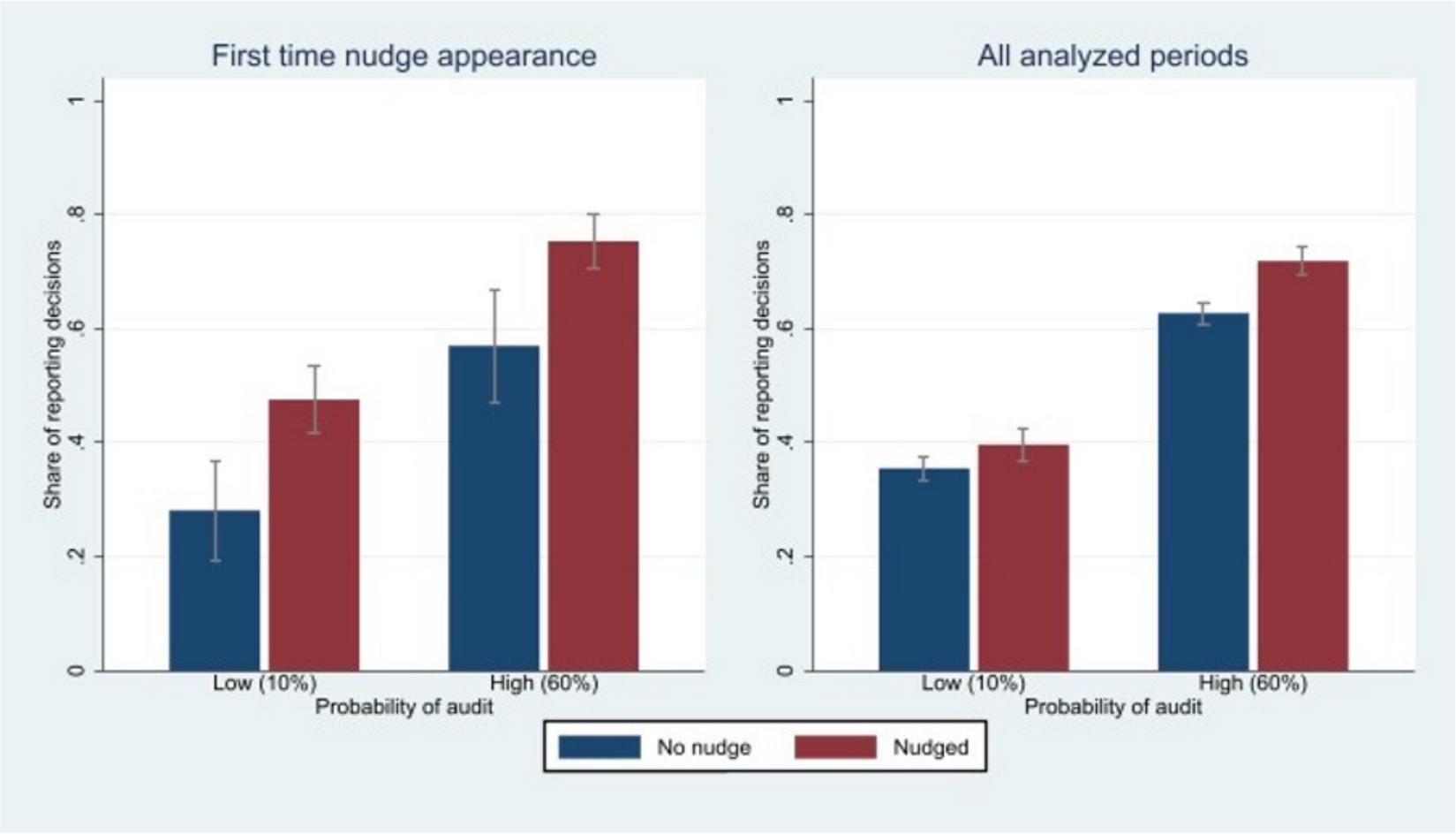

Estimates of task completion time
0
200
400
600
800
LOW
CONTROL
HIGH
Mean of estimates Round 1
Mean of estimates Round 2
Mean of estimates Round 3


Real completion times in seconds
0
200
400
600
800
LOW
CONTROL
HIGH
Mean of completion time R1
Mean of completion time R2
Mean of completion time R3

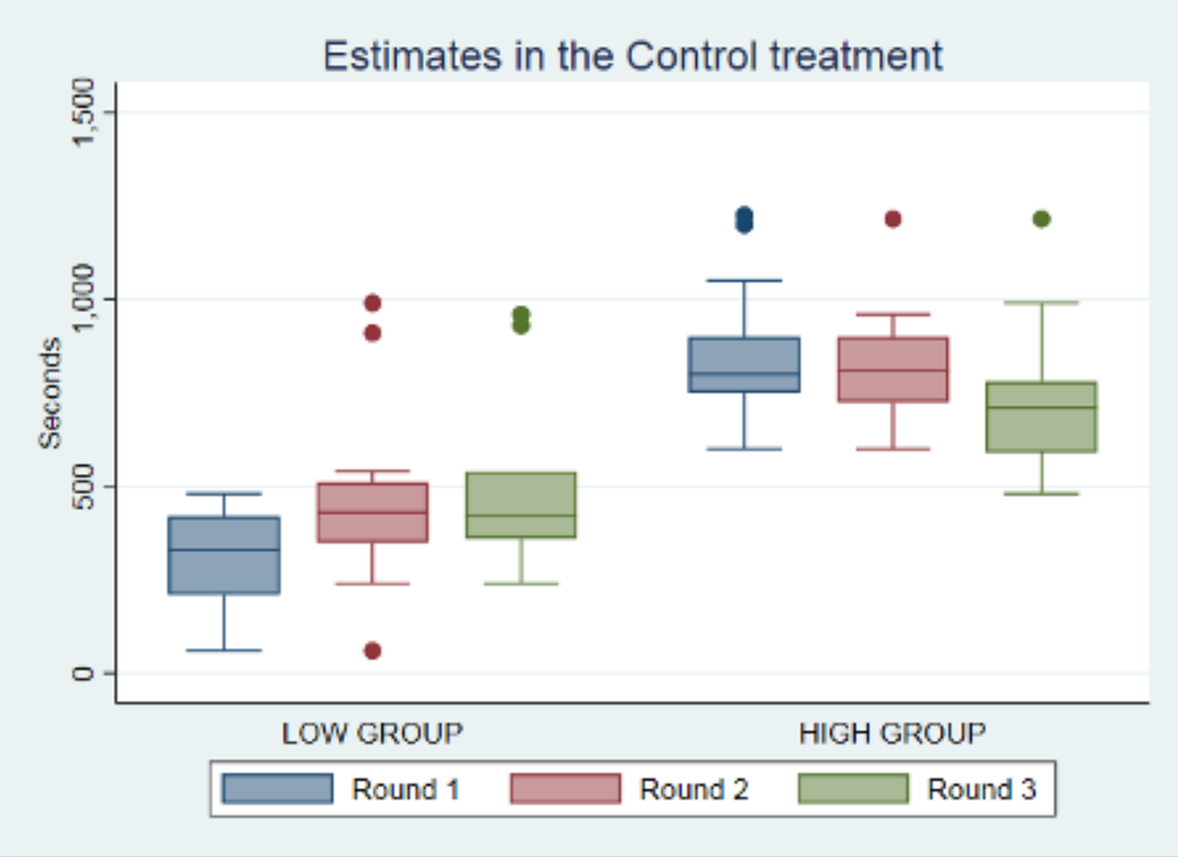
Bar chart

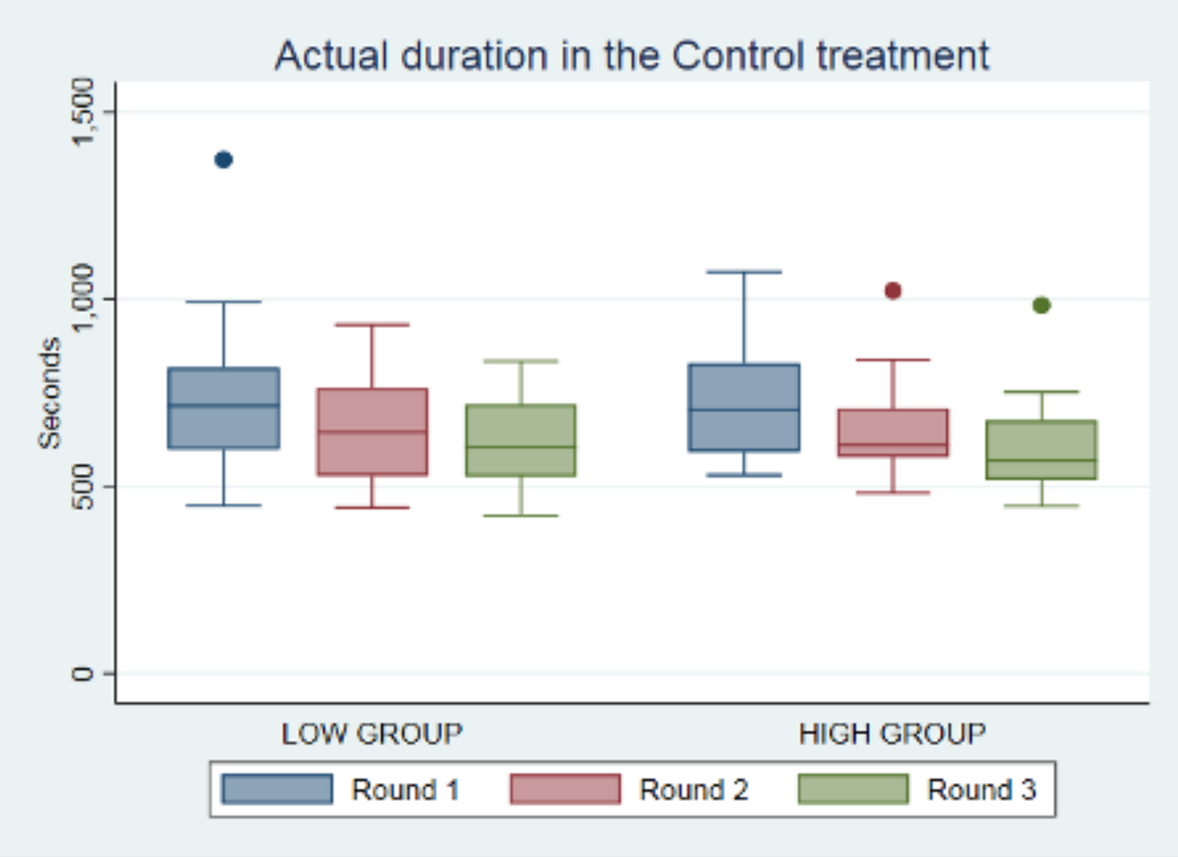

Mean supply in W2 by number of eligible subjects (E) and condition
Mean supply
0
.5
1
1.5
2
2.5
LI DI SI
E = 0
E = 1
E = 2
E = 3
E = 4
E = 5
E = 6

Bar chart

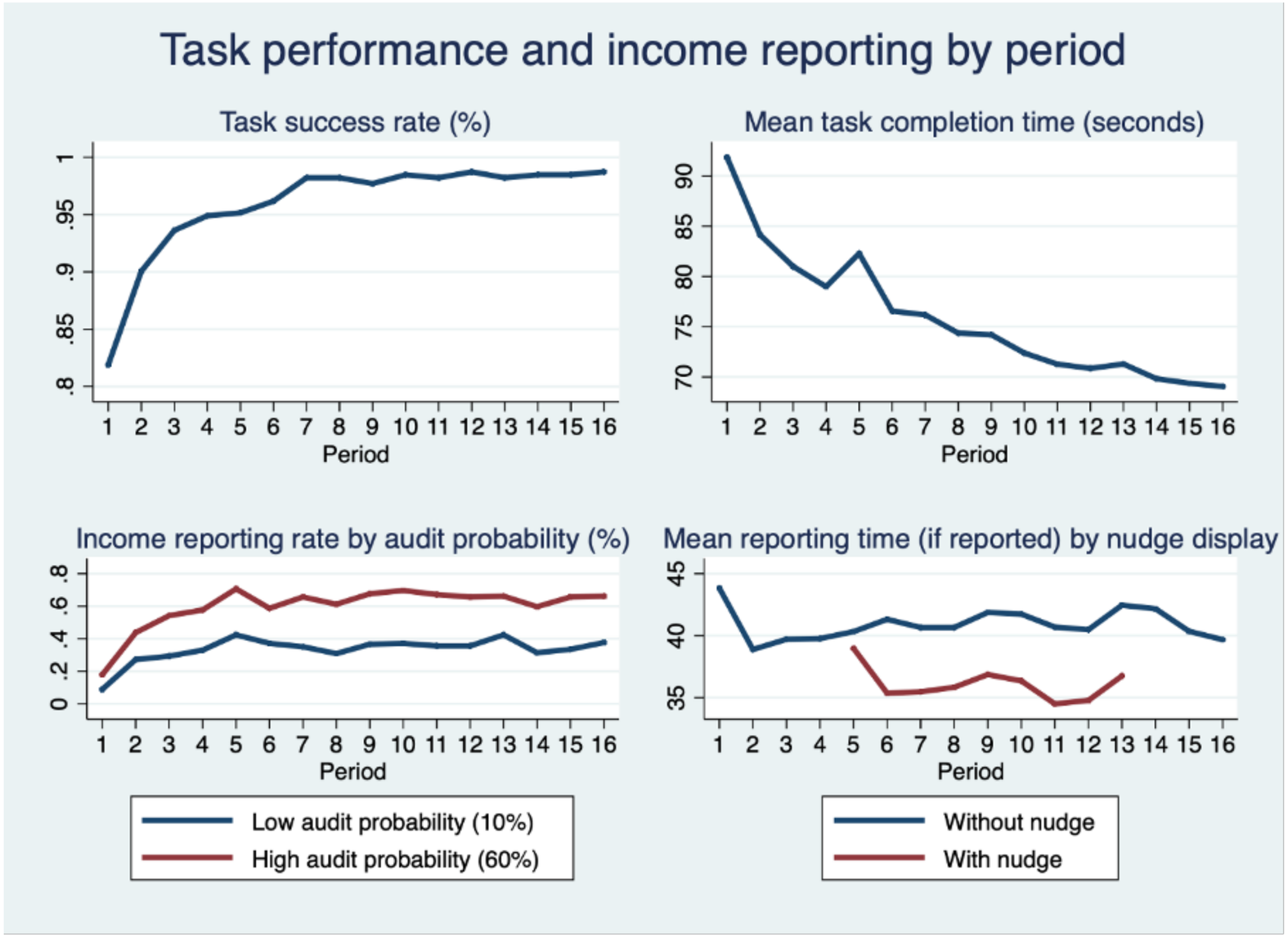

First time nudge appearance
All analyzed periods
Share of reporting decisions
0
.2
.4
.6
.8
1
Low (10%)
High (60%)
Probability of audit
No nudge
Nudged

Box plot

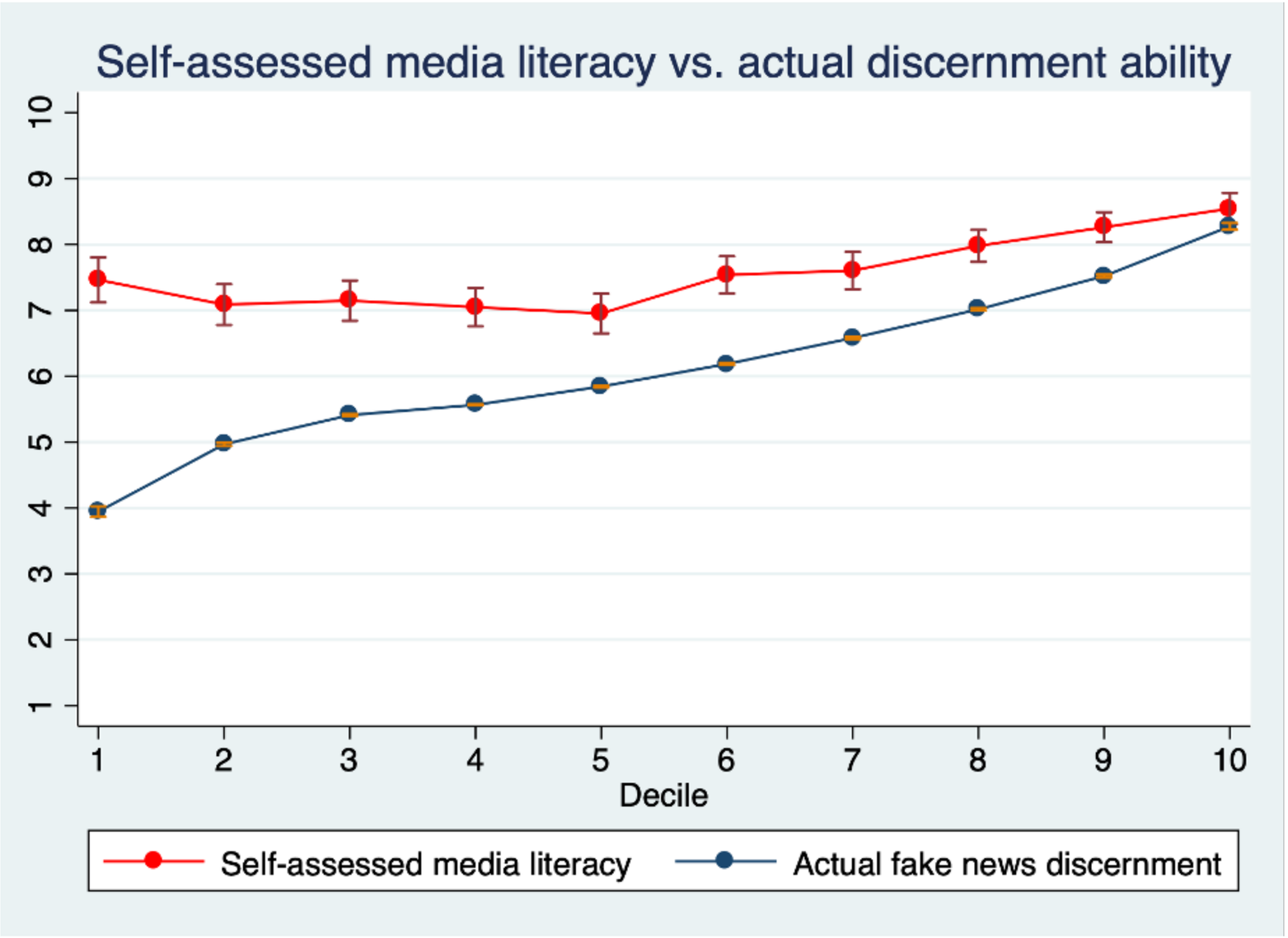

Estimates in the Control treatment
Seconds
0
500
1,000
1,500
LOW GROUP
HIGH GROUP
Round 1
Round 2
Round 3


Actual duration in the Control treatment
Seconds
0
500
1,000
1,500
LOW GROUP
HIGH GROUP
Round 1
Round 2
Round 3

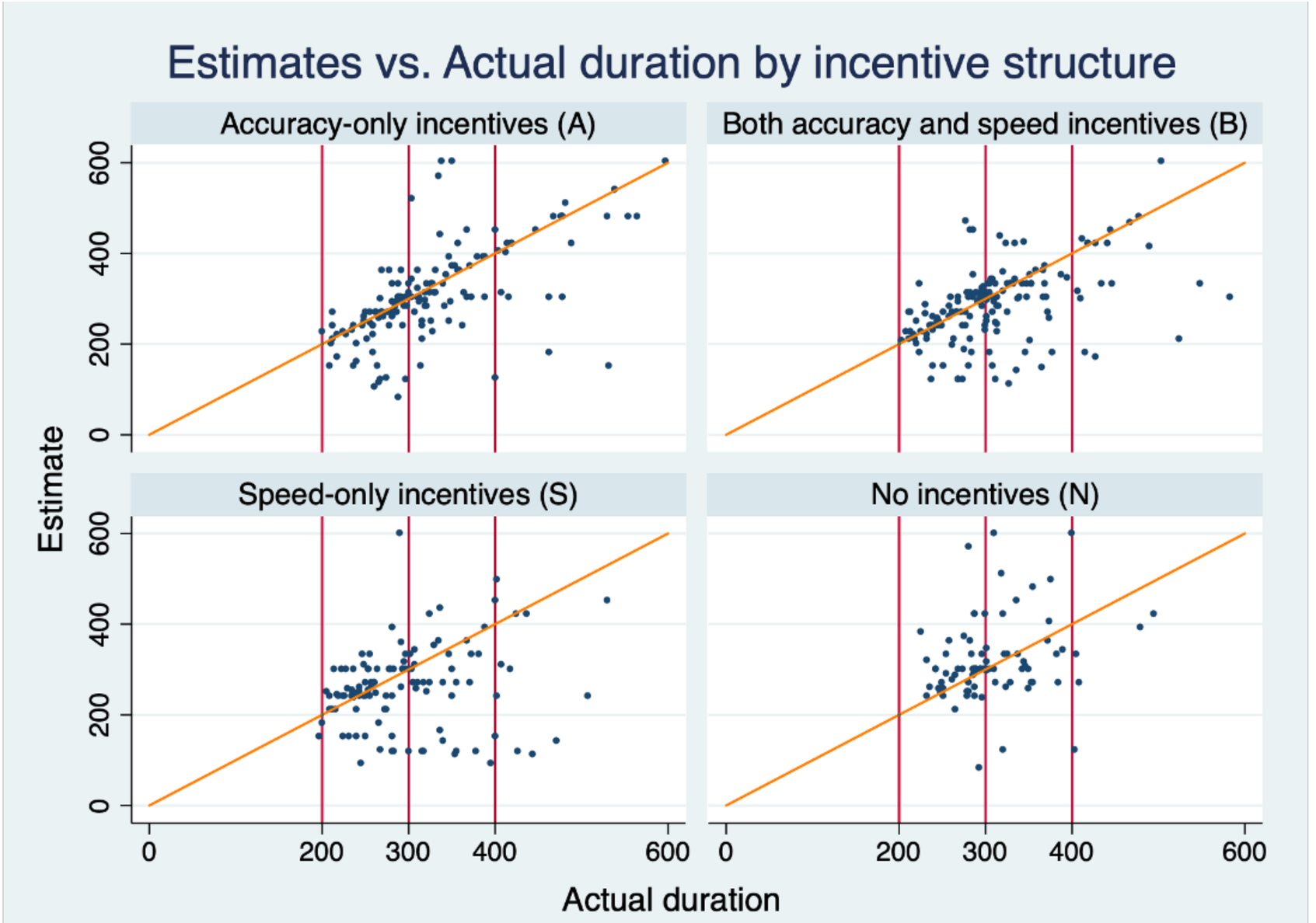
Line chart


Task performance and income reporting by period
Task success rate (%)
.8
.85
.9
.95
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Period
Mean task completion time (seconds)
70
75
80
85
90
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Period
Income reporting rate by audit probability (%)
0
.2
.4
.6
.8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Period
Low audit probability (10%)
High audit probability (60%)
Mean reporting time (if reported) by nudge display
35
40
45
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Period
Without nudge
With nudge

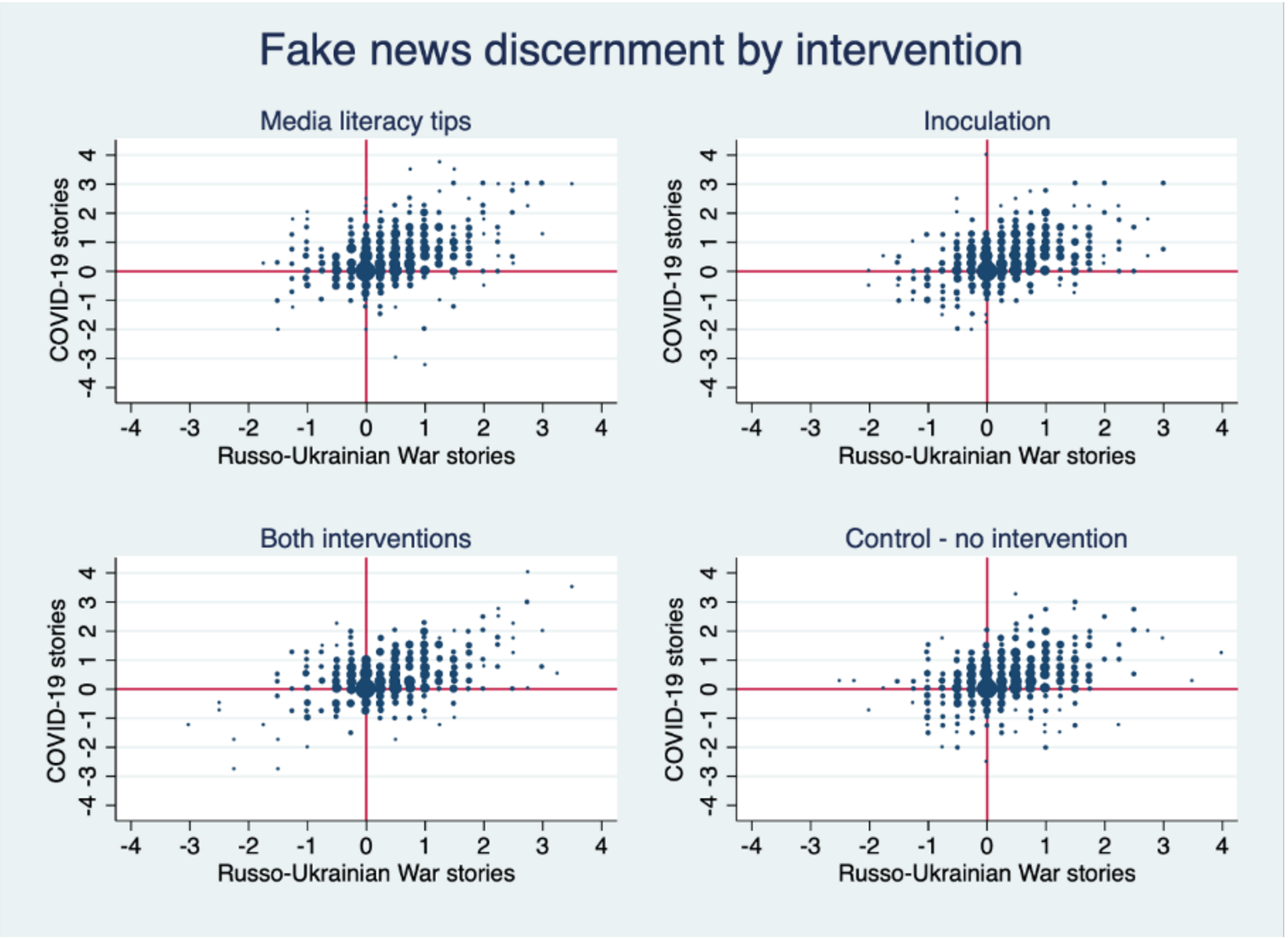
Line chart


Self-assessed media literacy vs. actual discernment ability
1
2
3
4
5
6
7
8
9
10
Decile
Self-assessed media literacy
Actual fake news discernment

Scatter plot



Scatter plot



Pearson correlation

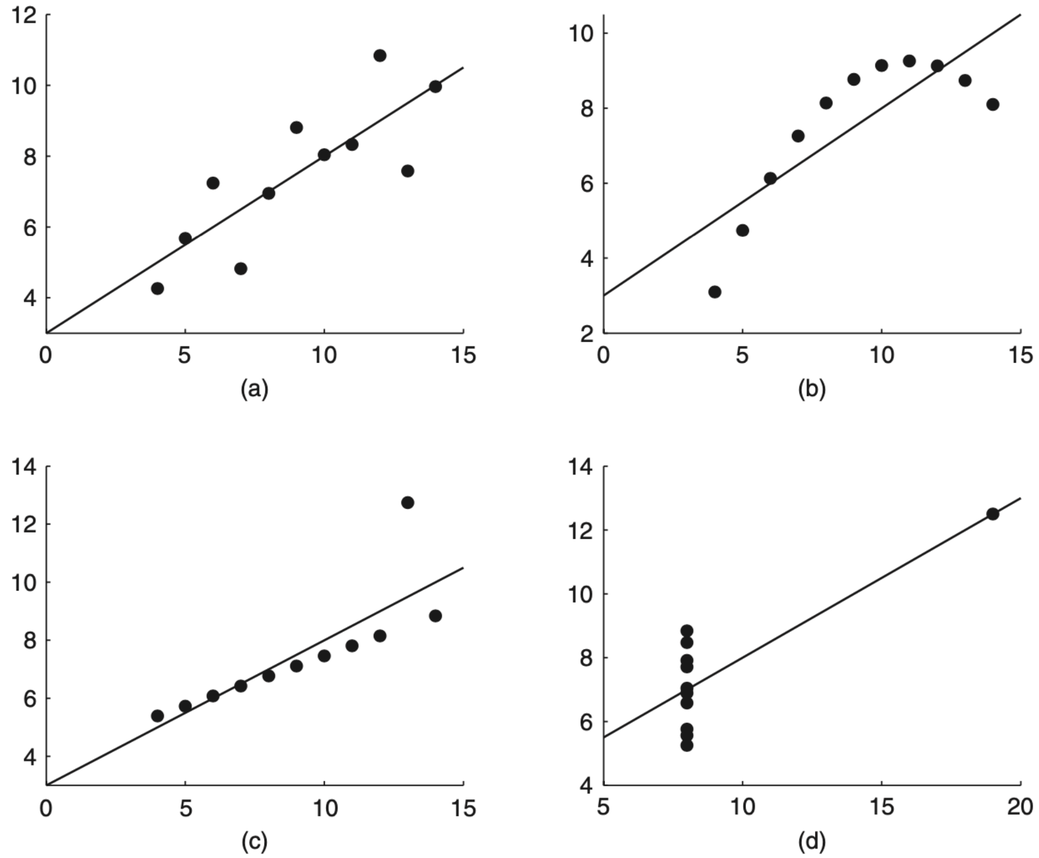
- The assessment of the degree of association between two variables is typically a measure of correlation. The basic measure here is the Pearson correlation, which measures the degree of linearity of the bivariate relationship between the two variables.

- The Pearson correlation is a standardised covariance (i.e. the ratio of the covariance between the two variables to the product of their variances). The Pearson correlation is bounded below by –1 (a perfect negative linear relation) and above by 1 (a perfect positive linear relation).

- The Pearson correlation is not affected by changes in either the scale (e.g. multiply variables by 2) or the location (e.g. add a constant) of the variables. Two independent variables have a correlation coefficient of 0. However, a finding of zero correlation between two variables can only be interpreted as indicating independence in special cases, such as the bivariate normal distribution. Using the Pearson correlation would be misleading here, due to its underlying assumption of linearity.

- Figure on the next slide presents Anscombe's quartet, a series of four data sets constructed by Anscombe (1973) to underline the importance of visualising the data before carrying out statistical analysis. Each data set consists of 11 pairs of points with a Pearson correlation of 0.816. The mean of the first variable (on the x-axis) is always 9, with a sample variance of 11. The mean of the second variable is approximatively 7.50, with a sample variance of between 4.122 and 4.127. Pearson correlation makes sense only in panel (a), for normally distributed data with a linear relationship. In panel (b) the relationship is clearly non-linear. Panels (c) and (d) show how the Pearson correlation is sensitive to outliers. In panel (c), without the outlier, the correlation is 1. One single outlier suffices to reduce this to 0.816. In panel (d), the correlation without the outlier is 0 but one single outlier suffices to increase this to 0.816.

Pearson correlation



Correlation Versus Causation

- The strength of association between two variables can be considered from a quantitative and qualitative point of view. In qualitative terms, the strongest relationship is a causal one. This means that the value of one variable causes a change in the value of another variable. For example, the force transmitted from one foot to a football is one of the reasons why the football flies so far.

- A correlation between variables is a qualitatively weaker form of relationship and exists when it can merely be observed that the increase of one variable is accompanied by an increase or decrease of the other variable. Even a perfect correlation does not necessarily mean that the variables are also causally related.

- For example, it may be observed that the amount of hair men have on their head and their respective income are inverse to each other, i.e. the less hair men have, the higher their income. If there were a causal relationship here, all men would probably shave off their hair in the hope of becoming richer. The actual causal relationship can be established easily if a third variable, age, is included.

- The older a man is, the more professional experience he has and, therefore, the higher the average income he earns. At the same time, it is in the nature of things that hair loss in men is also age-related. Age therefore has a causal effect on both the amount of hair and the average income of working men. Causation always means a correlation, but not every correlation means causation. To put it another way, if two variables are not correlated, there cannot be a causal relationship either. However, even if no causal relationship exists, there may well be a correlation.

- A simple statistical models merely measure the strength of a relationship and therefore provides purely quantitative information on the relationship between the variables. The relationship quantified by a statistical model is only ever causal to the extent that the experimental design, which was carried out in advance, has allowed it. The three factors of control, repetition and randomization are decisive for the causality in an experiment. Experiments in which randomization is not possible are called quasi-experiments. It is much more difficult to derive causal relationships in such experiments, but there are special statistical models and estimation methods that facilitate the determination of causalities (regression discontinuity designs). These include, in particular, instrumental variables estimation and the differences-in-differences (DiD) method.