Experimental economics

Lecture 6: Descriptive statistics

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.

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?

- Data processing: 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.

- Data analysis: 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?

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?

Operationalisation of variables

- The result of a measurement is a set of certain variables, i.e. properties that can take on different values. The rules by which we assign numerical values = operationalization. We cannot assume that there is always only one correct way to operationalise variables. On the contrary, it is up to us how we grasp the variables. Operationalization also depends on our research background and our topic.

- The operationalisation has to be consistent - i.e. we cannot measure one participant's age in years and another participant's age in months, etc.

- We have to be completely transparent about the operationalisation of our variables. In fact, it is common for student papers that readers are not able to understand how the variables are operationalized. Most of the time, this is a problem of the researcher, not the reader.

- In everyday life we too often rely on others to automatically understand what we mean. That it is self-evident, clear from the context, etc. To our surprise, this expectation is often completely wrong - misunderstandings arise on a daily basis. In science, however, we need to prevent misunderstandings. So nothing is self-evident. All information must be clearly and transparently stated.

- When operationalising a variable, it is also essential to reflect on what scale we are measuring the variable - as the same variable can be measured on different scales.

Types of variables

- A nominal variable (also categorical or qualitative) does not express quantity. It is merely a numerical designation of a phenomenon. An example might be the color of the eyes, which can be, for example, brown (we can code it as "1"); green ("2"); blue ("3"); or other ("4"). However, these are separate qualities that are in no way related to their numerical designation. The fact that blue is numbered "3" and green is numbered "2" does not mean that blue has more or less of anything.

- An ordinal variable, unlike a nominal one, already expresses some basic measure of quantity, i.e., some order. For example, if we rank all the students in a class in ascending order of height, we can say that the person with the higher number is taller. A typical example of an ordinal scale is a self-report scale, e.g., How satisfied are you with the quality of this course? 1 = I am not at all satisfied, 2 = I am rather not satisfied, 3 = I am neither satisfied nor dissatisfied, 4 = I am rather satisfied, 5 = I am very satisfied

- The problem with ordinal scales, however, is the fact that we cannot tell how much difference there is in the ranking, i.e. how much is one level different from another. We can arrange ordinal variables but cannot add, subtract multiply or divide.

- The quantitative variable (also cardinal) is a real number, unlike the nominal and ordinal variables. The differences between points on the scale are meaningful; we can say that the variable has a unit of measurement. We distinguish quantitative variables into two types - interval and ratio variables.

- Interval variables have a standard unit, but they do not have a true zero point to indicate the complete absence of a given quality. The most common example is temperature in degrees Celsius (°C). If today is 20 °C and yesterday was 23 °C, the difference between these values is 3 °C, so it is the same difference as between 17 °C and 20 °C. On the other hand, 0°C does not imply the absence of any temperature, it is just an agreed value corresponding to the melting point of water. So if it is 20 °C today and it was 10 °C last week, it would be strange to say that it is twice as warm today. Numerically it makes sense, but substantively it does not. Adding and subtracting interval variables doesn't cause problems, but multiplying and dividing might.

- The second type of quantitative variable is a ratio variable, which is a real number with a true zero point. Examples might be a number (e.g., the number of points on a test), a length measured in meters, a reaction time measured in milliseconds, or a heart rate measured in the number of beats per minute. The 0 in these is not the agreed thing but the absence of the phenomenon (points, length, duration, heartbeats). We can perform all mathematical operations with a ratio variable.

Measures of central tendecies

- If we want to characterize an entire population with a single data point, it is useful to describe the so-called central tendency (or center of the distribution). In addition to the arithmetic mean, the central tendency is also expressed by the modus and the median.

- The arithmetic mean (mean, $\bar{X}$, M) of a variable X is calculated by simply summing up all the elements in the dataset, which is then divided by the number of elements. The average has its own assumptions and specifics. It can only be used with quantitative data, since we calculate the mean by sum and proportion. One of the limitations of the mean is its sensitivity to outliers, i.e., data that are very significantly different from the rest of the set. An outlier can bias the mean so that it ceases to describe the center of the data well. A typical example of a variable where the mean is affected by outliers is financial income (wage) in the population. If we are trying to compare our wage with wages of other people in the country, the average wage figure will not be very useful to us, since a relatively small number of high-income individuals distort the mean so that its value is higher than what most people earn. Thus, if we were to think of the average wage as the middle wage of a person in this country, we would be wrong. The median would be more useful to us here.

- The median ($\tilde{X}$, Md) is a central value of the data in the set. If we arrange the values of all the elements in the set, the median will be exactly in the middle, dividing the set into two equally sized halves. In other words, half of the cases will have a value higher than the median and half of the cases will have a value lower than the median. So if I find that my wage is higher than the median, I know that majority of people in the country are earning less.

- If we had a set with the values: 12, 13, 15, 18, 21, the median value is 15. If we have an even number of elements, we will not have one middle value in the set, but two. In that case, the median is their average. For example, in a set with the values: 11, 13, 14, 15, 16, 21, the median is 14.5. The median is more suitable for presentation of ordinal data. At the same time, unlike the mean, the median is not sensitive to outliers in the data, so it may be a more appropriate measure of the central tendency when we are working with data that contains outliers. On the other hand, the median may underestimate the importance of important values at the extremes of the distribution.

- Modus (Mode) is a relatively simple statistic. It is the most frequently occurring value in the set. It can be applied to any data, including qualitative data. In some datasets we can obtain multiple modal values, we refer to them as multimodal (bimodal in the case of two modes). However, it is important to note that the informational value of a mode is limited because it does not reflect the distribution of the other values in any way, and so it may be that the mode is not a good representation of the whole dataset.

Measures of variability

- The statistics we have worked with so far have served to characterize the distribution based on its central tendency. Measures of variability answer the question of how far apart the data are, or how far from the mean our data occur.

- The range of a variable expresses the difference between the highest and lowest values. We can describe the data by the minimum and maximum. It is a very simple description of the variability, and it can be very imprecise since the range includes very distant values. Moreover, the range does not capture the relative frequencies of each value, so we cannot tell how frequent which values are.

- The interquartile range (IQR) is the range of the middle 50% of all data, which is the range between the first and third quartiles (or the 25th and 75th percentiles). If we divide ordered data into four equal parts, we obtain values of quartiles. If we arranged them by size and divided them into 100 parts, we would obtain the percentiles. The boundary of the 1st and 2nd quartiles, which represents the 25th percentile, bounds the lowest 25% of the values in the set. The 50th percentile is the median, since it is the value that divides the data into two equally sized halves. The 75th percentile is the boundary between the third and fourth quartiles, so it bounds the top 25% of values.

- The standard deviation (SD, s) is probably the most common indicator of variability, which is usually presented together with the mean. When working with quantitative data (interval, ratio), we can express variability in terms of deviations from the mean. It can be calculated as the difference between each value in the set and the mean. Since some deviations are positive and some are negative, summation would reset them to zero. Thus, we work with squares of the deviations, and we sum them.

- Although the sum of the squares of the variances gives us the total variability, it is dependent on the number of elements in the dataset. To standardize it, we need to divide it by the number of degrees of freedom (df), which is the number of observations minus 1 (N - 1). This gives the variance (variance, s2, σ2).

- Since it does not make much sense interpretatively to talk about squares of the variances, it is convenient to calculate the root square of the variance. This gives us the standard deviation, which is again expressed in the original units. The standard deviation thus tells us how much our data "deviates from the mean" on average.

- The symbols SD and s represent the standard deviation estimated from the sample. The symbol σ represents the hypothetical standard deviation in the population. In most cases we do not know it exactly, as we do not have data from the whole population, we only approximate it. It is the same with the symbols denoting the variance (s2, σ2).

Random Variables and Their Distribution

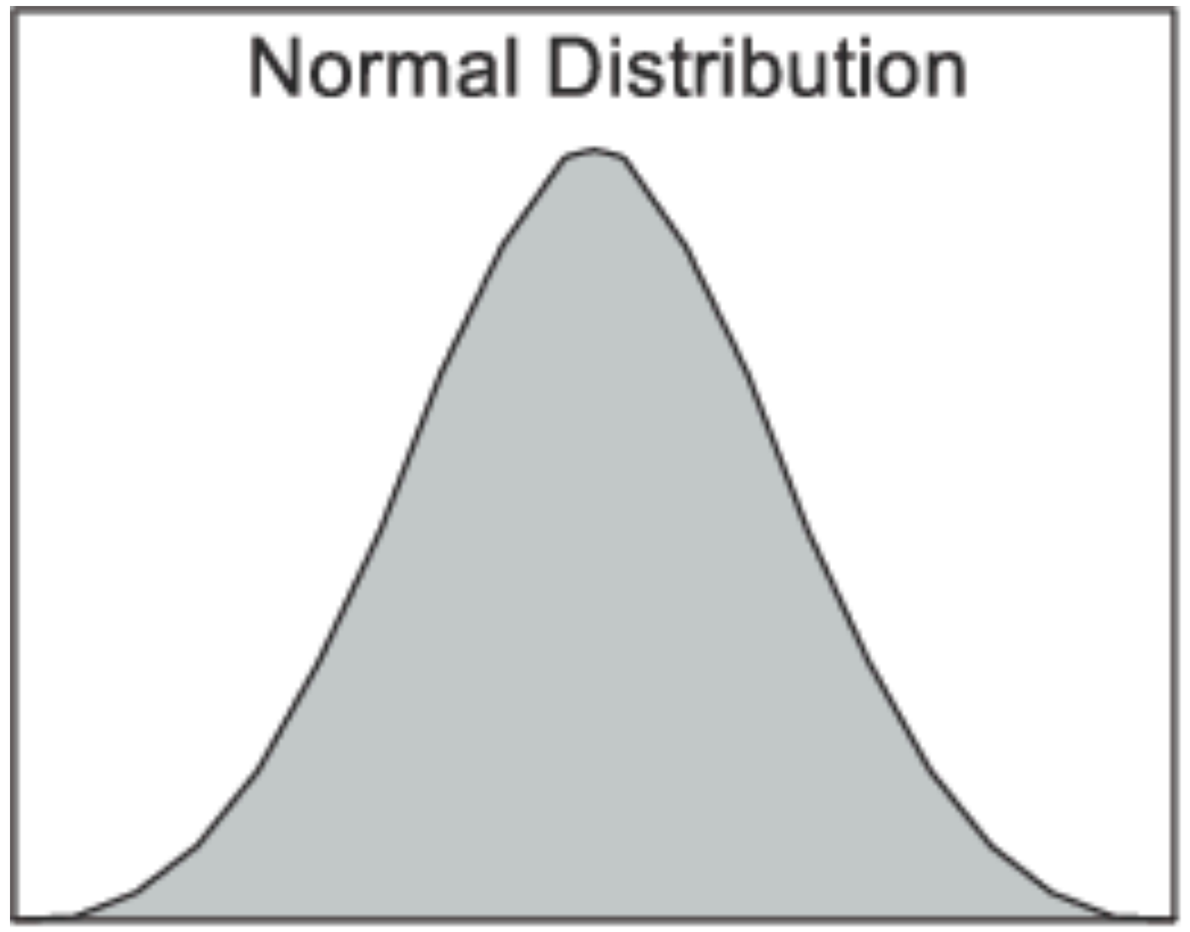
- In the statistical modeling of the relationship between variables, the dependent variable is interpreted as a random variable. Which values of a random variable are most likely and which are less likely is determined by their distribution. The so-called density function of a discrete random variable 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.

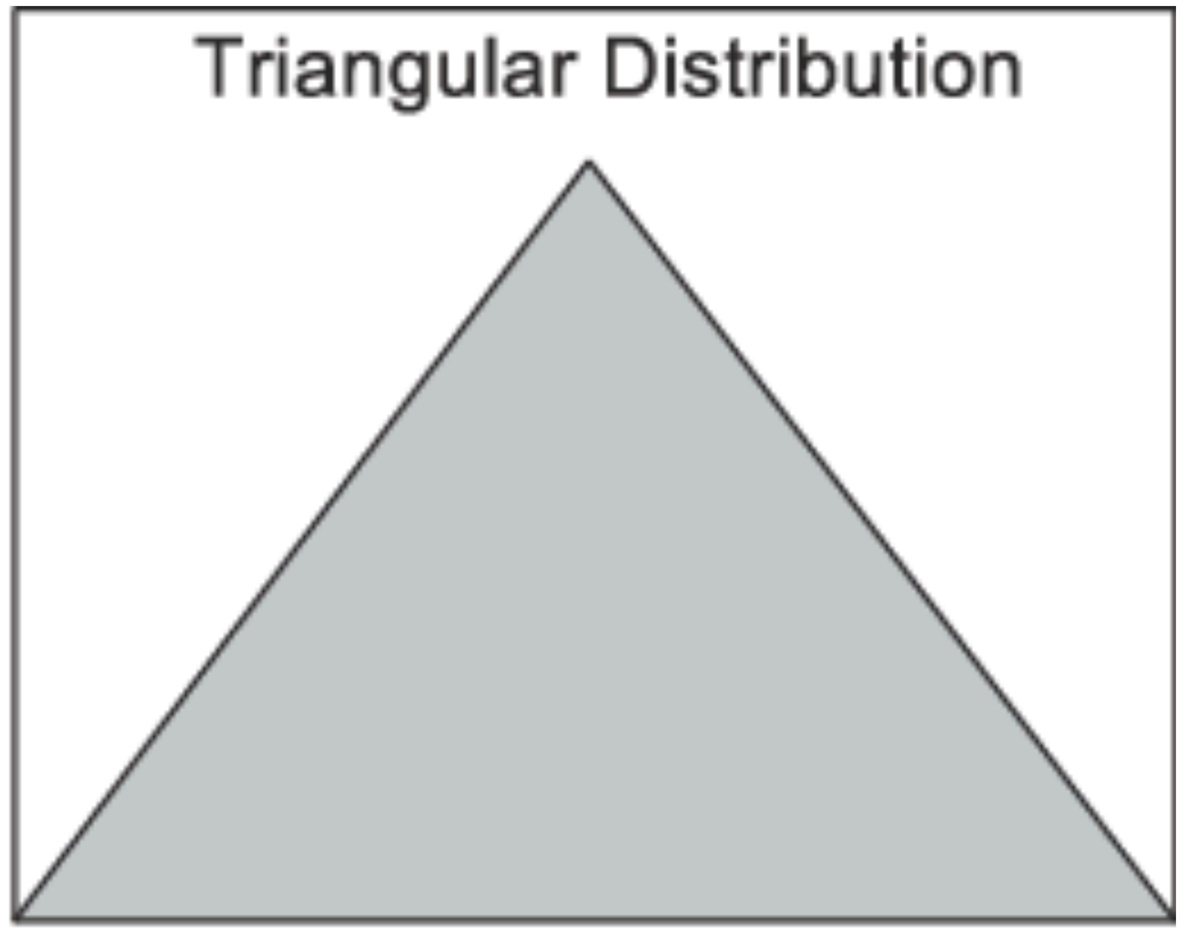
- Most statistical distributions have certain parameters which, depending on the value they have been set to, determine the shape of the density function. The three most important parameters are expected value, variance and degree of freedom. The expected value is the average of all the values drawn, 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 σ2. 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.

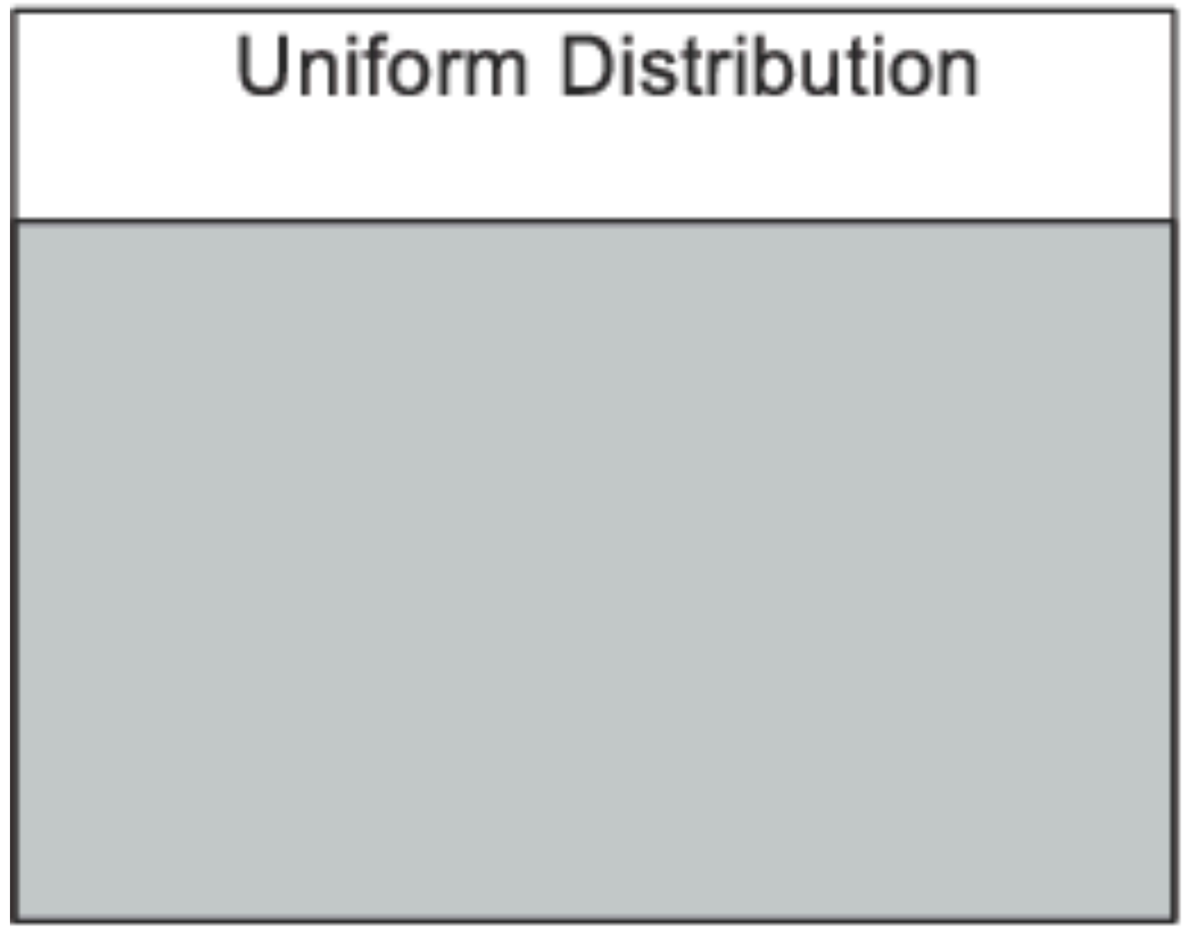
Random Variables and Their Distribution



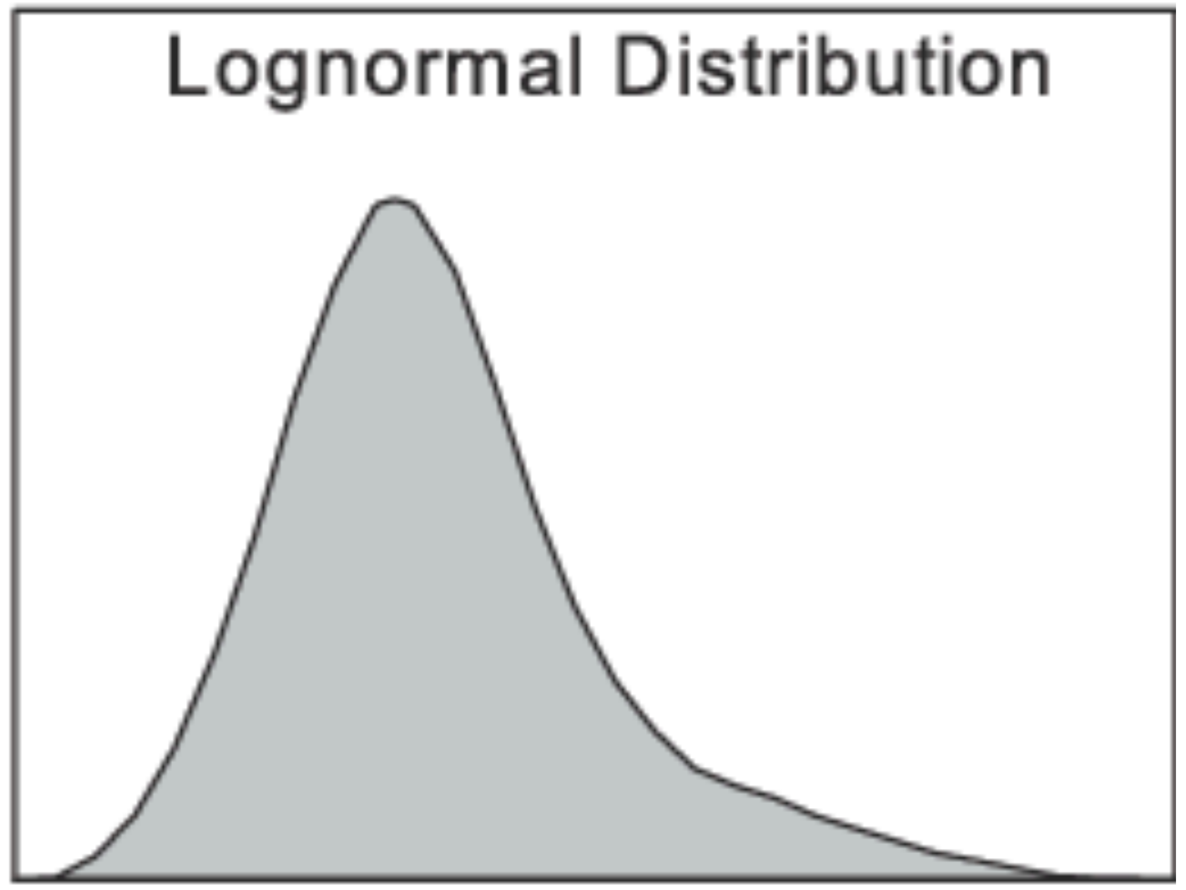
Normal distribution



Triangular distribution



Uniform distribution



Log-normal distribution

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

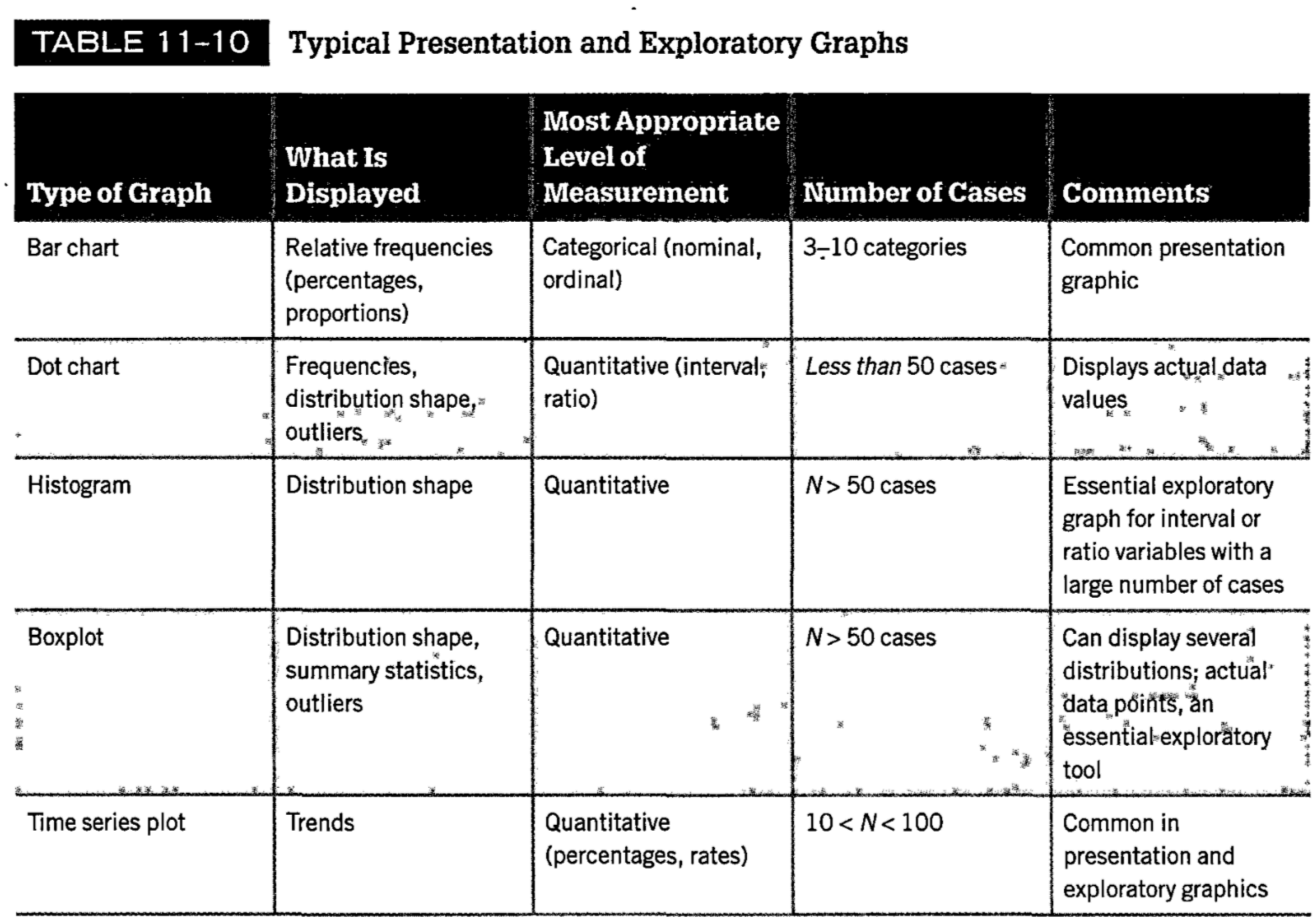


TABLE 11-10 **Typical Presentation and Exploratory Graphs**

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

- When working with quantitative variables, rather than displaying the frequencies of individual values, we may try to display the shape of the distribution and the data through the frequencies at certain intervals of values. Histograms are used for this purpose. In a histogram, the x-axis divides all possible values into intervals, and the height of the bar then shows the frequency (or relative frequency) of the measured values in that interval.

- The histogram thus gives us a good idea of the overall distribution of the data. Within the data distribution, we can assess its symmetry of densities, clusters, gaps (missing data), the occurrence of outliers, and the shape of the distribution.

- We also describe the shape of the distribution through numerical measure of skewness. We can also note the number of peaks (the most frequent interval of values) in the histogram. Although we most often expect one peak (unimodal shape), it may happen that there are two peaks (bimodal shape) or more than two peaks (multimodal shape). A histogram with two peaks (clearly distinguishable, of course, as if we had two bell shapes side by side) arises, for example, when the data come from two different distributions (e.g., they are not one but two different populations), and therefore these different distributions need to be identified and plotted separately.

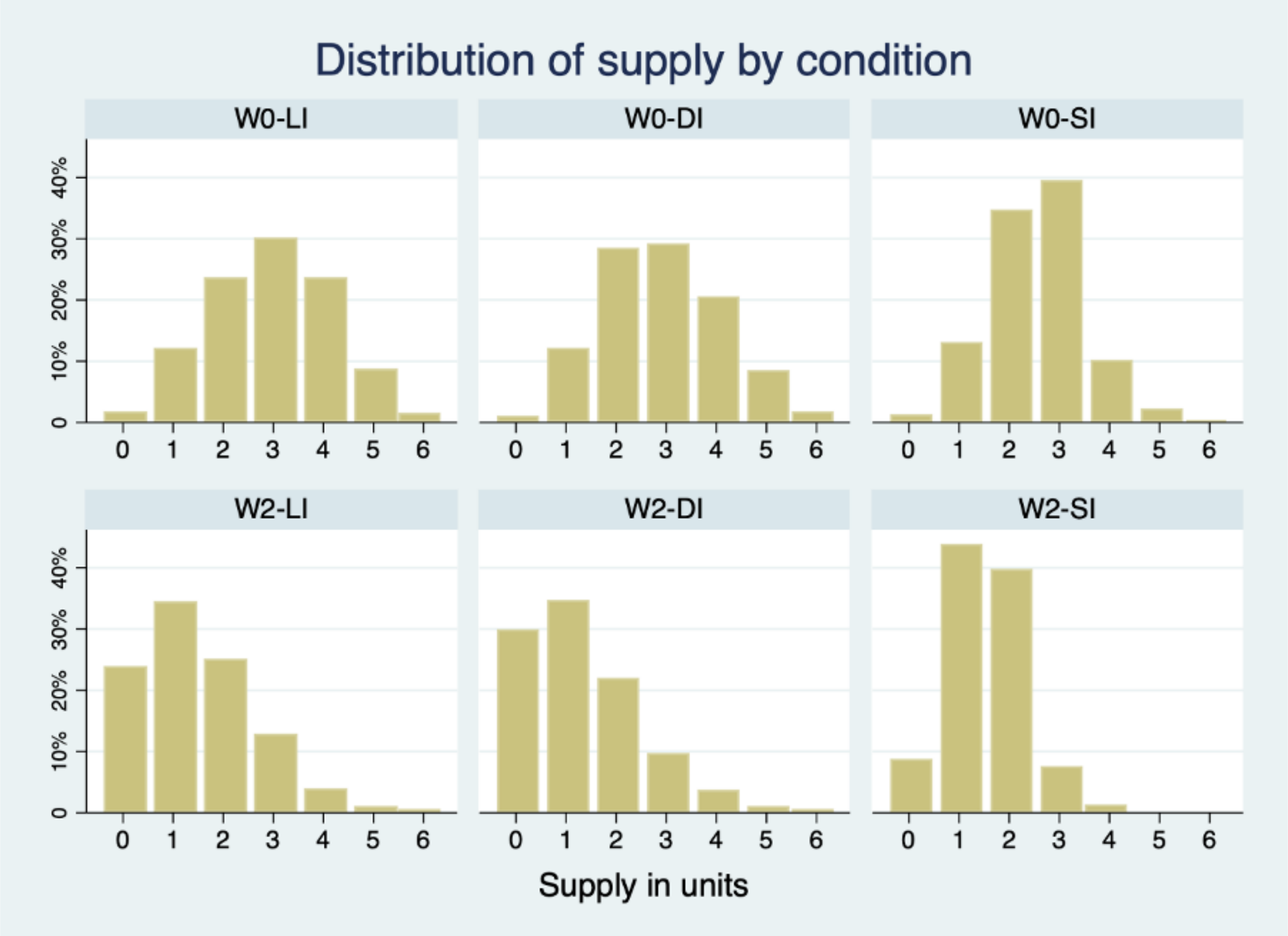
Histogram

Distribution of pooled estimates

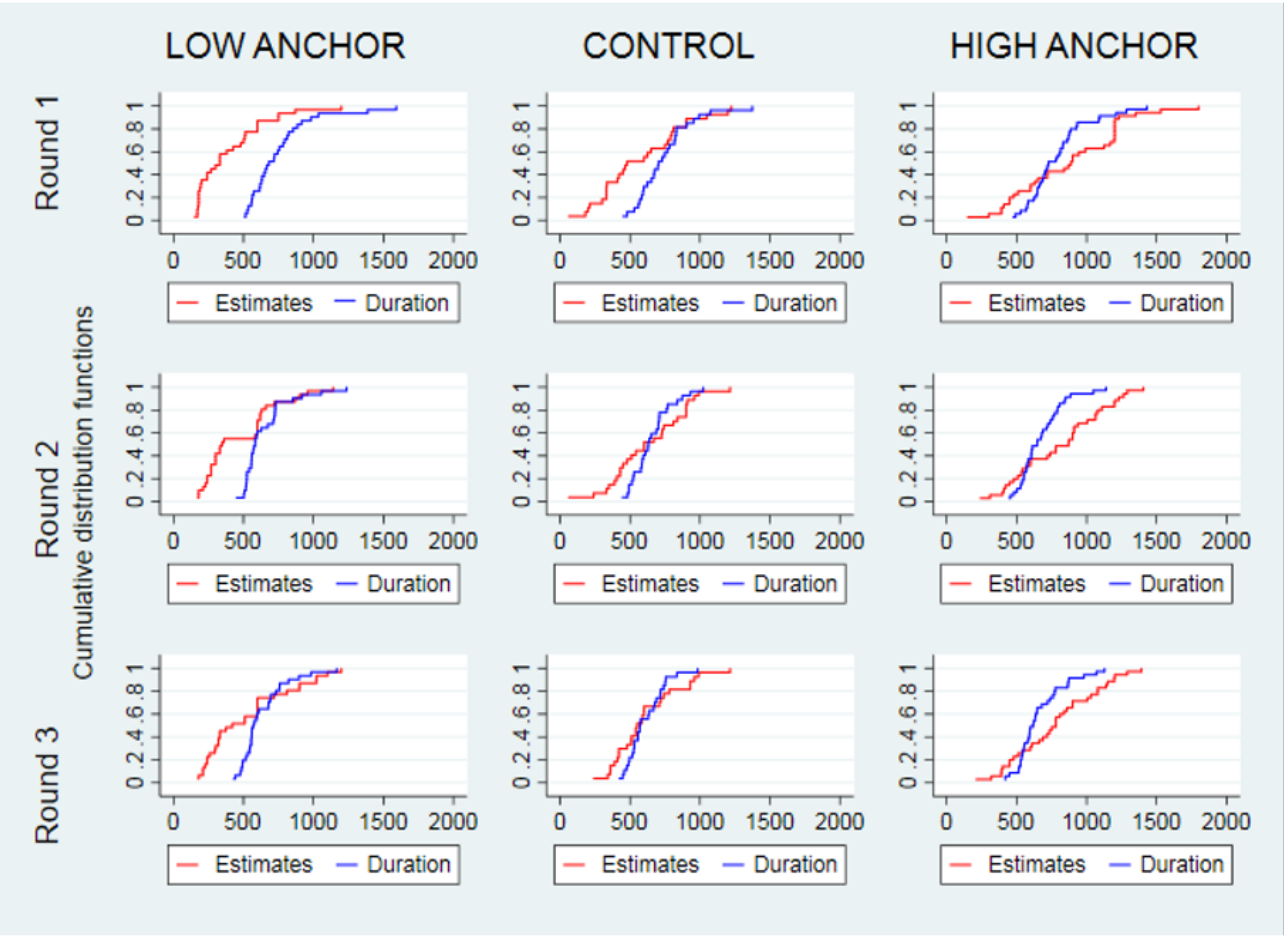
Distribution of pooled actual task duration

Frequency
0
.001
.002
.003
0
500
1000
1500
2000
Estimates (seconds)
Frequency
0
.001
.002
.003
0
500
1000
1500
2000
Actual duration (seconds)

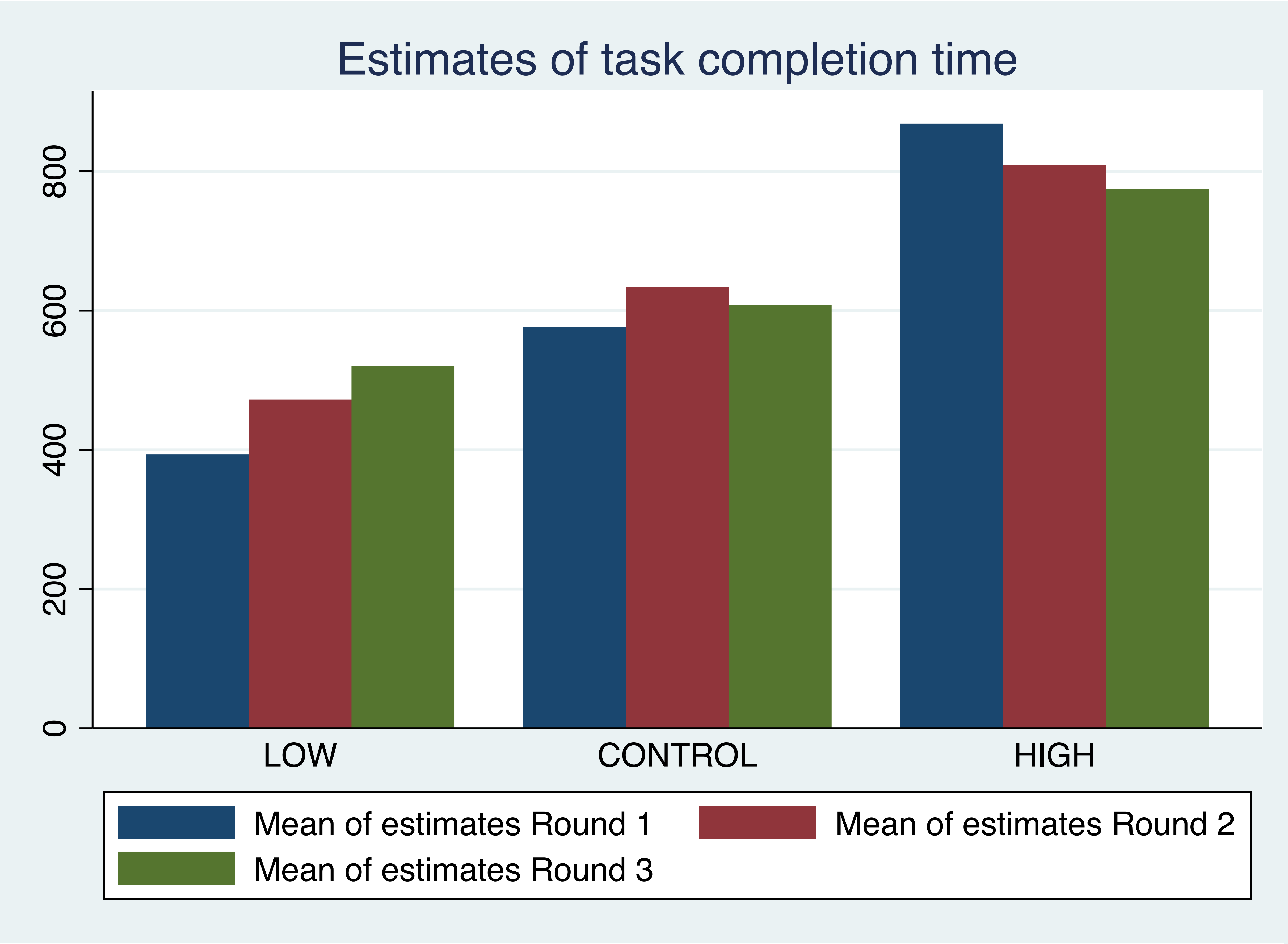
Histogram


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

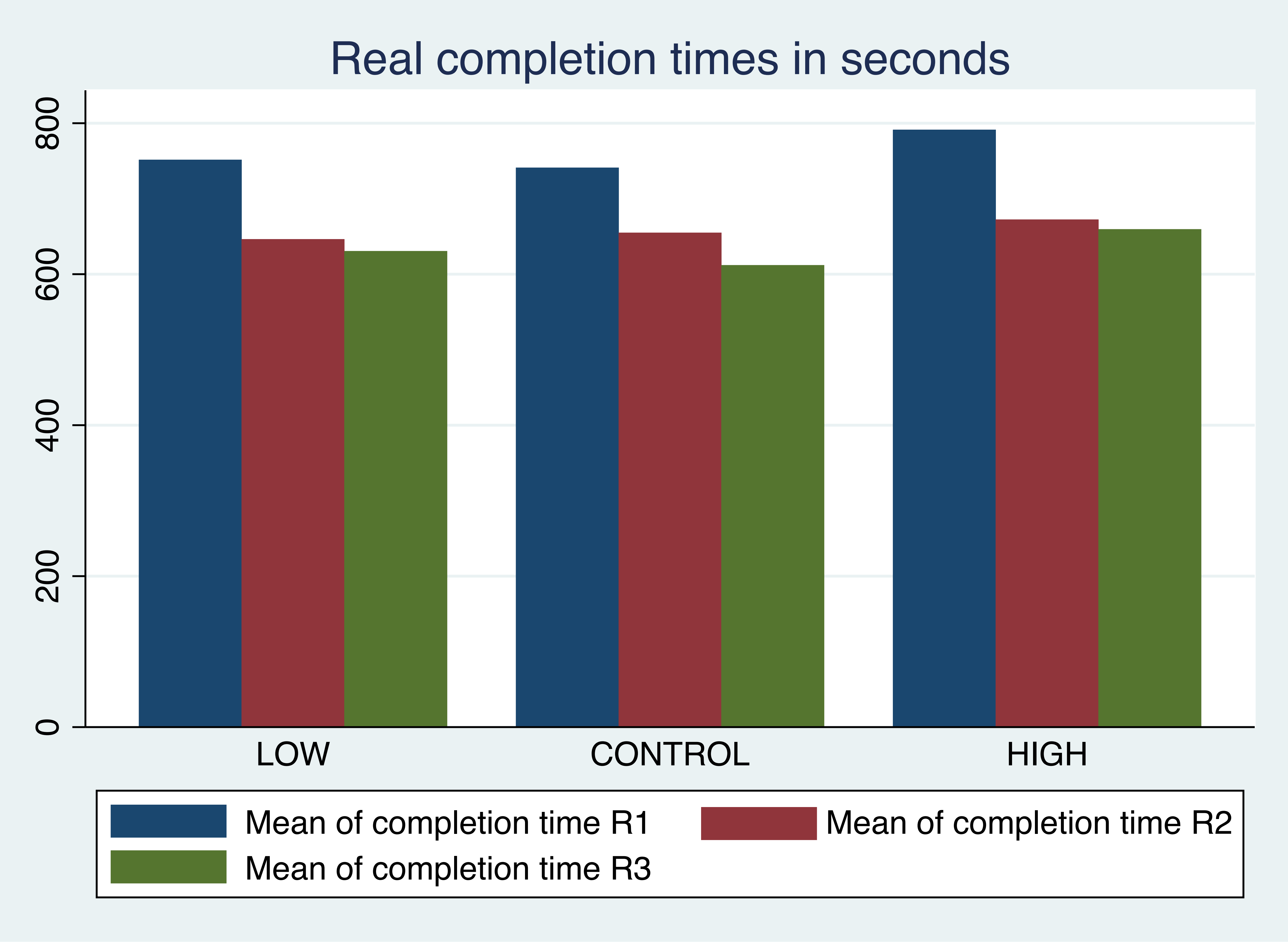
Cummulative distribution



Bar chart

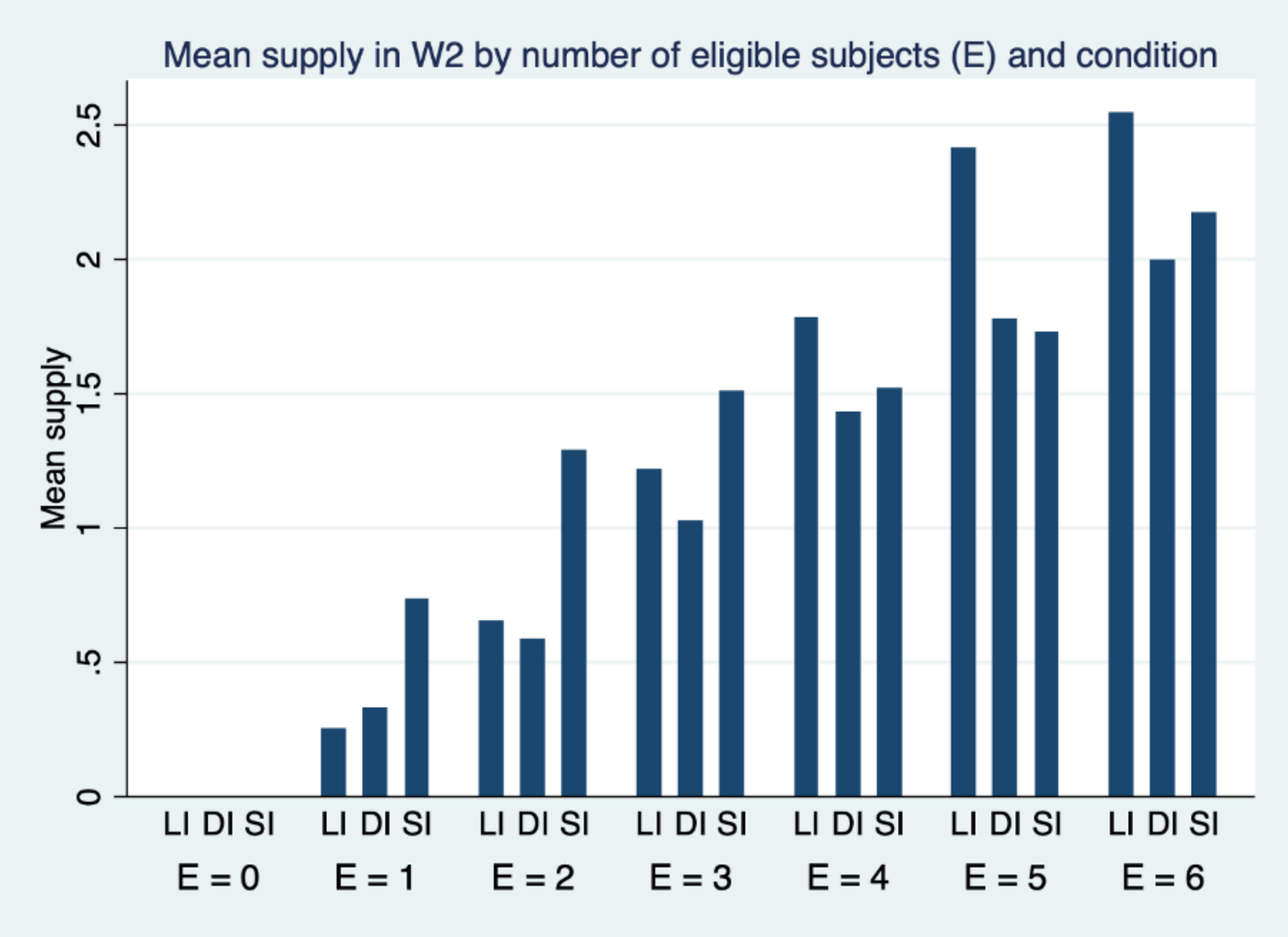


Estimates of task completion time

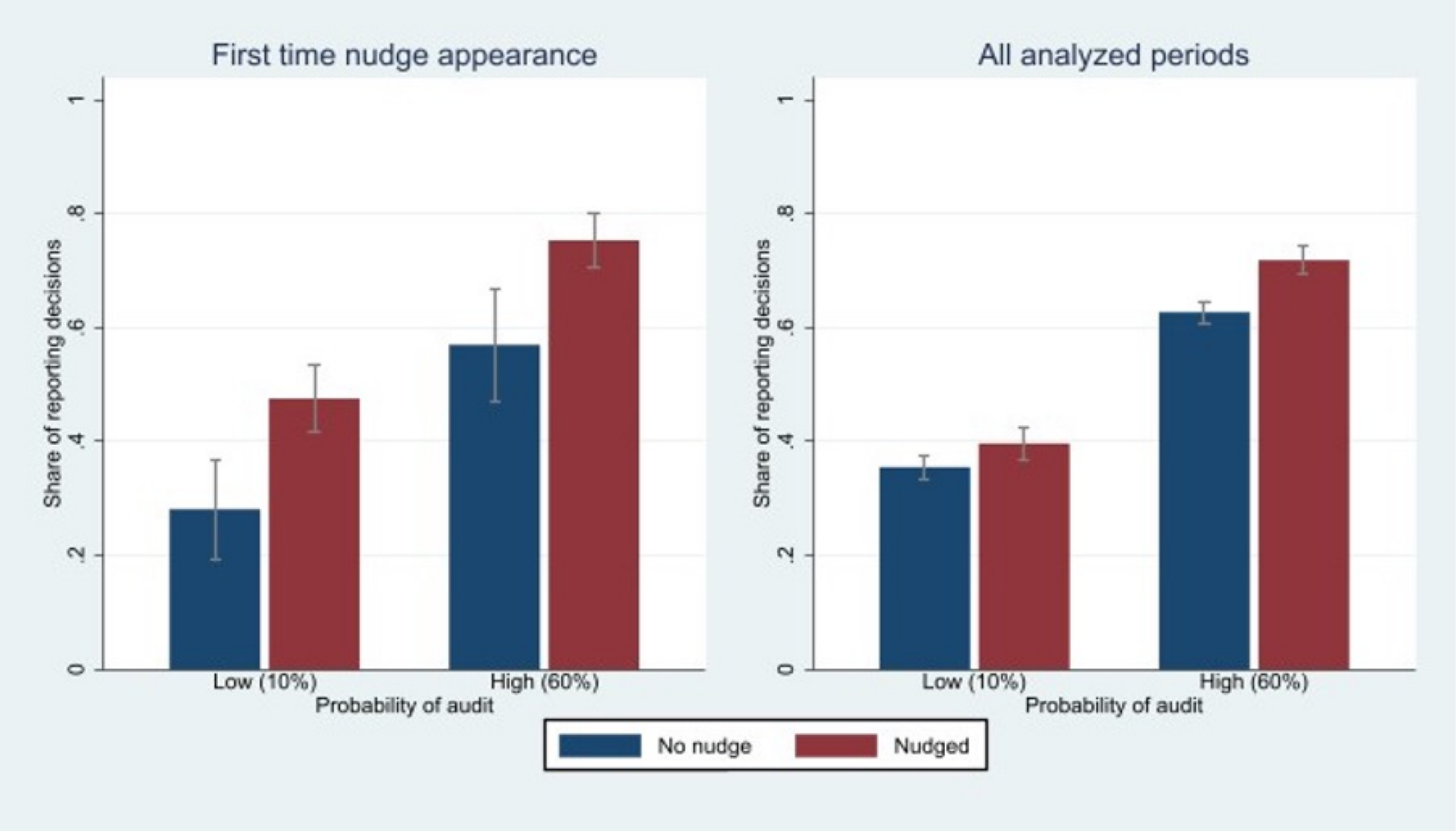


Real completion times in seconds

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
LI DI SI
E = 1
LI DI SI
E = 2
LI DI SI
E = 3
LI DI SI
E = 4
LI DI SI
E = 5
LI DI SI
E = 6

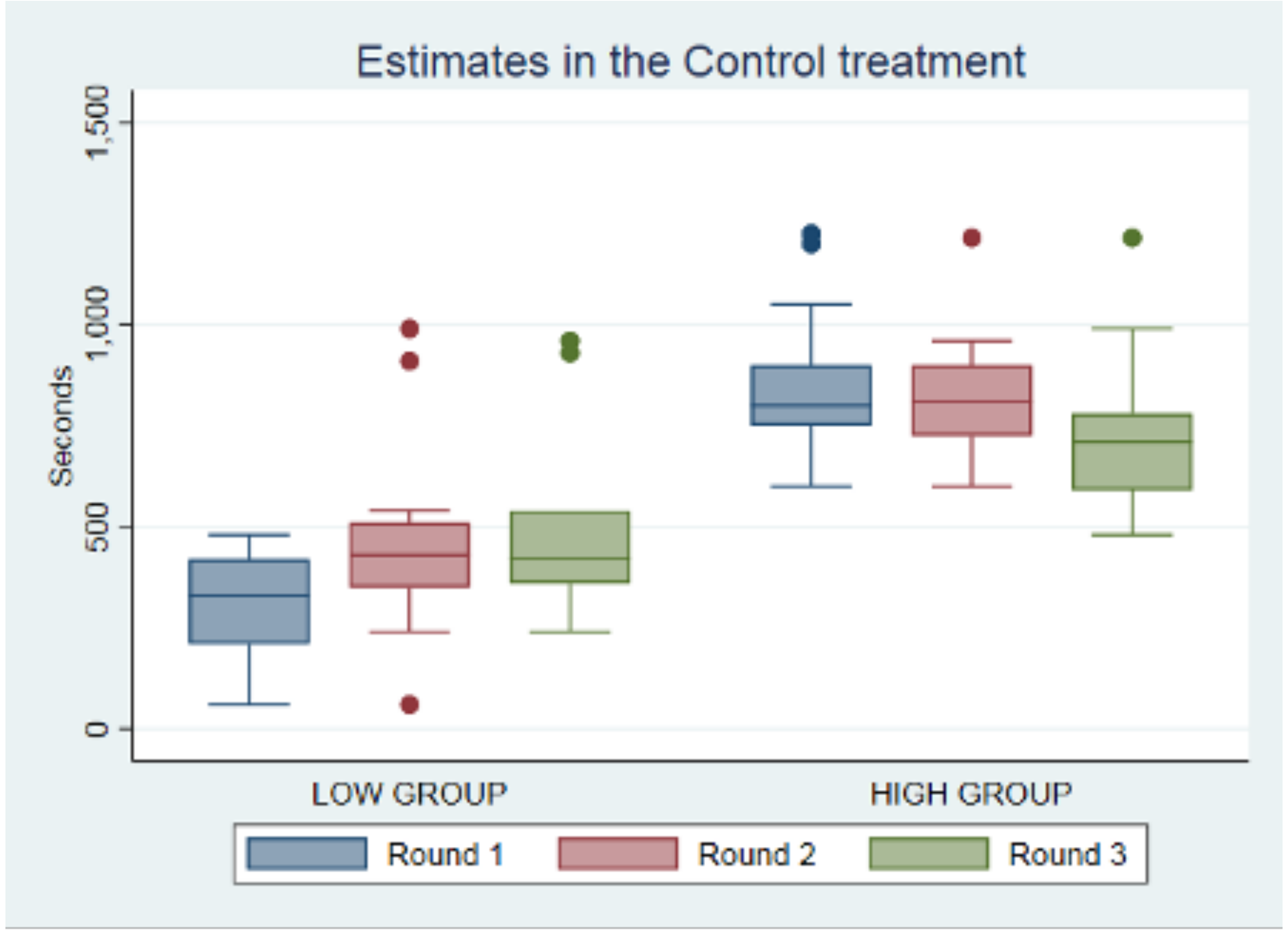
Bar chart

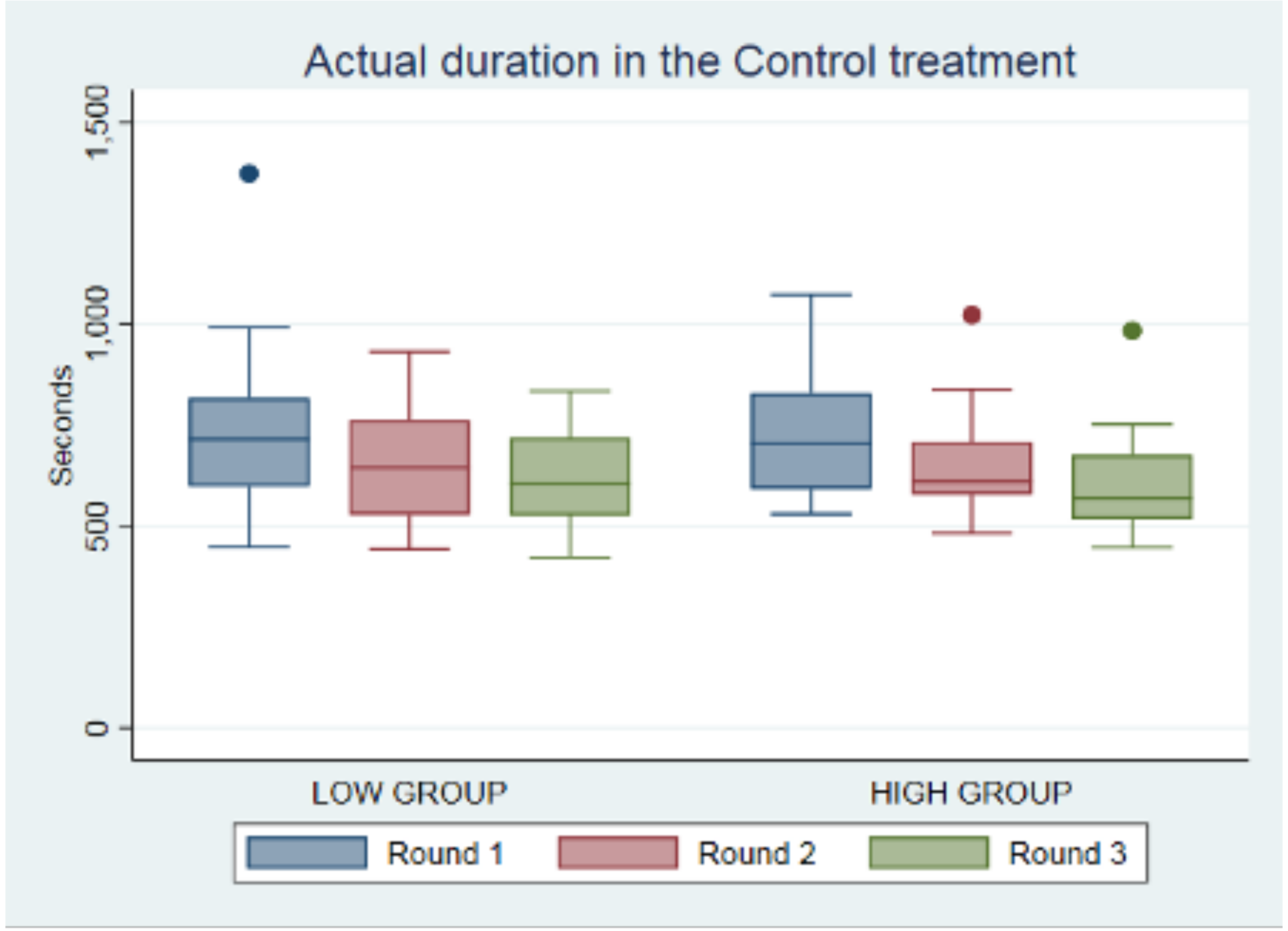

First time nudge appearance
All analyzed periods
Share of reporting decisions
0
.2
.4
.6
.8
1
Low (10%)
High (60%)
Probability of audit
No nudge
Nudged

Box plot

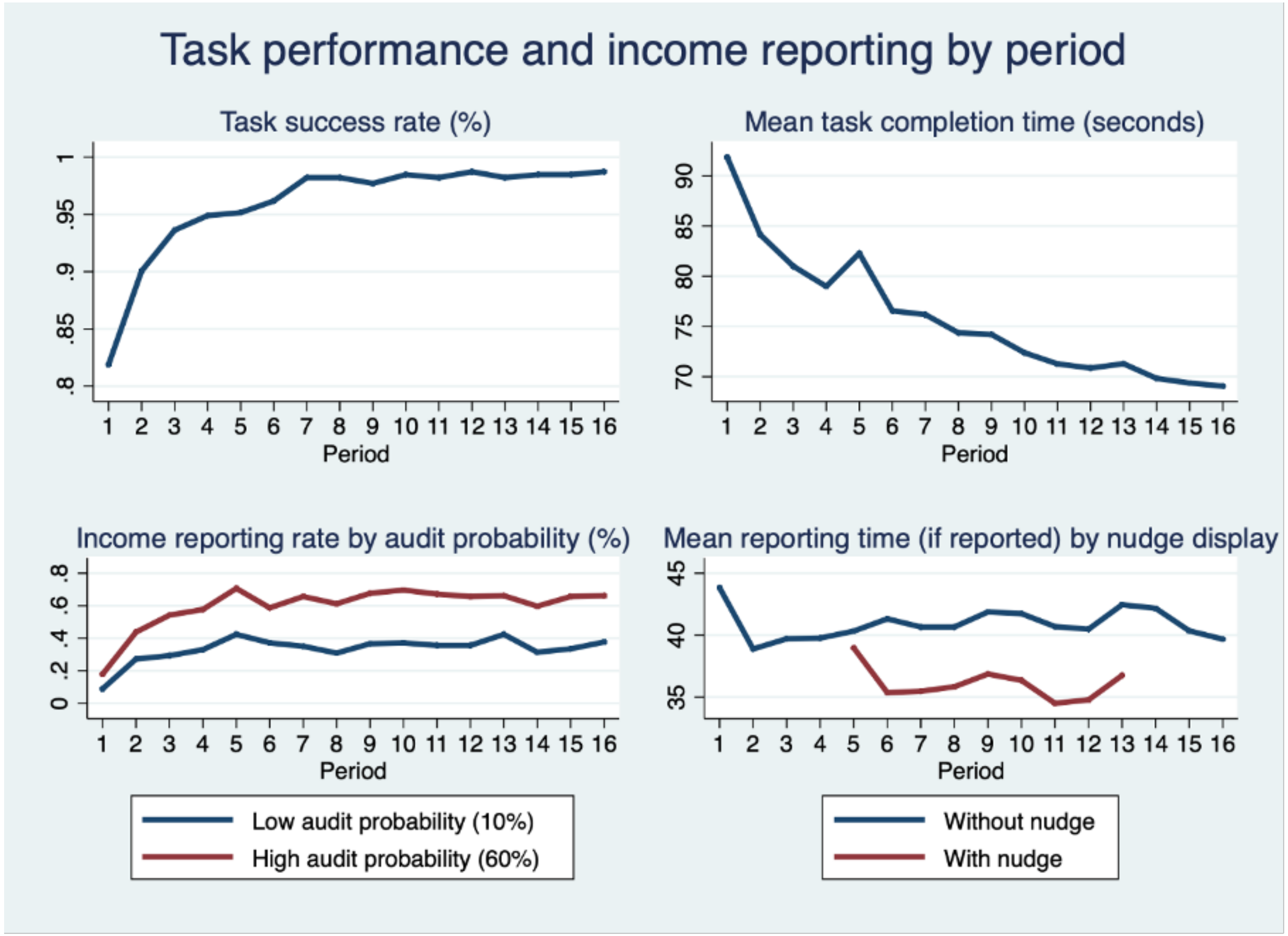
- A box plot is also one of the graphs used to display quantitative data. It is composed of five values:
  - minimum (the lowest value of a variable)
  - lower quartile (Q1, 25th percentile)
  - median
  - upper quartile (Q3, 75th percentile)
  - maximum (highest value of the variable).
- In other words, at the center of this graph is the median, which is surrounded at the top and bottom by a "box", and the space of the box makes up 50% of the observations (the inter-quartile range).
- The whiskers pointing up and down from the boxes denote scores of Q3 + 1.5 interquartile range (IQR) and Q1 + 1.5 IQR, respectively. The points above/below represent the so-called outliers, i.e., values further than 1.5 IQR from the value bounding Q3.

Box plot

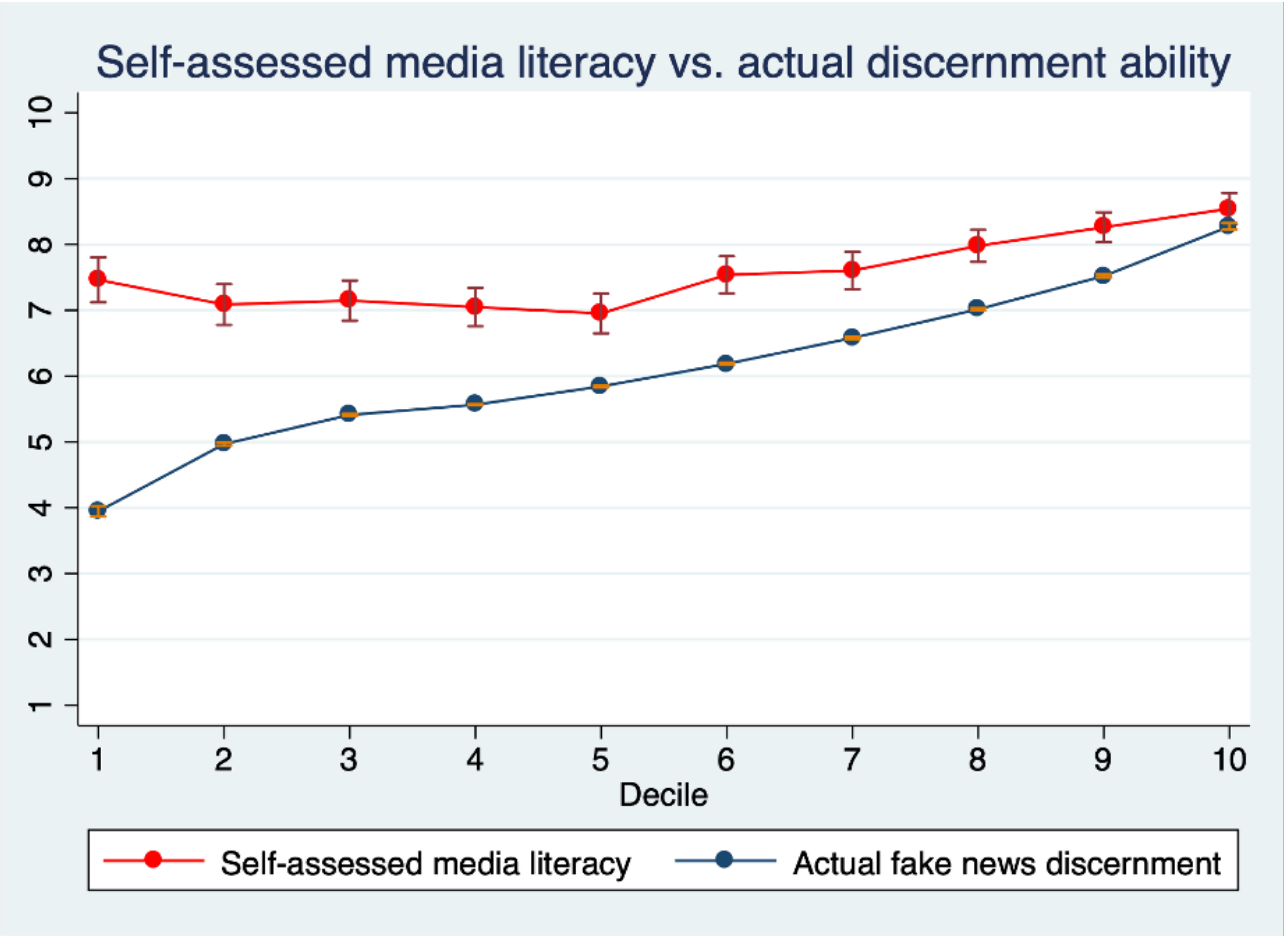

Estimates in the Control treatment
Seconds
1,500
1,000
500
0
LOW GROUP
HIGH GROUP
Round 1
Round 2
Round 3


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

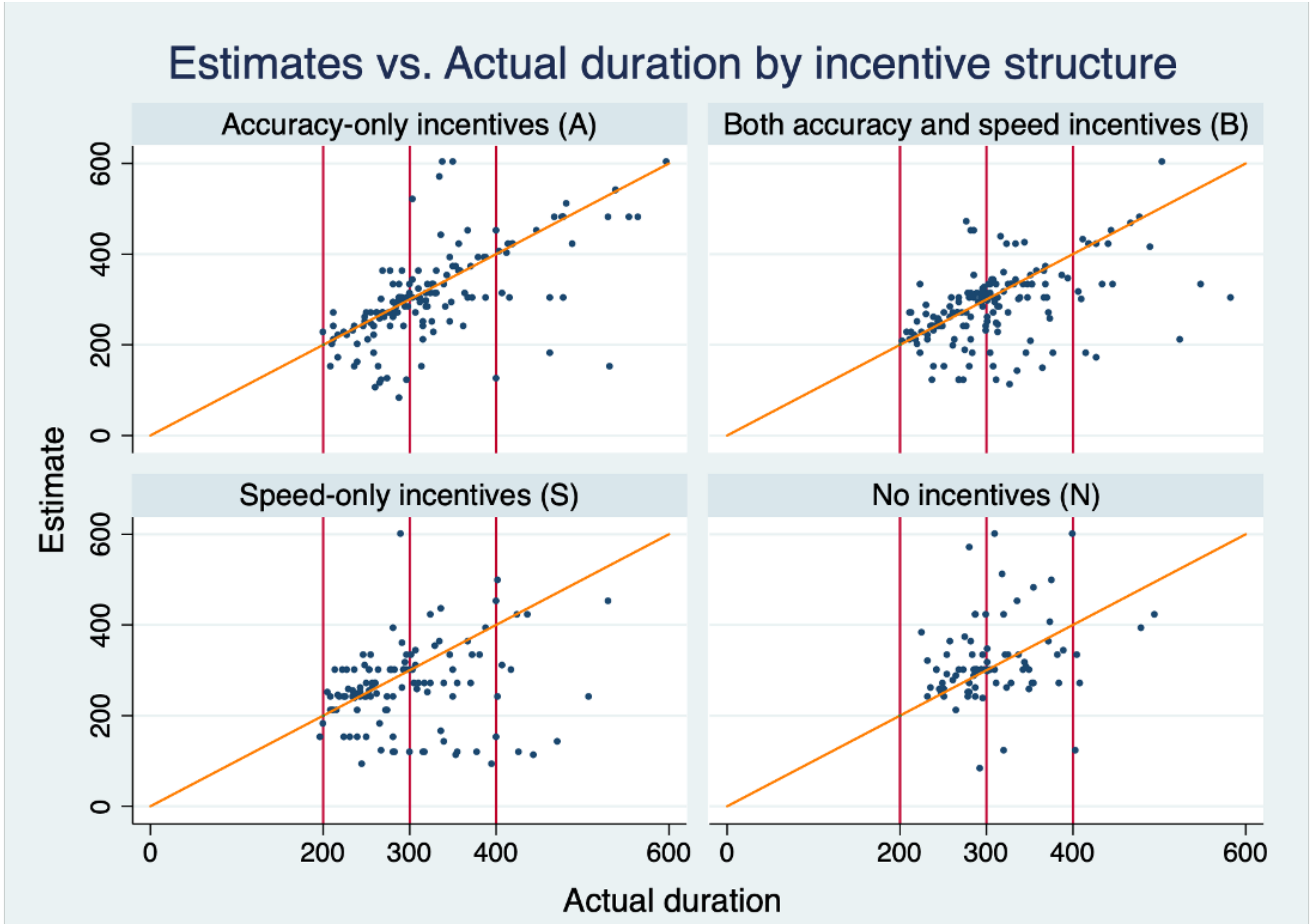
Line chart


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

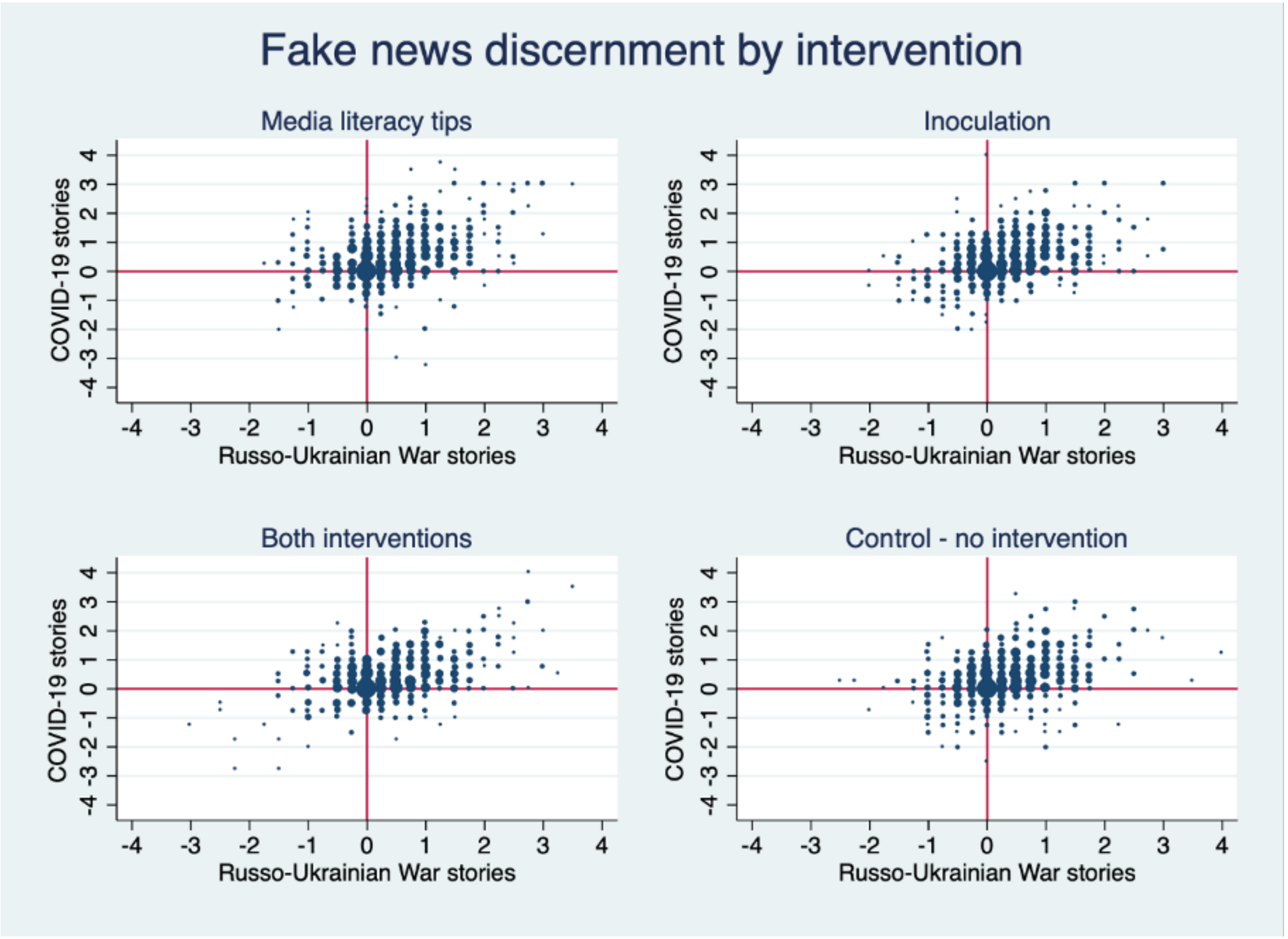
Line chart


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



Statistical inference

- In exploring the world around us, our goal may be to simply describe the phenomenon under investigation. We might ask, for example, what percentage of those surveyed were men? What was their minimum, maximum and average age? But sometimes we want to generalize our findings beyond our research set and the data we have available. This is called statistical inference. In inferential statistics, we make inferences about population-level relationships or differences based on the data collected in our research set.

- Using inferential statistics, we can ask many questions that we would like to explore, formulate a large number of different hypotheses that we would like to test, and use many different techniques and particular types of analysis. How can we simplify this? We can try to establish some general categories of inferential statistics - relational and differential statistics, and explore the dominant approach of statistical inference - testing the null hypothesis. Differential and relational statistics are two sides of the same coin, but this division can make it easier to think about some problems and bring some system to our thinking.

- For relational statistics, we will be interested in whether there is a relationship between two variables at the population level and how close that relationship is. For example, we can assume that the taller a person is, the more the person will weight. Thus, within the framework of relational statistics, we will focus on the joint variation - covariance - of the measured values. For example, we may conduct cross-sectional research where we administer a questionnaire to participants measuring five personality traits and other variables. In the context of the Big Five personality trait model, we might assume, for example, that the more open people are to experience (the higher their scores on a given dimension of the Big Five model), the more positive their attitudes will be toward less conservative methods of teaching statistics, or the higher their scores on the negative emotionality dimension, the less subjective well-being they will experience prior to enrolling in statistics course (i.e., we assume a negative relationship between negative emotionality and subjective well-being).

- In contrast, in differential statistics, we will be interested in the difference between groups or conditions. For example, we may assume that men will be taller than women. The question is posed differently here than in the previous case, and more along the lines of whether there is a difference between the groups, and how large that difference is. For example, in the context of designing interventions that improve the mental health of female and male students, we might assume that female and male statistics students who undergo training in a particular type of relaxation (e.g., Jacobson's progressive relaxation) will be better off in terms of their level of experienced anxiety than those who do not practice the relaxation method. We will design a simple experiment where we will randomly assign interested participants to experimental and control groups (between-subject experiment). With the experimental group we will practice the relaxation method once a day, while with the control group we will only meet, but will not practice relaxation. After one month, we will compare the two scores in subjectively experienced anxiety.

- However, we do not necessarily have to divide participants into groups. Instead, we may find it convenient to have them all complete all the interventions, but in a different order (a so-called within-subjects experiment). In the context of advertising effectiveness research, for example, we might assume that a humorous advertisement for a statistics book would be evaluated more positively than that of a neutral advertisement. We then show each participant/participant in our research a neutral and a funny advertisement in random order. For each of these, they will indicate how positively they rate the statistics book shown in the advertisement. Finally, we compare the product evaluations of the same participant under both the funny and neutral condition.

Pearson correlation

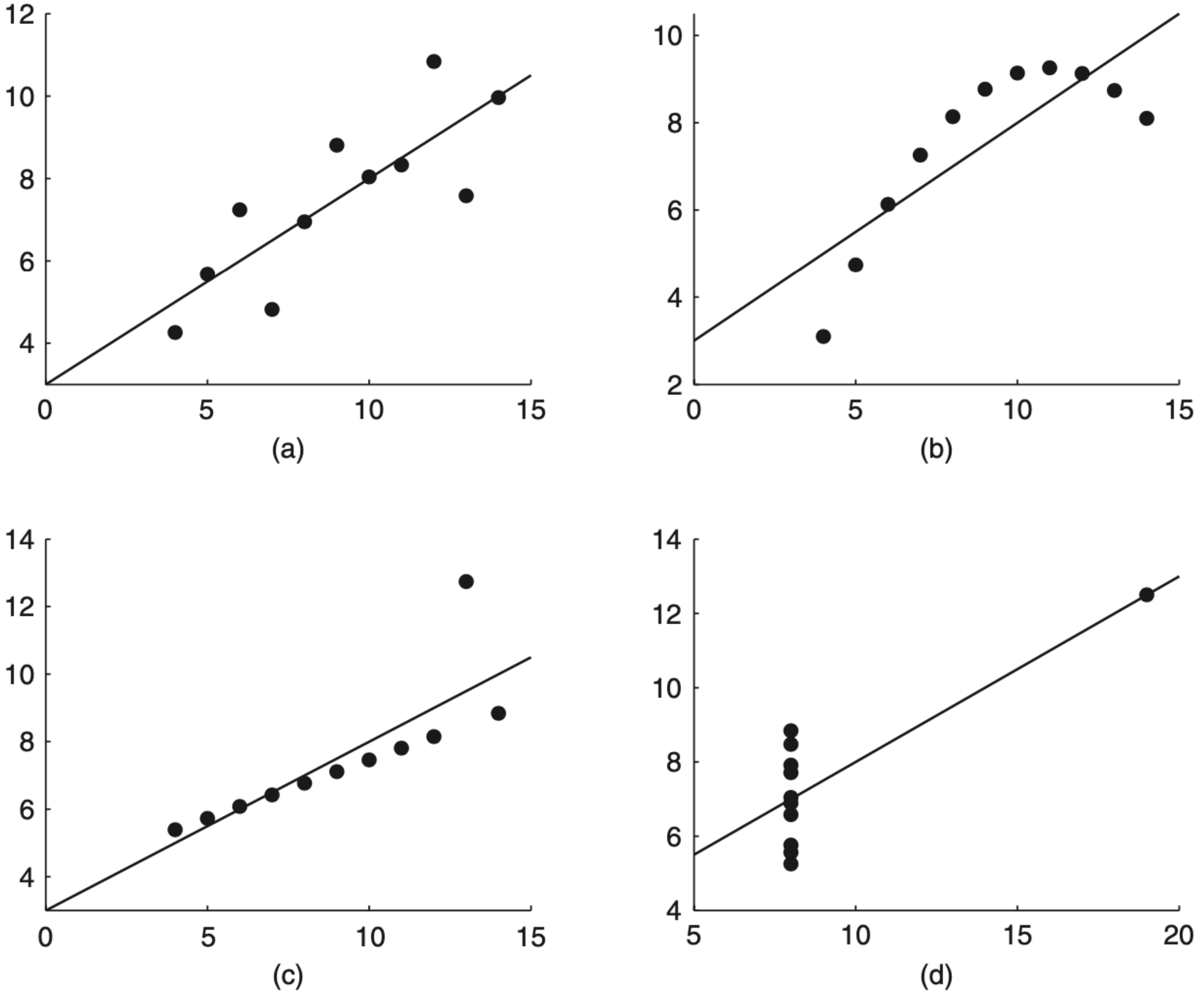
- The assessment of the degree of association between two variables is typically a measure of correlation. The basic measure here is the Pearson correlation, which measures the degree of linearity of the bivariate relationship between the two variables.

- 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


12
10
8
6
4
0
5
10
15
(a)
10
8
6
4
2
0
5
10
15
(b)
14
12
10
8
6
4
0
5
10
15
(c)
14
12
10
8
6
4
5
10
15
20
(d)