Experimental economics

Seminar VI - **Experiment from statistical perspective**

Matej Lorko
matej.lorko@euba.sk
Student resources: www.lorko.sk

References:

- Weimann, J., & Brosig-Koch, J. (2019). Methods in experimental economics. Springer International Publishing.

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 this chapter 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 deter- mined 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.

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.

Statistical tests

- In everyday life, we all too often find ourselves drawing completely unscientific invalid conclusions, such as “A friend of mine was once robbed in City A and so it is a criminal city” or “A seatbelt isn’t necessary. After all, I’ve never had an accident”.

- Even without formal analysis, we can be fairly certain these conclusions generalize far too much, since they are based on only one observation. But how can concrete statements be made about the quality of a conclusion? How certain can an experimenter be that an observed effect is not completely random?

- In such situations, tools from inferential statistics come to our assistance. The focus is on what is known as the statistical hypothesis testing. This can be used to check how consistent a general statement about the characteristics of a population is with the observed laboratory data or with the sample.

Formulating Testable Hypotheses

- The starting point of a hypothesis test is what is known as the research hypothesis. It usually postulates the content of the research question, i.e. a difference or an effect with regard to a scientifically interesting characteristic of the population under consideration.

- Basically, what we assume to be true is formulated as the null hypothesis H0 and the opposite or complement of this as the alternative hypothesis H1. The null hypothesis must therefore always include the unambiguous case of equality, leaving only the “more complicated”, indirect approach B as an option.

- This principle of statistical testing is comparable to the presumption of innocence in a court case. The initial or null hypothesis is: “The defendant is innocent.” Instead of showing directly that a defendant is guilty, more or less strong evidence that is not consistent with the innocence of the defendant is presented by the prosecutor. If this evidence is strong enough, the assumption of innocence is no longer valid and the defendant is found guilty. If, however, it is not possible to produce sufficiently strong evidence against the assumption of innocence, the defendant is not found guilty because his previously assumed innocence could not be called into question beyond reasonable doubt.

- The null hypothesis is assumed to be true until the data collected are sufficiently strong against it and it must be rejected. As soon as this is the case, the alternative hypothesis is indirectly accepted. However, if the data cannot refute the null hypothesis, it must still be assumed that it is true and the research hypothesis is not accepted. Since only the null hypothesis is tested in a hypothesis test and evidence is sought against it, a null hypothesis can only be rejected or not rejected but, strictly speaking, not accepted.

How Inferential Statistics Works

- If the research question is formulated in the form of a statistical hypothesis, hypothesis tests can be used to draw statistical conclusions regarding this hypothesis. There is, however, always a certain probability of errors.

- It must always be kept in mind that no statistical test can determine whether a hypothesis is actually true or false. Even if the test statistic of the sample is in the critical region and we come to the conclusion that the null hypothesis should be rejected, it can still be true.

- The larger we choose the critical area or the significance level, the more likely this so-called Type I error is. Now let us imagine that the null hypothesis is in fact false. In this situation, we would be making an error by not rejecting the null even though it is false (Type II error).

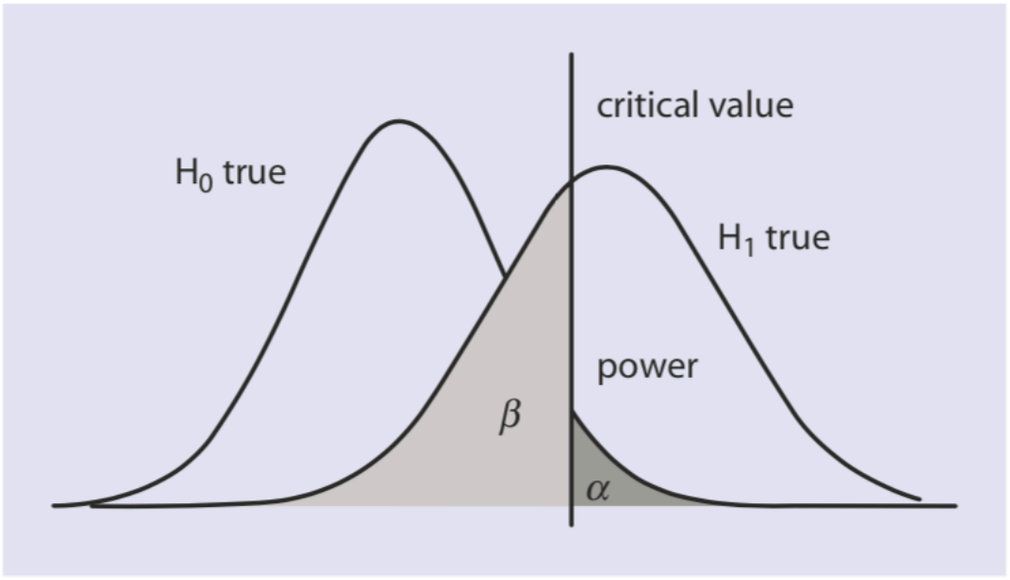
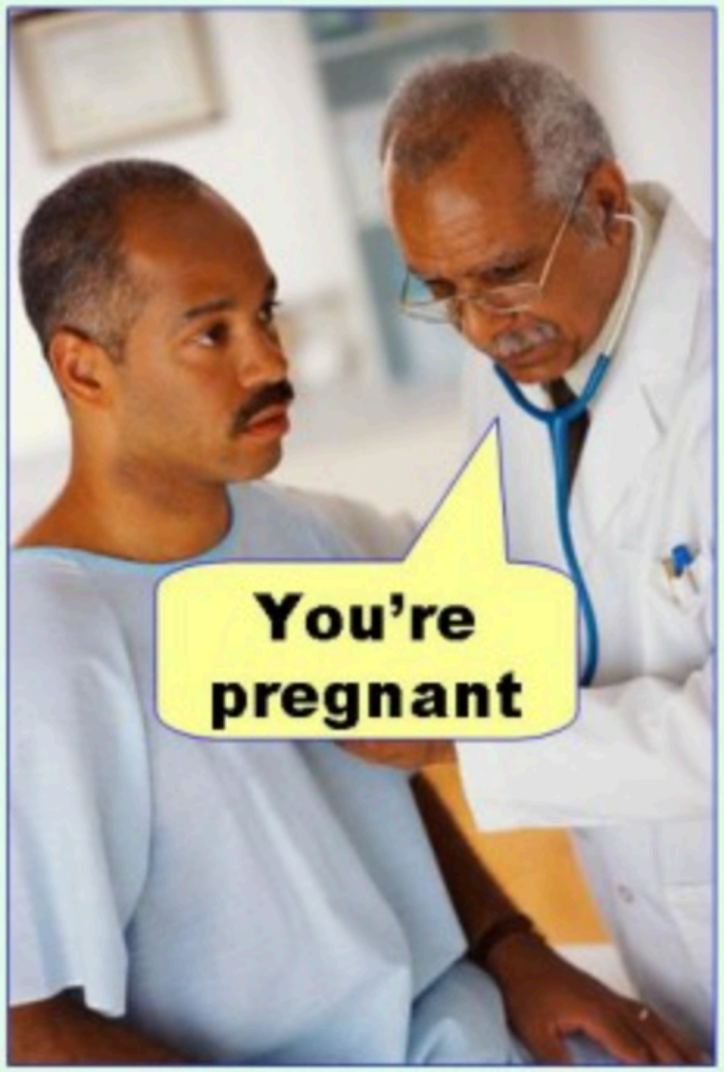


Table 4.1 Summary of error probabilities

	Truth	
	H_0 true	H_0 not true
Rejection H_0	**Type I error**	*correct*
	(Prob. α)	
Non-rejection H_0	*correct*	**Type II error**
	(Prob. $1 - \alpha$)	(Prob. $1 - \beta$)

Type I error

(false positive)



Type II error

(false negative)