Experimental economics

Seminar VII - **Statistical tests**

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.

Choosing Statistical Tests

- The “right” choice of methods for the analysis of experimental data always lies between two extremes. One extreme is a completely arbitrary decision to use a particular method of analysis, which is then applied entirely without reflection. The other extreme is the assumption that there is only one method of analysis that is perfectly suitable for each experiment. Both approaches are, of course, equally wrong. On the one hand, it is certainly possible and necessary to limit the number of methods that can be used. All experimental data have certain characteristics that rule out certain statistical analyses while allowing others to be performed. On the other hand, an experiment is never so specific that only one optimal method of analysis can be used.

- The basic approach for choosing suitable methods of statistical analysis first of all involves matching the formal requirements for the application of a method with the given characteristics of the data. All methods of inferential statistics, correlation analysis and regression analysis are based on certain assumptions.

- The main objective is therefore to avoid the most serious errors in choosing a method of statistical analysis. It is particularly important to see these considerations as part of the experimental design, which takes place before the actual experiment. Once the data are available and it is only then noticed that no suitable procedure to analyze them exists, it is usually too late for corrections.

- As a matter of principle, statistical data analysis should always be based on expert knowledge, and a statistical method should only be used if the results can provide a real insight into the research question being investigated experimentally. An ad-hoc application of a method “for the sake of the method only” should be avoided, since the statistical analysis then often misses the point of the original question.

Classifying Test Methods

- Statistical hypothesis tests can be categorized using several criteria. One of the most basic distinguishing features of statistical hypothesis tests is the number of groups or samples the test is comparing. If only one group is being examined, it is possible, for example, to test whether its mean is consistent with a certain population parameter that is assumed to be true. In this way, a comparison is made between the specific sample and a postulated true value of the population using one-sample tests.

- If, on the other hand, two groups are to be compared, for example in a classical control and treatment group comparison, it is assumed that the samples were taken from two separate populations. In this case, other tests must be used. Other tests have been developed for comparisons between more than two groups.

- As soon as several groups are to be compared, the choice of a suitable test depends on whether the groups are statistically independent of each other (unrelated or unpaired) or not (related or paired). This question is largely answered by the experimental design used. Testing individual subjects in a number of experimental conditions or groups unavoidably leads to related samples.

- By their very nature, two successive decisions of the same person cannot be independent of each other. It does not matter whether the person makes the decisions one after the other in two different treatments (cross-over design) or in one and the same treatment (longitudinal design). If, on the other hand, each subject is a decision-maker only once, it can be assumed under conditions of full anonymity and no feedback that the decision of one person does not influence the decision of another person.

Classifying Test Methods

- The third criterion influencing the choice of the statistical methods is the question as to which assumptions about the probability distribution of the variables apply. Two broad classes of methods are available, parametric and nonparametric, depending on the answer. Parametric methods only provide meaningful results if specific assumptions about the form (e.g. normal distribution) and the parameters (e.g. mean, variance, degrees of freedom) of the distribution apply.

- Sometimes finding this out is quite straightforward but in most other cases at least some uncertainty remains. As long as the sample is very large (about 100 or more), this uncertainty hardly plays a role due to the central limit theorem. Even if the true distribution is not normally distributed and a parametric test requires the normal distribution, this test will still provide reliable results for large samples. For this reason, it is said that parametric tests are robust (against deviations in the distribution) for large samples. For small samples, however, it is highly advisable to be sure that the assumptions concerning the distribution of a test are correct. Even small deviations from the assumed distribution can make a test result completely unusable.

- Nonparametric (also distribution-free) methods are an alternative to this. They do not depend on the form and the parameters of the distribution of the population from which the sample was taken. However, this does not mean, of course, that nonparametric procedures do not require any assumptions. The assumptions are only less restrictive than in the parametric case.

- As long as large samples are involved, there is no need to worry too much about which class is the better choice. A parametric test then has only a slightly higher power than its nonparametric counterpart, but the latter may be somewhat easier to perform.

How Do I Choose a Specific Test?

- For the initial choice of a statistical hypothesis test, the following criteria at least must be considered:
  - One or more groups?
  - Related or unrelated groups?
  - Parametric or nonparametric data or scales of measurement of the data?
-

Table 4.3 A simple classification of test methods. The word "test" was omitted from every name for space reasons

Design			
	1-sample	*2-sample*	
Scale		*independent/between-subject*	*dependent/within-subject*
metric	*z*, *t*	*t*	*t*
ordinal	Kolmogorov	Wilcoxon rank-sum, Mann-Whitney *U*	Wilcoxon signed-ranks
nominal/ categorical	binomial, multinomial	Fisher's exact X^2 ($2 \times k$)	McNemar

The z-Test und t-Test for One Sample

- The z-test for one sample examines whether the mean x of a random sample is suffi- ciently consistent with a given population mean μ0 that is assumed true. If the difference between x and μ0 is significant, the data do not support the hypothesis that the sample was drawn from a population with a mean μ = μ0. Accordingly, the null hypothesis is H0: μ = μ0 and the alternative hypotheses are H1: μ ≠ μ0 or H1: μ < μ 0 or H1: μ > μ 0.

- Since this is a parametric method, an important prerequisite is that the sample was taken from a normally distributed population with a known variance σ2. The sample size of a z-test should comprise at least 30 observations.

- Unlike the null distribution of the z-test, the null distribution of the t-test is different for different sample sizes, as it depends on the degrees of freedom n-1.

- Example

    - The scores of a nationwide math test are normally distributed with a mean of μ = 78 points and a standard deviation of σ = 12 points. The teacher of a particular school wants to test whether his newly introduced method of teaching math has a positive significant influence on the point score students achieve. His research hypothesis is therefore H1: μ > 78.

    - The 36 students in his course obtained an average score of x = 82 from the values 94, 68, 81, 82, 78, 94, 91, 89, 97, 92, 76, 74, 74, 92, 98, 70, 55, 56, 83, 65, 83, 91, 76, 79, 79, 86, 82, 93, 86, 82, 62, 93, 95, 100, 67, 89. The test statistic is then z = (82–78)/(12/ 36) = 2.

    - The p-value is 2.5% < 5% and we reject H0: μ = 78 at a significance level of 5%.

t-Test for Two Independent Samples (Between-Subject Comparison)

- In order to compare two samples, we need to modify the one-sample t-test. First, we assume that no one is represented in both samples at the same time and that the realizations of one sample are not in any way influenced by those of the other sample.

- The test will determine whether the means x1 and x2 of these two independently drawn samples differ so much that it can be concluded that a significant difference between the population means exists.

- If the difference between x1 and x2 is significant, the data do not support the hypothesis that the samples were taken from populations with the same mean, $\mu 1 = \mu 2$. Therefore, the null hypothesis is H0: $\mu 1 - \mu 2 = \mu 0$, generally with “no difference”, i.e. $\mu 0 = 0$, being tested. The alternativehypothesesarethenH1:$\mu 1 - \mu 2 \neq \mu 0$ orH1:$\mu 1 - \mu 2 < \mu 0$ orH1:$\mu 1 - \mu 2 > \mu 0$.

- Since we are still in the realm of parametric methods, it is necessary to assume that the each sample was randomly selected from its own normally distributed population. The two populations have the same, albeit unknown, variance $\sigma 2$, but it is not necessary for the samples to be of equal size. It is crucially important that the subjects are randomly assigned to the different treatments in a between-subject design. Only a successful randomization can ensure that selection effects can be avoided.

t-Test for Two Dependent Samples (Within-Subject Comparison)

- A further modification of the t-test is required if the realizations of one sample are not independent of those of the other sample. This is always the case in a within-subject design of an experiment, since one subject makes decisions in two different treatments or samples.

- Therefore, there are pairs of measured values in which the decision of the same subject is found in both treatments. The null hypothesis is H0: $\mu 1 - \mu 2 = \mu 0$, with $\mu 0 = 0$ usually being tested, and the alternative hypotheses are H1: $\mu 1 - \mu 2 \neq \mu 0$ or H1: $\mu 1 - \mu 2 < \mu 0$ orH1:$\mu 1 - \mu 2 > \mu 0$.

- Once again, the samples are randomly drawn from each of the normally distributed populations of unknown but equal variance $\sigma 2$. The test statistic is the same as in the two-sample case using independent samples. The standard error is calculated from a weighted mean of the sample variances, corrected by the degree of correlation between the two samples

Kolmogorov Test

- The Kolmogorov test is one of what is termed goodness-of-fit tests. These tests examine whether the distribution of the values of a sample are those that would be expected based on a specific, pre-defined distribution. This means that this test provides statistical evidence as to whether or not the assumption of a particular distribution is fulfilled.

- For this purpose, the empirical distribution function Fx of the sample, i.e. the proportion of observed x-values that are smaller than or equal to a specific x-value (for all real x-values), is compared with the pre-defined or presumed distribution function F0. The test statistic D measures the degree of agreement and is the maximum distance between Fx and F0.

- The null hypothesis postulates concordance between the theoretical and the empirical distributions, and the alternative hypothesis states that the sample does not originate from the theoretical distribution. For this reason, a two-tailed hypothesis, which allows a deviation in both directions, is usually used in practice.

- In contrast to most other tests, with the Kolmogorov test we do not want the null hypothesis to be rejected, since we usually expect the assumption concerning a particular distribution to be confirmed (e.g. normal distribution). The more dissimilar the data are to the reference distribution, the higher the probability that the null hypothesis will be rejected.

The Wilcoxon Rank-Sum Test and the Mann-Whitney U Test

- The Wilcoxon rank-sum test is a popular alternative to the t-test when it does not appear realistic to assume a normal distribution and/or the data are not scaled metrically.

- Like the t-test, it compares the equality of the “central points” of two independent samples. Arithmetic means no longer exist for these data and we generally speak of “central tendencies” to compare groups.

- An alternative method that always leads to the same test result as the Wilcoxon rank-sum test is the Mann-Whitney U test.

- Wilcoxon Signed-Rank Test (Two Dependent Samples)

- Just as the Wilcoxon rank-sum test can be seen as a nonparametric counterpart to the t-test with two independent samples, the Wilcoxon signed-rank test can be used as a nonparametric alternative to the t-test with two dependent samples. It is one of the standard tests for ordinal data in a within-subject or matched-pairs design.

- The hypotheses are the same as those in the Wilcoxon rank-sum test. If the null hypothesis is valid, it is assumed that the differences originate from a population that is symmetrically distributed around the median of 0.

The Binomial Test

- Many variables in experiments have only two possible outcomes, such as “accept offer/reject offer” in the ultimatum game, “cooperate/defect” in the prisoner dilemma game, or “choose an even number/choose an odd number” in the matching pennies game. A coin toss with the results heads or tails can also be represented by such a dichotomous variable. We call the one-off performance of such an experiment a Bernoulli trial and the two results success and failure. The probability of one of the two results of a one-off Bernoulli trial is the probability of success or failure, which is 0.5 for flipping a fair coin, for example.

- In a laboratory experiment involving decision-making, the probability of the subjects deciding on one or the other alternative action is generally not known in advance. Yet it is precisely this which is often of particular interest. If a theory specifies a particular value, the laboratory data and a suitable hypothesis test could be used to check whether the laboratory data statistically support the specific theoretical value or not.

- In the matching pennies game mentioned above, for example, game theory predicts an equilibrium in which both players play both alternatives with equal probability, i.e. with p = P(choosing an even number) = 1 – p = P(choosing an odd number) = 0.5. If this game is played sufficiently frequently in the laboratory, a relative frequency for “even number” (“success”) and “odd number” (“failure”) is obtained by simply counting the respective realizations. This frequency is also referred to as the empirical probability of success pÙ.

- The binomial test examines whether the observed value of pÙ is that which would be expected if it is assumed that in reality the probability of success takes on a specified value p = p0, which in the case of the matching pennies game is p = 0.5. If the difference between π and p0 is sufficiently large, then the null hypothesis is rejected, i.e. taking into account a given probability of error, the specified value p is not consistent with the observed sample. If, however, the null hypothesis cannot be rejected, the experimental data support the theoretical prediction.

- The variable under consideration is either dichotomous, i.e. it can by definition only have two values, such as the result of a coin toss, or it is categorically scaled with 2 categories, e.g. the amounts given in the dictator game, which are “high” if they exceed a certain amount, and otherwise “low”.

The Multinomial Test (1 × k)

- The multinomial test is the generalization of the binomial test to categorical variables with k > 2 categories. For example, it might be desirable to classify amounts given in the dictator game not only in “high” and “low”, but rather more refined in “high”, “medium” and “low”, which would correspond to a categorical variable with three categories.

- Otherwise, the test principle of the multinomial test is completely analogous to that of the binomial test. The test examines whether the empirical frequencies π1, ... , πk of the k categories are those that would be expected on the premise that in reality the probabilities of success of the categories assume certain given values p1, ... , pk (null hypothesis).

-

Fisher’s Exact Test (2 × 2)

- The multinomial test compared the frequencies of a single sample over k categories with the expected values of a reference distribution (e.g. uniform distribution over all k categories). If we now want to compare two independent, categorically scaled samples (or groups or treatments) with each other, then Fisher’s exact test offers a good solution.

- As before, the observed frequencies are first calculated and summarized in a contingency table. The rows and columns of this table contain the respective values of the two categorical variables.

- Fisher’s exact test now checks whether the frequencies are sufficiently different to indicate a significant difference between the groups. The null hypothesis assumes that the population frequencies are equal or, alternatively, that the two samples originate from the same population.

Table 4.13 Contingency table 1 for Fisher’s exact test in the example

		Measured categorical variable		
		High school diploma	No high school diploma	
Gender	Male	$x_{11} = 2$	$x_{12} = 6$	$n_1 = 8$
	Female	$x_{21} = 9$	$x_{22} = 3$	$n_2 = 12$
		$N_1 = 11$	$N_2 = 9$	$N = 20$

- Fisher’s exact test, discussed in the last section, quickly becomes impractical when the number of classes of the categorical variable or the number of observations increases. The χ2 test offers a simplifying approximation for these cases

Statistical Models

- Testing the statistical significance of a treatment effect in the form of a hypothesis is not the only reason behind many experimental studies. Collecting data on variables experimentally in order to estimate relationships between the variables is another. For this second purpose, a statistical model is developed in order to explain the data obtained as well as possible. This model can be used to answer further questions, such as:

- How can the attributes of a variable be explained using other variables and how well can this be done? What value would a variable presumably have if it were influenced by the attribute of another variable that was not elicited in the experiment?

- The starting point for a statistical model is the desire to model the changes in an experimentally observed variable y or to at least explain these changes with the help of a model. The variable y is therefore also called the variable to be explained or the endogenous variable. The information we use to explain the endogenous variable originates from one or more explanatory variables (exogenous variables). The basic assumption of each statistical model is that there is a true relationship between the two variables, but this is unknown. In particularly simple cases, it may be appropriate to assume a true, linear relationship between the endogenous variable y and exactly one exogenous variable x. This would then have the form $y = a + bx$, where the constant parameters a and b of this line are unknown.

- Such a model is always subject to certain assumptions. The most important of these are:

    - There are no relevant exogenous variables missing in the econometric model and the exogenous variables used are not irrelevant.

    - The true relationship between the exogenous variable and the endogenous variable is linear.

    - The intercept and slope parameters are constant for all the observations, i.e. they have no index t or i.

    - The disturbance is normally distributed with $u_i \sim N(0, \sigma^2)$ for all the observations i and the disturbances of all the observations i are statistically independent of each other.

    - The values of the independent variable x are statistically independent of the disturbance variable u.

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.