Foundations of Experimental Economics

University of Kassel

Winter Term 2025/2026

Lecture 7: December 5, 2025

Agenda

A quick reminder of what we will do together over the course of the next weeks:

- **Today (05.12.2025):** Statistical hypothesis testing I
- **12.12.2025:** Statistical hypothesis testing II & Regression analysis I
- **19.12.2025:** NO LECTURE!
- **16.01.2026:** Regression analysis II
- **23.01.2026:** Stated choice experiments

Objectives of this lecture

- **A quick recap** of relevant stuff
- **Understand** the differences between the different measurement scales of variables
- **Explain** what should be considered when creating a statistical design
- **Formulate** testable hypotheses
- **Understand** the differences between the various statistical tests
- **Decide** when to use which statistical test
- **Learn** to interpret the results of a statistical test
- **Explain** what types of errors might be encountered in statistical tests

Recommended readings

- Moffatt, P. (2015). *Experimetrics: Econometrics for experimental economics.* Red Globe Press. (Chapter 3, p. 35-70)

- Weimann, J., & Brosig-Koch, J. (2019). *Methods in experimental economics: An introduction*. Springer Cham. (Chapter 4, p. 169-243)

Introduction

The experiment from a statistical perspective

"To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination … he may be able to say what the experiment died of."

R. A. Fisher, Indian Statistical Congress, Sankhya, 1938

https://www.sciencephoto.com/media/225081/view/sir-ronald-fisher

The experiment from a statistical perspective

- The quote by Fischer underscores the importance of involving a statistician before (and during) the experiment, and not after it has been conducted.

- If we have a research question that we want to answer experimentally using statistical methods, then we must design the experiment so that the statistical methods of choice suit the type of data generated.

The experiment from a statistical perspective – A recap

What is a *construct valid* experiment?

- *Construct valid* experiments are experiments that produce measurable variables for the constructs of the research question being examined that best capture this construct.
- In other words: *Construct validity* refers to how well an experiment or measurement actually reflects the theoretical concept (or *construct*) it is intended to capture.
- A construct is an abstract idea like intelligence, fairness, trust, or aggression.

The experiment from a statistical perspective – A recap

Types of variables:

- *Dependent variables*, e.g. the amount offered by the first mover in an ultimatum game.
- *Explanatory variables*, e.g. whether the opponent in an ultimatum game is a computer or not.
- *Confounding variables*, e.g. time of the day (only if not explicitly controlled for), at which an experimental treatment is conducted.

The experiment from a statistical perspective – A recap

Control, randomization, and sample size:

- *Control* → All the unwanted influences that can be held constant, should be held constant.
- *Randomization* → Creating comparison groups that are homogeneous on average by leaving it to chance, which subject is placed into which group.
- *Sample size* → Ensuring a sufficient number of independent observations in a treatment, i.e. sufficiently large groups of subjects who do not systematically exhibit the same behavioral pattern.

The experiment from a statistical perspective – A recap

Scales of measurement:

- *Nominal scales* are used for categorical variables without any meaningful order, e.g. eye color (green, blue, brown, black).

- *Ordinal scales* are used for variables whose values always represent a meaningful ordering or a sequence of sorts, **but** the differences (i.e. distances) between the values of the variable are **meaningless**, e.g. Life satisfaction measures (low, medium, high).

The experiment from a statistical perspective – A recap

Scales of measurement:

- *Metric (cardinal) scales* are used for variables whose values always represent an ordering or a sequence of sorts, **and** the differences in the values of the variable give **meaningful information**.
- *Ratio scales* are metric scales that possess absolute zero points (a zero value is defined and makes sense), e.g. Minutes of a day spent watching TV.
- *Interval scales* are metric scales that do not possess absolute zero points, e.g. Body weight or height.

Random variables

- A Random variable is a numerical quantity that is generated by a random experiment, a process, or an event.
- Usually, we denote a random variable with capital letters, e.g. X and the actual values it may take with lowercase letters, e.g. x.
- There are two types of random variables
  1. A random variable is called **discrete** if its range (i.e. the set of possible values that it can take) is finite or at most countably infinite. (e.g. The number of kids in a family)
  2. A random variables is called **continuous** if it takes an infinite number of possible values in a whole interval of numbers. (e.g. The **exact** time it takes a player to make her decision in the ultimatum game.)

Random variables

- Examples of random variables:

Experiment / Process / Event	Random variable X	Possible values x	Type of random variable
Roll a die once	X = The number that appears on the die	x = 1, 2, 3, 4, 5, 6	?
Flip a coin repeatedly	X = Number of tosses until heads appears	x = HH, HT, TH, TT	?
Ultimatum game with 10 coins	X = Proposer's choice of a number of coins	x = 0, 0.5, 1, 1.5, 2, 2.5, …, 10	?
Operate a light bulb until it burns out	X = Time until the bulb burns out	$0 \leq x < \infty$	?

Probability distribution of random variables

Probability distribution of a discrete random variable

- The probability distribution of a discrete random variable X is a list of each possible value of X together with the probability that X takes that value in one trial.
- For a discrete random variable, these probabilities are captured by the probability mass function (PMF) of X, denoted by pa.
- Specifically, if a represents any possible value of X, then the probability mass of a, denoted by $p\mathrm{X}(a)$, is the probability that the event (X = a) occurs, consisting of all outcomes that give rise to a value of X equal to a :

$$p\mathrm{X}(a) \ = \ \mathrm{P}\ (\{\mathrm{X} \ = \ \mathrm{a}\})$$

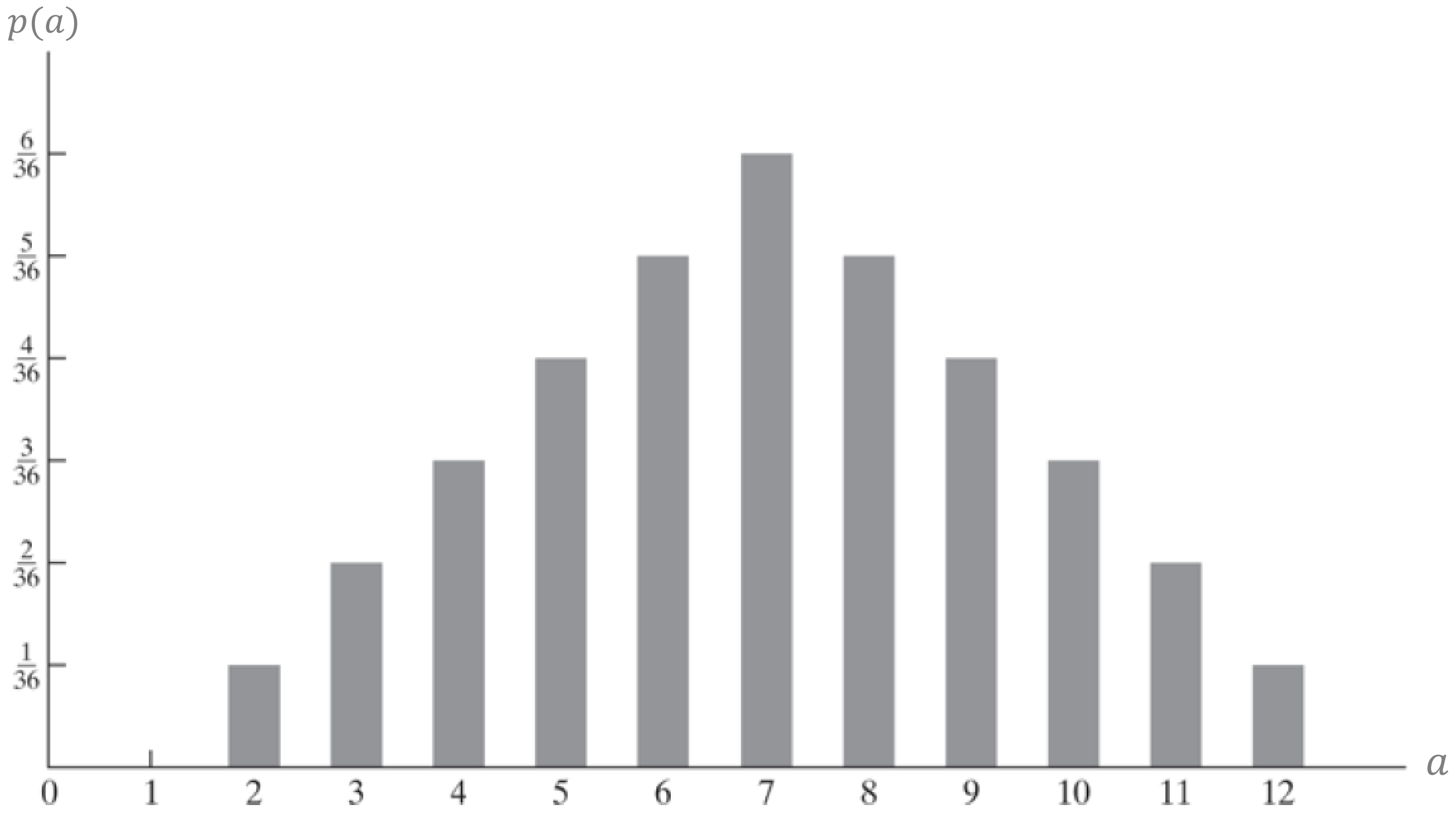
Probability distribution of random variables

- For example, let the experiment consist of rolling two fair dice, and let X (i.e. discrete random variable) be the sum when the two dice are rolled. Then the PMF of X is:

$$p\mathrm{X}(a) = \begin{cases} \frac{1}{36}, & \mathrm{a} = 2 \text{ or } \mathrm{a} = 12 \\ \frac{2}{36}, & \mathrm{a} = 3 \text{ or } \mathrm{a} = 11 \\ \frac{3}{36}, & \mathrm{a} = 4 \text{ or } \mathrm{a} = 10 \\ \frac{4}{36}, & \mathrm{a} = 5 \text{ or } \mathrm{a} = 9 \\ \frac{5}{36}, & \mathrm{a} = 6 \text{ or } \mathrm{a} = 8 \\ \frac{6}{36}, & \mathrm{a} = 7 \\ 0, & \text{otherwise.} \end{cases}$$

Probability distribution of random variables

The graph of the PMF of the sum of rolling two fair dice:



Source: Ross, S. M. (2020). *A first course in probability*. Harlow, UK: Pearson. P. 206

Probability distribution of random variables

Probability distribution of a continuous random variable

- The probability distribution of a continuous random variable X is an assignment of probabilities to intervals of decimal numbers using a non-negative function $f(x)$, called the probability density function (PDF) of X, which satisfies the following:

$$P(X \in B) = \int_{B} fX\ (x)\, dx$$

- It follows that the probability that X assumes a value in the interval $[a, b]$ is equal to the area of the region that is bounded above by the graph of the equation $y = f(x)$, bounded below by the x-axis, and bounded on the left and right by the vertical lines a and b, i.e.

$$P(a \leq X \leq b) = \int_{a}^{b} fX(x)\, dx$$

Probability distribution of random variables

Probability distribution of a continuous random variable

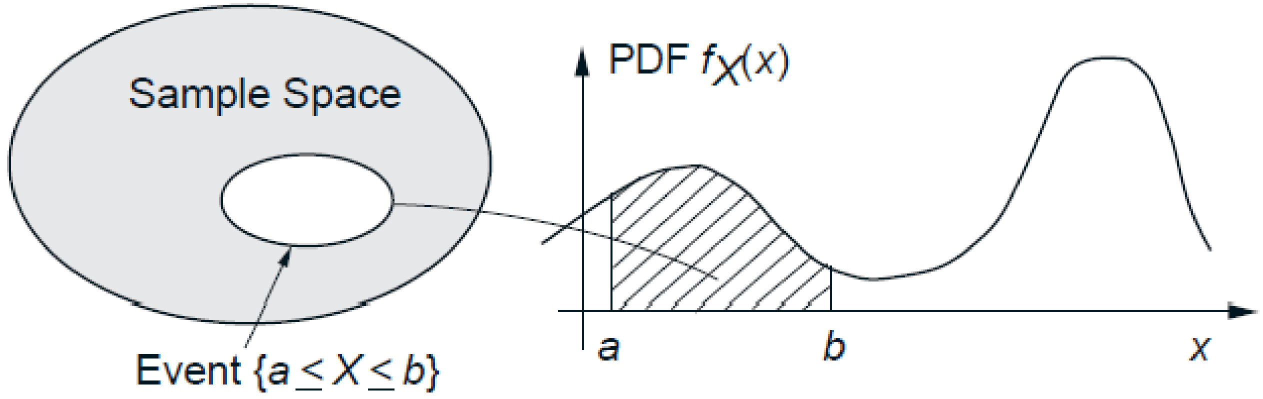


Figure 3.1: Illustration of a PDF. The probability that X takes value in an interval $[a, b]$ is $\int_a^b f_X(x)\,dx$, which is the shaded area in the figure.

Source: Bertsekas, D., & Tsitsiklis, J. N. (2008). *Introduction to probability* (*Vol. 1*). Athena Scientific. P. 206

5-minute exercise

- **Assign the following variables to their appropriate measurement scales:**

  - Religion ➡
  - Education level ➡
  - Age of a person ➡
  - IQ test score ➡
  - Number of children ➡

- **Discuss which measurement scale is most suitable for the variable "Temperature in Celsius". Why?**

5-minute exercise

- **Discuss which measurement scale is most suitable for the variable “Temperature in Celsius”. Why?**

Creating the Statistical Design

Creating the statistical design

- Compiling the observation units, i.e. sampling:

  1. Decide from where the sample should be drawn
  2. Determine how many subjects should be drawn (i.e. sample size)
  3. Determine the type of sampling, i.e. probability sampling vs. non-probability sampling

Creating the statistical design

1. Decide from where the sample should be drawn:
   - In every experiment, the set of subjects, from which a sample is drawn, is called the *population*.
   - For example, for most experiments that take place at universities, the largest possible population is the number of all enrolled students.
   - The smaller and more specific the population is, the more specific the statistical statements that can be made about this population.
   - However, this means that such statements cannot be generalized well to larger populations, i.e. the external validity of the results is low.

Creating the statistical design

2. Determine how many subjects should be drawn (i.e. sample size):
   - Ideally, not too large sample and not too small samples. However, in practice the sample size is often determined by the available budget.
   - For example, in neurosciences the lab experiments are extremely expensive so that samples are often very small. This, however, can be problematic from the point of view of inferential statistics since the probability that a statistical hypothesis test correctly identifies an actual effect as present decreases dramatically with smaller samples.
   - On the other hand, too large samples can also be problematic since they can make statistical hypothesis test too sensitive, i.e. even the smallest, probably scientifically irrelevant effects can become statistically significant.

Creating the statistical design

3. Probability vs. non-probability sampling:

   I. **Probability sampling**: The probability of a subject being drawn can be specified.

   II. **Non-probability sampling**: The probability of a subject being drawn cannot be specified.

Creating the statistical design

I. Probability sampling:

– There are three variants of probability sampling:

1. Random sampling
2. Systematic random sampling
3. Stratified sampling

1. In **random sampling**, n subjects are randomly selected from a population of size N. The probability of each subject being drawn is thus $1/N$, i.e. every subject has the same likelihood of being drawn. However, if this is not the case, then the sample is not representative of the population (biased sample).
2. In **systematic random sampling**, every tenth subject is drawn from a randomly ordered list of subjects. Thus, every person has a selection probability of 1/10. However, this variant works only if the list of subjects or population itself is not subject to any order.

Creating the statistical design

I. Probability sampling:

3. In **stratified sampling**:

– The population is first divided into subpopulations (i.e. strata), with the subjects in each stratum having at least one common characteristic that distinguishes them from the subjects of the other strata.

– 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.

– However, stratified sampling only works if the most important characteristics of all the subjects of the entire population are known, so that suitable strata can be formed.

Creating the statistical design

II. Non-probability sampling:

- Non-probability sampling always leads to biased (i.e. unrepresentative) samples since not every subject has the same probability of being selected.
- In **convenience samples** only those subjects who can easily be reached by the experimenter are invited.
- In **voluntary response samples**, participation is on the subject’s own voluntary initiative. Such samples are also unrepresentative since they suffer from self-selection bias.

Creating the statistical design

- **Classifying statistical designs of experiments:**

    1. Completely randomized design (CRD)

    2. Randomized block design (RBD)

    3. Repeated measure design (RMD)

Creating the statistical design

1. Completely randomized design (CRD):
   - The assignment of subjects is completely random across all groups.
   - It is advisable to divide the subjects into equally large groups. However, it is not an insoluble problem if the groups are not equally large.
   - CRDs can theoretically be applied to any number of treatments, they allow unequal group sizes, and they are well-suited for relatively simple statistical methods of analysis.
   - However, this design shall not be used if the experimenter is unsure about the success of the randomization process. This is, e.g., the case when many treatments are tested in small groups or when the subjects are highly heterogeneous.

Creating the statistical design

2. Randomized block design (RBD):
   - RBDs are recommended when the subjects are highly heterogeneous with respect to a measurable confounding factor.
   - In RBDs the subjects are divided into “blocks” according to their various characteristics. In each block, the similarity of two subjects within a block must be greater than that of two subjects between two blocks.
   - It is important to assign the subjects to the control and treatment groups in the same way within each block.

Creating the statistical design

3. Repeated measure design (RMD):
   - In RMDs each subject undergoes several measurements, either in one and the same treatment at different times (i.e. longitudinal design) or in different treatments and also at different times (cross-over design). The sequence of treatments a subject goes through is, again, randomized.
   - RMDs entail lower costs due to fewer subjects, higher statistical power than comparable between-subject designs due to lower error spread, and the possibility of measuring treatments over time.
   - On the other hand, RMDs involve considerably more complex methods of analysis.

Formulating Testable Hypotheses

Formulating testable hypotheses

- Formulating testable hypotheses:
  - The starting point of a hypothesis test is the so-called *research hypothesis*. It usually postulates the content of the research question.
  - For example, in the ultimatum game we could assert that:
  - *RH1: Northern Germans do not offer exactly half of the endowment on average, but either more or less than half.*
  - *RH2: Northern German men offer less on average than northern German women do.*

Formulating testable hypotheses

- Formulating testable hypotheses:
  - There are basically two possible approaches to testing the research hypothesis:
  - **Approach A**: We assume that the research hypothesis is true and try to find evidence in favor of it
  - **Approach B**: We assume that the opposite of the research hypothesis is true and try to find evidence against the opposite of the research hypothesis.
  - What we assume to be true is formulated as the ***null hypothesis*** H_0 and the opposite (or complement) of this is formulated as the ***alternative hypothesis*** H_1.

Formulating testable hypotheses

- Formulating testable hypotheses:
  - For our **first research hypothesis** from the previous example (i.e. *RH1: Northern Germans do not offer exactly half of the endowment on average, but either more or less than half),* this means:
  - **Approach A**:
    1. H_0: $\mu \neq 0.5$ (H_0, i.e. research hypothesis, assumed to be true)
    2. H_1: $\mu = 0.5$ (alternative hypothesis)
    (Target: Do not reject H_0)

  - **Approach B**:
    1. H_0: $\mu = 0.5$ (H_0 assumed to be true)
    2. H_1: $\mu \neq 0.5$ (research hypothesis)
    (Target: Reject H_0)

Formulating testable hypotheses

- Formulating testable hypotheses:
  - For our **second research hypothesis** from the previous example (i.e. *RH2: Northern German men offer less on average than northern German women do),* this means:
  - **Approach A**:
    H_0: $\mu_m - \mu_f < 0$ versus H_1: $\mu_m - \mu_f \geq 0$
    H_0 (i.e. research hypothesis) is assumed to be true, and the target is to **not reject** H_0.
  - **Approach B**:
    H_0: $\mu_m - \mu_f \geq 0$ versus H_1: $\mu_m - \mu_f < 0$
    H_0 is assumed to be true, H_1 is our research hypothesis, and the target is to **reject** H_0.
  - Note that H_1 in approach A (i.e. H_0 in approach B) can be alternatively formulated as follows: ($\mu_m - \mu_f = 0$). Why?

Formulating testable hypotheses

- Formulating testable hypotheses:
  - For *RH1*, we have formulated a so-called two-tailed hypothesis (also two-sided or non-directional). Why?
  - In contrast, for *RH2*, we have formulated a one-tailed hypothesis (also one-sided or directional). More precisely, RH2 is a left-tailed hypothesis. Why?
  - Which approach should be preferred? Why?

Formulating testable hypotheses

- Understanding hypothesis testing:
  - So, as we saw, the null hypothesis is always the starting point. It usually says that there is no relationship, no difference, or no effect.
  - What we do is: We test this null hypothesis to see if our data give us enough evidence to reject it.
  - Whether we reject or fail to reject a null hypothesis can be determined through two main criteria:
    1. *p-values*
    2. *Test statistics*

Formulating testable hypotheses

- Understanding hypothesis testing:
  - Each hypothesis test produces, among other things, a test *p*-value and a test statistic.
  - The smaller the test *p*-value, the less likely our obtained data are under the null hypothesis. Thus, we consider rejecting the null hypothesis.
  - The larger the test statistic, the less likely our obtained data are under the null hypothesis. Thus, we consider rejecting the null hypothesis.
  - But how small or large should the p-value and test statistic be?

Formulating testable hypotheses

- Understanding hypothesis testing:

Common significance levels

***P*-values**	**test statistics**	**Inference**
$p > 0.10$	$t < 1.645$	We have no statistical evidence to reject the null hypothesis at common significance levels
$0.05 < p < 0.10$	$1.645 < t < 1.96$	We have statistical evidence to reject the null hypothesis at the 10% significance level
$0.01 < p < 0.05$	$1.96 < t < 2.576$	We have statistical evidence to reject the null hypothesis at the 5% significance level
$0.001 < p < 0.01$	$2.576 < t < 3.291$	We have statistical evidence to reject the null hypothesis at the 1% significance level
$p < 0.001$	$t > 3.291$	We have statistical evidence to reject the null hypothesis at the 0.1% significance level

Formulating testable hypotheses

- Possible errors:
  - As clear as our test decision might be (i.e. reject H_0 or not), it must always be kept in mind that no statistical test can determine whether a hypothesis is *actually* true or false.
  - With that being said, there will always be a certain probability of errors.
  - However, the smaller the critical *p*-value we choose, the smaller this probability of error becomes (but it will never be zero).
  - There are two types of errors that we may encounter, Type I error and Type II error:

Formulating testable hypotheses

- Possible errors:

Decision	Null hypothesis is …	
	True	**False**
Reject H_0	Type I error (false positive)	Correct decision (true positive)
Fail to reject H_0	Correct decision (true positive)	Type II error (false negative)

Formulating testable hypotheses

- Possible errors:
  - If our null hypothesis is **true**, yet we mistakenly **reject** it, then we have made a **type I error** (also called **false positive**).
  - If our null hypothesis is **false**, yet we **fail to reject** it, then we have made a **type II error** (also called **false negative**).

5-minute exercise

- Think of an example, with which you can illustrate type I and type II errors.
- Formulate the null and alternative hypotheses
- Explain when a type I or type II error might occur

Choosing Statistical Tests

Choosing statistical tests

- Classifying statistical tests:
  - For the initial choice of a statistical hypothesis test, we must consider the following criteria:

1. How many groups or samples is the test method comparing?
2. In case multiple groups are to be compared, are they all statistically independent of each other?
3. Parametric or non-parametric tests?

Choosing statistical tests

1. How many groups or samples is the test method comparing?

– If only one group is being examined, then it is possible, for example, to test whether its mean is consistent with a certain population parameter that is assumed to be true.

– If, however, two groups are being compared (e.g. the classical control vs. treatment), then it is assumed that the two groups are taken from two separate populations and, thus, other tests must be used.

– For more than two groups, also other tests must be used.

Choosing statistical tests

2. In case multiple groups are to be compared, are they all statistically independent of each other?

– As soon as several groups are to be compared, then the choice of a suitable test depends on whether the groups are statistically paired (i.e. related or dependent on each other) or whether they are unpaired (i.e. unrelated or independent of each other).

– For example, two successive decisions of the same person can never be independent of each other.

– On the other hand, if each subject makes a decision only once, then it can be assumed (under conditions of full anonymity and no feedback) that the decision of one subject does not affect the decision of another subject.

Choosing statistical tests

3. Parametric or non-parametric tests?

– The choice of a parametric or non-parametric test depends on the probability distribution of the variables.

– Parametric tests only provide meaningful results if certain assumptions about the form (e.g. normal distribution) and the parameters (e.g. mean, variance) of the distribution apply.

– As long as the examined sample is large enough (100 or more), then parametric tests are robust against deviations in the distribution.

Choosing statistical tests

3. Parametric or **non-parametric** tests? (*continued*)

– For small samples, we must be sure that the distribution assumptions of a test are correct. Otherwise, our test results will mostly be unusable.

– If we are not sure, then we can use non-parametric tests (i.e. distribution-free), since they do not depend on the form and the parameters of the distribution of the population from which the sample was taken.

Choosing statistical tests

Design			
Sample / **Scale**	One sample	Two samples	
	/	Independent / between-subject	Dependent / within-subject
Metric	*z*-test, *t*-test	*t*-test	*t*-test
Ordinal	Kolmogorov-Smirnov test	Wilcoxon rank-sum test (Mann-Whitney *U* test)	Wilcoxon signed-ranks test
Nominal / Categorical	Binomial test	Fisher's exact test, χ^2 (2 x k)	McNemar

Choosing statistical tests

I. The ***z*-test** and the ***t*-test** for one sample:

- Both are **parametric** tests that examine whether the mean $\bar{x}$ of a random sample is sufficiently consistent with a given population mean μ_0 that is assumed to be true.
- If the difference between $\bar{x}$ and μ_0 is significant, the data do not support the hypothesis that the sample was drawn from a population with a mean $\mu = \mu_0$.
- Thus, the null hypothesis is: H_0: $\mu = \mu_0$, and
- the alternative hypothesis is: H_1: $\mu \neq \mu_0$ (or: H_1: $\mu < \mu_0$ or: H_1: $\mu > \mu_0$)
- Since these are parametric tests, it is a prerequisite that the sample was drawn from a normally distributed population.

Choosing statistical tests

II. The ***z*-test** and the ***t*-test** for one sample. When to use which test?

– Use the *z*-test when:
(a) the variance σ^2 of the population from which the sample was drawn is **known**, **(b)** the sample size comprises **at least 30** observations, **(c)** the variable is **metric**, and **(d)** only **one sample** is considered.

– Use the *t*-test when:
(a) the variance σ^2 of the population from which the sample was drawn is **NOT known**, **(b)** the variable is **metric**, and **(c)** only **one sample** is considered.

– Note that the *t*-test has no restrictions on sample sizes as it does well with any sample size.

Choosing statistical tests

III. The ***z-test*** for one sample. Example:

```
. ztest mpg == 20, sd(6)

One-sample z test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
     mpg |      74     21.2973    .6974858           6    19.93025    22.66434
------------------------------------------------------------------------------
    mean = mean(mpg)                                              z =   1.8600
H0: mean = 20

    Ha: mean < 20               Ha: mean != 20               Ha: mean > 20
 Pr(Z < z) = 0.9686        Pr(|Z| > |z|) = 0.0629         Pr(Z > z) = 0.0314
```

Choosing statistical tests

IV. The ***t*-test** for one sample. Example:

```
. ttest mpg == 20

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
     mpg |      74     21.2973    .6725511    5.785503     19.9569    22.63769
------------------------------------------------------------------------------
    mean = mean(mpg)                                              t =   1.9289
H0: mean = 20                                    Degrees of freedom =       73

    Ha: mean < 20                Ha: mean != 20                 Ha: mean > 20
 Pr(T < t) = 0.9712         Pr(|T| > |t|) = 0.0576          Pr(T > t) = 0.0288
```

Choosing statistical tests

V. The ***z*-test** and the ***t*-test** for one sample. Interpretation of the results:

- ➢The *p*-value of the two-sided *z*-test is 0.0629 < 0.10. Thus, we have statistical evidence to reject the null hypothesis (that the mean equals 20) at the 10% significance level.
- ➢The *p*-value of the two-sided t-test is 0.0576 < 0.10. Thus, we have statistical evidence to reject the null hypothesis (that the mean equals 20) at the 10% significance level.

Choosing statistical tests

I. The ***t*-test** for two **independent** samples (i.e. between-subject comparison):

- The *t*-test for two independent samples is a modification of the *t*-test for one sample. Here, we want to examine whether the means $\bar{x}_1$ and $\bar{x}_2$ of two independently drawn samples are equal.
- If the difference between $\bar{x}_1$ and $\bar{x}_2$ is significant, the data do not support the hypothesis that the two sample were drawn from two populations with the same mean $\mu_1 = \mu_2$.
- Thus, the null hypothesis is: H_0: $\mu_1 - \mu_2 = \mu_0$ and
- the alternative hypothesis is: H_1: $\mu_1 - \mu_2 \neq \mu_0$ (or: H_1: $\mu_1 - \mu_2 < \mu_0$ or: H_1: $\mu_1 - \mu_2 > \mu_0$)
- Since this is also a parametric test, it is still a prerequisite that each of the samples was drawn from its own normally distributed population.

Choosing statistical tests

II. The ***t*-test** for two **independent** samples (i.e. between-subject comparison). When to use?

– Use the *t*-test when:
(a) the two samples are **unpaired** (i.e. **independent** of each other),
(b) the two samples have the **same** (but unknown) **variance** σ^2,
(c) the subjects are **randomly assigned** to the samples, and
(d) the variables are **metric.**

– Note that it is not necessary for the two samples to be of equal size.

Choosing statistical tests

III. The ***t*-test** for two **independent** samples (i.e. between-subject comparison). Example:

```
. ttest bp_before, by(sex)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
    Male |      60    159.2667    1.473469    11.41344    156.3183    162.2151
  Female |      60    153.6333     1.38596     10.7356      150.86    156.4066
---------+--------------------------------------------------------------------
Combined |     120      156.45    1.039746    11.38985    154.3912    158.5088
---------+--------------------------------------------------------------------
    diff |            5.633333    2.022868                1.627503    9.639163
------------------------------------------------------------------------------
    diff = mean(Male) - mean(Female)                              t =   2.7848
H0: diff = 0                                     Degrees of freedom =      118

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9969         Pr(|T| > |t|) = 0.0062          Pr(T > t) = 0.0031
```

Choosing statistical tests

IV. The ***t*-test** for two **independent** samples (i.e. between-subject comparison). Interpretation of the results:

- The *p*-value of the two-sided *t*-test is $0.0062 < 0.01$. Thus, we have statistical evidence to reject the null hypothesis (that the means of the two groups are equal to each other) at all common significance levels (i.e. 10%, 5%, and 1%).

Choosing statistical tests

I. The ***t*-test** for two **dependent** samples (i.e. within-subject comparison):

– The t-test for two dependent samples is a further modification of the t-test for one sample. Here, we want to examine whether the means $\bar{x}_1$ and $\bar{x}_2$ of two related samples are equal.

– If the difference between $\bar{x}_1$ and $\bar{x}_2$ is significant, the data do not support the hypothesis that the two sample were drawn from two populations with the same mean $\mu_1 = \mu_2$.

– Thus, the null hypothesis is: H_0: $\mu_1 - \mu_2 = \mu_0$, and

– the alternative hypothesis is: H_1: $\mu_1 - \mu_2 \neq \mu_0$ (or: H_1: $\mu_1 - \mu_2 < \mu_0$ or: H_1: $\mu_1 - \mu_2 > \mu_0$)

– Since this is also a parametric test, it is still a prerequisite that each of the samples was drawn from its own normally distributed population.

Choosing statistical tests

II. The ***t*-test** for two **dependent** samples (i.e. within-subject comparison). When to use?

– Use the *t*-test when:
(a) the two samples are **paired** (i.e. **dependent** on each other),
(b) the two samples have the **same** (but unknown) **variance** σ^2,
(c) the subjects are **randomly assigned** to the samples, and
(d) the variables are **metric.**

– Note that it is not necessary for the two samples to be of equal size.

Choosing statistical tests

III. The ***t*-test** for two **dependent** samples (i.e. within-subject comparison). Example:

```
. ttest bp_before == bp_after

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
bp_bef~e |     120      156.45    1.039746    11.38985    154.3912    158.5088
bp_after |     120    151.3583    1.294234    14.17762    148.7956     153.921
---------+--------------------------------------------------------------------
    diff |     120    5.091667    1.525736     16.7136    2.070557    8.112776
------------------------------------------------------------------------------
     mean(diff) = mean(bp_before - bp_after)                      t =   3.3372
 H0: mean(diff) = 0                              Degrees of freedom =      119

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9994         Pr(|T| > |t|) = 0.0011          Pr(T > t) = 0.0006
```

Choosing statistical tests

IV. The ***t*-test** for two **dependent** samples (i.e. within-subject comparison). Interpretation of the results:

- The *p*-value for the two-sided *t*-test is $0.0011 < 0.01$. Thus, we have statistical evidence to reject the null hypothesis (that the means of the two groups are equal to each other) at all common significance levels (i.e. 10%, 5%, and 1%).

Choosing statistical tests

I. The **Kolmogorov-Smirnov test** for one sample:

– The Kolmogorov-Smirnov test for one sample is a **non-parametric "goodness-of-fit"** test. It compares the empirical distribution of a sample with a fully specified and a pre-defined theoretical (reference distribution) (e.g. normal, uniform, …etc.).

– Thus, the null hypothesis is: H_0: $F(x) = F_0(x)$ for all x, i.e. *The sample comes from the reference distribution*

– The alternative hypothesis is: H_1: $F(x) \neq F_0(x)$ for at least one x, i.e. *The sample does **not** come from the reference distribution*

Where: $F(x) =$ the cumulative distribution function (CDF) of the population

$F_0(x) =$ the theoretical cumulative distribution function we are testing against

Choosing statistical tests

III. The **Kolmogorov-Smirnov test** for one sample. When to use?

– Use the Kolmogorov-Smirnov test when:

  (a) you have a continuous random variable (with at least ordinal values), as the test does not work well with discrete random variables (without further modifications) and

  (b) you do not assume normality (recall the conditions for when to use non-parametric tests.

Choosing statistical tests

IV. The **Kolmogorov-Smirnov test** for one sample. Example:

```
. sum x

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
           x |          7    4.571429    3.457222          0         10

. ksmirnov x = normal((x-4.571429)/3.457222)

One-sample Kolmogorov-Smirnov test against theoretical distribution
     normal((x-4.571429)/3.457222)

Smaller group                 D      p-value
---------------------------------------------
x                        0.1650        0.683
Cumulative              -0.1250        0.803
Combined K-S             0.1650        0.991
```

Choosing statistical tests

V. The **Kolmogorov-Smirnov test** for one sample. Interpretation of the results:

- Here, we have three *p*-values. The third *p*-value is the one we are interested in. Since this *p*-value 0.991 > 0.1, then we do not have statistical evidence to reject the null hypothesis (that our variable (x) comes from a normal distribution with the pre-defined mean and standard deviation) at any common significance level.

Choosing statistical tests

I. The **Wilcoxon rank-sum test (Mann-Whitney *U* test)** for **two independent** samples:

- The Wilcoxon rank-sum test is the **non-parametric** alternative to the t-test for two independent samples. Like the t-test, it examines whether the “central points” (not the means anymore, why?) of two independently drawn samples are equal.
- Thus, the null hypothesis is: H_0: *The samples were drawn from populations with the same distribution*
- The alternative hypothesis is: H_1: *The samples were **not** drawn from populations with the same distribution*

Choosing statistical tests

II. The **Wilcoxon rank-sum test (Mann-Whitney *U* test)** for **two independent** samples. When to use?

– Use the test when:

(a) the samples are independent (i.e. unpaired),

(b) you have a continuous random variable that is measured at least on an ordinal scale, and

(c) you do not assume normality (recall the conditions for when to use non-parametric tests.

Choosing statistical tests

III. The **Wilcoxon rank-sum test (Mann-Whitney *U* test)** for **two independent** samples. Example:

```
. ranksum bp_before, by(sex)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

         sex |      Obs    Rank sum    Expected
-------------+---------------------------------
        Male |       60      4168.5        3630
      Female |       60      3091.5        3630
-------------+---------------------------------
    Combined |      120        7260        7260

Unadjusted variance    36300.00
Adjustment for ties      -44.75
                     ----------
Adjusted variance      36255.25

H0: bp_bef~e(sex==Male) = bp_bef~e(sex==Female)
         z =   2.828
Prob > |z| =  0.0047
Exact prob =  0.0044
```

Choosing statistical tests

IV. The **Wilcoxon rank-sum test (Mann-Whitney *U* test)** for **two independent** samples. Interpretation of the results:

- Since the exact *p*-value 0.0044 < 0.01, then we have statistical evidence to reject the null hypothesis (that the distribution of blood pressure values is the same for males and females (i.e. the two groups are drawn from populations with the same distribution)) at all common significance levels.

Choosing statistical tests

I. The **Wilcoxon signed-rank test** for **two dependent** samples:

- The Wilcoxon signed-rank test is the **non-parametric** alternative to the t-test for two dependent samples. The test examines whether the median difference between the two paired samples is zero (i.e. no systematic difference).
- Thus, the null hypothesis is: H_0: *The median difference between paired observations is zero*
- The alternative hypothesis is: H_1: *The median difference between paired observations is **not** zero*

Choosing statistical tests

II. **Wilcoxon signed-rank test** for **two dependent** samples. When to use?

– Use the test when:

(a) the samples are related (i.e. paired),

(b) you have a continuous random variable that is measured at least on an ordinal scale, and

(c) you do not assume normality (recall the conditions for when to use non-parametric tests.

Choosing statistical tests

III. The **Wilcoxon signed-rank test** for **two dependent** samples. Example:

```
. signrank bp_before = bp_after

Wilcoxon signed-rank test

        Sign |      Obs   Sum ranks    Expected
-------------+---------------------------------
    Positive |       73      4843.5        3625
    Negative |       43      2406.5        3625
        Zero |        4          10          10
-------------+---------------------------------
         All |      120        7260        7260

Unadjusted variance   145805.00
Adjustment for ties      -73.88
Adjustment for zeros      -7.50
                     ----------
Adjusted variance     145723.62

H0: bp_before = bp_after
         z =   3.192
Prob > |z| = 0.0014
Exact prob = 0.0013
```

Choosing statistical tests

IV. The **Wilcoxon signed-rank test** for **two dependent** samples. Interpretation of the results:

- Since the exact *p*-value $0.0013 < 0.01$, then we have statistical evidence to reject the null hypothesis (that the median difference between the two groups is zero) at all common significance levels.

Choosing statistical tests

I. The **binomial test** for **one** sample:

– Many variables in experiments have only two possible outcomes, such as “accept offer/reject offer” in an ultimatum game or “cooperate/defect” in the prisoner’s dilemma game. Can you think of other similar examples?

– We refer to such variables as “dichotomous variables”.

– A single, random experiment with exactly two possible outcomes is referred to as a “Bernoulli trial” (also Binomial trial), and the two results (i.e. outcomes) are called “success” and “failure”.

– The probability that one of these two results occurs in a Bernoulli trial is referred to as the probability of success or failure.

Choosing statistical tests

II. The **binomial test** for **one** sample:

- In a lab experiment involving decision-making, the probability of the subjects deciding on one or the other alternative action is generally not known in advance. However, it is precisely this probability which is often of particular interest.
- If a theory specifies a particular value, the lab data and a suitable hypothesis test could be used to check whether the data generated in the lab statistically support the pre-specified theoretical value or not.
- Consider, for example, the matching pennies game. In this game, player A and player B both have a penny each and must secretly turn their penny to either heads or tails. The players then reveal their choices simultaneously. If the pennies match (i.e. both heads or both tails), player A wins and can keep both pennies to herself. If the pennies do not match (i.e. one heads and one tails), player B wins and can keep both pennies to herself.

Choosing statistical tests

III. The **binomial test** for **one** sample:

– Game theory predicts an equilibrium in which both players play both alternatives with equal probability, i.e.:

$$p = P(choosing\ heads) = 1 - p = P(choosing\ tails) = 0.5$$

– If this game is played sufficiently frequently in the lab, a relative frequency for "choosing heads" (i.e. success) and "choosing tails" (i.e. failure) is obtained by simply counting the corresponding realizations.

– This frequency is referred to as the "*empirical probability of success* $\hat{\pi}$".

Choosing statistical tests

IV. The **binomial test** for **one** sample:

- Accordingly, the binomial test (which is a **non-parametric** test) examines whether the observed value of the empirical probability of success $\hat{\pi}$ in a Bernoulli trial is equal to a pre-specified value $p = p_0$.
- For example, when we toss a coin, we have two possibilities, i.e. head or tail. The probability of one of these two outcomes is the probability of success or failure, which is 0.5 (this would be our $p = p_0$ in this case) for tossing a fair coin.
- If the difference between $\hat{\pi}$ and p_0 is sufficiently large, then the null hypothesis is rejected, i.e. the pre-specified value p is not consistent with the observed data.
- If we fail to reject the null hypothesis, this implies that the experimental data support the theoretical pre-specified value.

Choosing statistical tests

IV. The **binomial test** for **one** sample:

– It follows that the possible hypotheses in the binomial test are:

	Two-tailed	Left-tailed	Right-tailed
$H_0:$	$p = p_0$	$p \geq p_0$	$p \leq p_0$
$H_1:$	$p \neq p_0$	$p < p_0$	$p > p_0$

Weimann, J., & Brosig-Koch, J. (2019). *Methods in experimental economics: An introduction*. Springer Cham. pp. 228

Choosing statistical tests

II. The **binomial test** for **one** sample. When to use?

– Use the test when:

(a) the sample size is small,

(b) the data consists of independent trails (each trial = success or failure),

(c) the number of trials *n* is fixed,

(d) the probability of success p is constant across trials, and

(e) each trial has exactly two outcomes: Success or failure.

Choosing statistical tests

III. The **binomial test** for **one** sample.
Example:

```
. bitest quick == 0.25

Binomial probability test

    Variable |         N   Observed k   Expected k   Assumed p   Observed p
-------------+-------------------------------------------------------------
       quick |        15           7          3.75     0.25000      0.46667

  Pr(k >= 7)           = 0.056620  (one-sided test)
  Pr(k <= 7)           = 0.982700  (one-sided test)
  Pr(k <= 0 or k >= 7) = 0.069984  (two-sided test)
```

Choosing statistical tests

IV. The **binomial test** for **one** sample. Interpretation of the results:

- Here, we have three different *p*-values representing different tests each.
- $Pr(k \geq 7) = 0.057$: This is a one-sided (right-tailed) *p*-value testing H1: $p > 0.25$, specifically the probability of seeing 7 or more successes if the true p were 0.25.
- $Pr(k \leq 7) = 0.983$: This is a one-sided (left-tailed) *p*-value testing H1: $p < 0.25$, i.e., the probability of seeing 7 or fewer successes if the true p were 0.25.
- $Pr(k \leq 1 \text{ or } k \geq 7) = 0.070$: This is the two-sided *p*-value (combining the tails), which tests H1: $p \neq 0.25$.

Choosing statistical tests

V. The **binomial test** for **one** sample. Interpretation of the results:

	Two-tailed	Left-tailed	Right-tailed
$H_0:$	$p = 0.25$	$p \geq 0.25$	$p \leq 0.25$
$H_1:$	$p \neq 0.25$	$p < 0.25$	$p > 0.25$

Choosing statistical tests

VI. The **binomial test** for **one** sample. Interpretation of the results:

- Pr(k ≥ 7) = 0.057 (one-sided (right-tailed) *p*-value): We can reject the null hypothesis (that, on average, 25% of University students would get a “success” on the visual quickness test) at the 10% significance level since *p*-value = 0.057 < 0.10.
- Pr(k ≤ 7) = 0.983: Irrelevant in this case.
- Pr(k ≤ 1 or k ≥ 7) = 0.070: (two-sided *p*-value): We can reject the null hypothesis (that, on average, 25% of University students would get a “success” on the visual quickness test) at the 10% significance level since *p*-value = 0.057 < 0.10.

Choosing statistical tests

I. **Fisher's exact** and **Pearson's χ^2 test** for **two independent** samples:

– Both are **non-parametric** tests for two independent samples.

– They both examine relationships in categorical data (often represented in so-called contingency tables) to test whether two (or more) categorical variables are independent.

– However, the two tests differ in their underlying assumptions, the calculation of their test statistic, and their typical use cases.

Choosing statistical tests

II. Fisher's exact test for **two independent** samples:

– Fisher's exact test is commonly applied to 2x2 contingency tables.

– It is particularly useful for small sample sizes.

– The test calculates the exact probability of observing a table as extreme or more extreme under the null hypothesis of independence.

– The test calculates exact *p*-values.

Choosing statistical tests

III. Fisher's exact test for **two independent samples:**

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	$x_{21} = 9$	$x_{22} = 3$	$n_2 = 12$
		$N_1 = 11$	$N_2 = 9$	$N = 20$

Weimann, J., & Brosig-Koch, J. (2019). *Methods in experimental economics: An introduction*. Springer Cham. pp. 236

Choosing statistical tests

IV. Pearson's χ^2 test for **two independent** samples:

- Pearson's χ^2 test is applied to 2x2 or larger contingency tables.
- It is particularly useful for moderate and/or large sample sizes.
- The test compares the observed frequencies with expected frequencies under the null hypothesis of independence.
- The test calculates approximate *p*-values.

Choosing statistical tests

V. **Pearson's χ^2 test** for **two independent** samples:

Table 4.17 Contingency table 2 for the χ^2 test in the example

		Reaction time		
		Fast	Slow	
IQ	Low	52	48	100
	High	49	51	100
		101	99	200

Weimann, J., & Brosig-Koch, J. (2019). *Methods in experimental economics: An introduction*. Springer Cham. pp. 240

Choosing statistical tests

VI. **Fisher's exact** and **Pearson's χ^2 test** for **two independent** samples. Example:

```
. tab region agecat, exact chi2

    Census |     Age category
    region |     19-29      30-34 |     Total
-----------+----------------------+----------
     South |       301        233 |       534 
      West |       206        216 |       422 
-----------+----------------------+----------
     Total |       507        449 |       956 

          Pearson chi2(1) =   5.3972   Pr = 0.020
           Fisher's exact =                 0.022
   1-sided Fisher's exact =                 0.012
```

Choosing statistical tests

VII. Fisher's exact and **Pearson's χ^2 test** for **two independent** samples. Example:

- *Null hypothesis* $\boldsymbol{H_0}$: Age categories distribution is the same in the South and the West, i.e. age is independent of region.
- *Alternative hypothesis* $\boldsymbol{H_1}$: Age categories distribution differs between the South and the West.

Choosing statistical tests

VIII. Fisher's exact and **Pearson's χ^2 test** for **two independent** samples. Interpretation of the results:

- **Fisher's exact test**: Since the *p*-value = 0.022 < 0.05, we can reject the null hypothesis (that the age categories distribution is the same in the Sout and the West) at the 5% significance level.

- **Pearson's χ^2 test**: Since the *p*-value = 0.022 < 0.05, we can reject the null hypothesis (that the age categories distribution is the same in the Sout and the West) at the 5% significance level.

Choosing statistical tests

I. The **McNemar test** for **two paired** samples:

– The McNemar test is a **non-parametric** test used for comparing two paired (i.e. dependent) samples with respect to a dichotomous attribute.

– The test checks the marginal homogeneity of two dichotomous variables, i.e. the test examines whether there is a statistically significant change in responses between the two related samples/conditions.

– Thus, the null hypothesis is: H_0: *There is no systematic difference in the proportion of subjects who change from one category to another.*

– The alternative hypothesis is: H_1: *There is a systematic difference in the proportion of subjects who change from one category to another.*

Choosing statistical tests

I. The **McNemar test** for **two paired samples**:

Table 4.20 Contingency table 2 in the McNemar test

		With info		
		High	Low	
Without info	High	21	15	36
	Low	6	10	16
		27	25	52

Weimann, J., & Brosig-Koch, J. (2019). *Methods in experimental economics: An introduction*. Springer Cham. pp. 243

Choosing statistical tests

II. The **McNemar test** for **two paired** samples. When to use?

– Use the test when:

(a) you have paired (i.e. dependent) nominal/categorical data,

(b) the data can be represented in a 2x2 contingency table, and

(c) you want to examine whether the proportions of responses have changed across conditions/treatments.

Choosing statistical tests

III. The **McNemar test** for **two paired** samples. Example:

```
. mcc smoke1 smoke0

                 |        Controls        |
Cases            |   Exposed   Unexposed  |      Total
-----------------+------------------------+-----------
         Exposed |         8          22  |         30
       Unexposed |         8          18  |         26
-----------------+------------------------+-----------
           Total |        16          40  |         56

McNemar's chi2(1) =      6.53    Prob > chi2 = 0.0106
Exact McNemar significance probability       = 0.0161

Proportion with factor
        Cases       .5357143
        Controls    .2857143     [95% conf. interval]
                   ---------     --------------------
        difference       .25     .0519726   .4480274
        ratio          1.875     1.148685   3.060565
        rel. diff.       .35     .1336258   .5663742

        odds ratio      2.75     1.179154   7.143667   (exact)
```

Choosing statistical tests

IV. The **McNemar test** for **two paired** samples. Interpretation of the results:

- ➢Since the exact *p*-value 0.0161 < 0.05, then we do have statistical evidence to reject the null hypothesis (that there is no systematic differences in the proportion of mothers who smoked) at the 5% significance level that the proportion of mothers who smoked during pregnancy is the same across the two groups.

- ➢Additionally, the test provides odds ratio. In this case, this means that mothers who smoked during the pregnancy were about 2.75 times more likely to have a baby with low birth weight compared to mothers who did not smoke during the pregnancy.

Questions

Paper reading

THE ECONOMIC JOURNAL

The Economic Journal, **130** (*November*), 2569–2595 DOI: 10.1093/ej/ueaa055
Advance Access Publication Date: 21 May 2020

THE STRUCTURE AND BEHAVIOURAL EFFECTS OF REVEALED SOCIAL IDENTITY PREFERENCES*

Florian Hett, Mario Mechtel and Markus Kröll

A large body of evidence shows that social identity affects behaviour. However, our understanding of the substantial variation of these behavioural effects is still limited. We use a novel laboratory experiment to measure differences in preferences for social identities as a potential source of behavioural heterogeneity. Facing a trade-off between monetary payments and belonging to different groups, individuals are willing to forego significant earnings to avoid belonging to certain groups. We then show that individual differences in these foregone earnings correspond to the differences in discriminatory behaviour towards these groups. Our results illustrate the importance of considering individual heterogeneity to fully understand the behavioural effects of social identity.

Paper discussion

1. What are the authors investigating in the paper?
2. Describe how they did that experimentally. What experimental design and which “game(s)” did they implement?
3. Is their sample representative? Why/why not?
4. Name all statistical tests mentioned in the paper.

Relevant for the exam

- **Name and describe** types of measurement scales of variables
- **Explain** random variables and **differentiate** between their different types
- **Explain** the key elements of creating a statistical design
- **Formulate** testable hypotheses
- **Choose** the appropriate statistical test
- **Interpret** the results of a statistical test given a Stata output
- **Name and Describe** the different possible errors of a statistical test