Introduction to Experimental and Behavioral Economics
Lecture 2: Foundations of Experimental Economics

Dr. Jonathan Stäbler
Masaryk University
Faculty of Economics and Administration

Spring 2025

Outline of Lecture 2

Economics and Experiments

Understanding Social Reality

- Economics aims to understand social reality: How people behave and why?
- Can we collectively do better than we currently do? If yes, how?

The Role of Formal Models and Experiments

- Unlike most social sciences, economics builds formal theoretical models.
- Laboratory experiments stand between theory and reality.

Three Purposes of Laboratory Experiments

1. Testing Theory

- Assess empirical relevance of theoretical models.
- Example: Test selfishness assumption with Altruism games.
- Measure individual preferences and behavior in controlled settings.

2. Searching for Facts

- Some economic environments are too complex for theory and math.
- Example: Auctions with specific rules that theory cannot easily model.
- Experiments provide empirical knowledge by replicating real-world settings.

3. Whispering in the Ears of Princes

- Inform decision-makers about policy impacts.
- Test-bed for market mechanisms, policy changes, or organizational designs.
- Laboratory experiments are cost-effective and allow precise control over variables.

Categories of Experimental Parameters

- **Behavioural Parameters**: Describe behaviour in a controlled environment and their determinants.
  - *Example*: Comparing observed contributions in a public-good game to predicted Nash equilibria (testing the general validity of the theory).
  - Additionally, exploring reciprocity and altruism as explanations for deviations.
- **Comparative Statics**: Comparing parameters between different experimental conditions (treatment effects).
  - Isolate the causal effects of one parameter compared to the other.
  - *Example*: Does observability change giving behavior of dictators?
  - *Example:* How do contributions to a public good change if the number of team members increases from 2 to 5?
- **Discover Individual Heterogeneity in Behaviour**:
  - *Example:* Gender differences in giving behavior of dictators.

The Two Inferential Problems in Data Analysis

Data analysis relies on samples, but the goal is to infer characteristics of the whole population.

Two key challenges:

- **Inference**: How well does the sample represent the whole population?
  + **Statistical Inference**: How does sample variability affect our conclusions about the population? *[[Watch Video]]*
- **Identification**: Are the conclusions we draw from the results correct?
  [[Watch Video]]

Data Generating Process

Data Generating Process (DGP)

- The DGP refers to the underlying mechanism that produces the data we observe. It includes:
  - The structural relationships between variables (e.g., causal mechanisms).
  - Randomness or noise that influences the data (e.g., errors in measurement or external shocks).
- Understanding the DGP is essential for interpreting data and drawing valid conclusions.

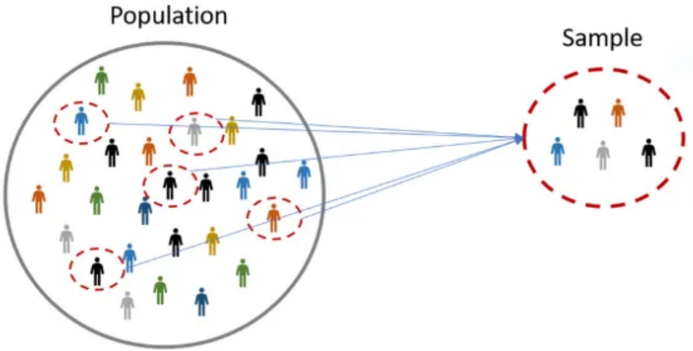
Experiments and Data-Generating Processes

- With experimental data, the researcher can design the decision-making environment and thus determine the data-generating process (DGP).
- Unlike observational studies, experiments allow researchers to select the DGP based on the research question.
- This reverses the typical econometric challenge: instead of inferring the DGP, experiments allow its direct specification.
- The art is to specify the DGP in the best way.

Outline of Lecture 2

Sampling and Inference

Since we cannot run experiments on the entire population, we have to rely on samples.



The goal is to draw conclusions from the sample that generalize to the entire population (inference).

Potential Problems in Sampling

Sampling Bias: Occurs when some groups are more likely to be selected, leading to unrepresentative samples and biased conclusions.

- Non-Random Sampling: Any selection process that isn't purely random (e.g., purposive, convenience).
- Selection Effects: Individuals with certain traits (e.g., stronger opinions, participants who need more money) may be more likely to participate.

Random Sample: Each individual of the entire population has an equal chance of being chosen.

→ But: Even with random sampling, chance can still lead to over- or under-representation of groups. Thus, we always require statistical testing for inference.

→ The larger your sample, the more likely that you get the behavior that you would get from the full population.

Statistical Inference

Are our observations typical for the population (assuming that we have random-sampling) or are they driven by chance by the specific random sample that we draw.

- By chance it could be that we draw majorly females, younger individuals, or individuals with very specific traits and behavior.
- Luckily, we can test and find probabilities for the possibility of seeing an effect by chance.

Example: Hypothesis Testing in the Dictator Game

Assume a dictator game, where the dictator can either give money to the recipient, or take money away.
Null Hypothesis (H_0): The dictator keeps all the money (no transfer).
Alternative Hypothesis (H_1): The dictator either transfers some money or takes money away.

$\longrightarrow$ Use a statistical test to examine whether observed transfers differ significantly from zero.

Example: Hypothesis Testing in the Dictator Game

Suppose that the dictator can either give (up to A) or take money (up to A) from the recipient, such that $x_i = [-A, A]$. Let $\bar{x}$ be the average amount that all dictators give (or take) that we observe in the sample.

- $H_0 : \mu = 0$
- $H_1 : \mu \neq 0$

Now we are testing assuming that H_0 is true by looking at the data. After our test, we will make one of the following decisions:

1. Reject the null hypothesis.
2. Fail to reject the null hypothesis.

With our decision, we could make two types of errors:

- Type-I Error: We reject the null hypothesis even though it is true.
- Type-II Error: We fail to reject the null hypothesis even though it is false.

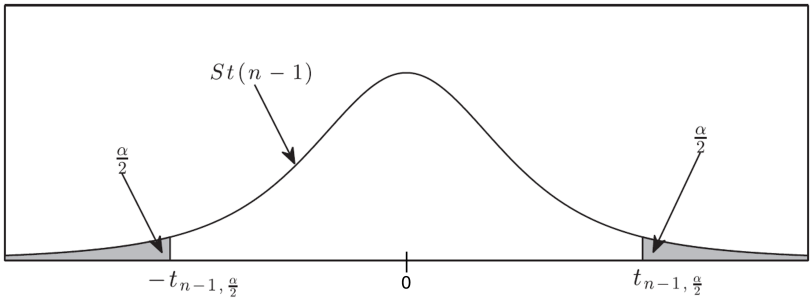
Example: Hypothesis Testing in the Dictator Game

Level of significance: Maximum allowable probability to make a type-I error (rejecting the null even though it is true). Denoted by α and common thresholds are $\alpha = 0.1, 0.05$ and 0.01.

Power of a test: Denotes the probability to reject the null hypothesis, when it is indeed false. Denoted by (1-β), where β represent the probability to make a type-II error. A common power threshold is 0.8.

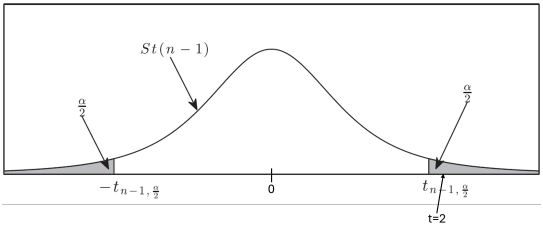
Example: Hypothesis Testing in the Dictator Game

Remember, we want to see whether $H_0 : \mu = 0$ holds. For a t-test, we assume that our observed data is randomly distributed. Suppose that we take a large number of samples and always compute the mean $\bar{x}$. By the CLT, the sample mean $\bar{x}$ is approximately normally distributed for large n. We can compute the t-value $t = \frac{\bar{x}}{s/\sqrt{n}}$, which will follow a t-distribution with $n - 1$ degrees of freedom.



Example: Hypothesis Testing in the Dictator Game

Suppose we observe a mean of 4 and an associated value of $t = 2$. We can then compute a *p-value* based on the distribution of t. The *p-value* tells us how likely it is to see the value of t (or more extreme), given that $H_0 : \mu = 0$ is true.



If $p < 0.05$, we reject H_0 and conclude that there is statistical evidence for a difference between the observed mean and the hypothesized mean.

Example: Hypothesis Testing in the Dictator Game

If we find a significant difference between the theorized and the observed mean (e.g., positive transfers), we can say that we find evidence in favor of a specific theory.

- We find evidence for refuting the benchmark theoretical prediction of 0 transfers.
- We do not find evidence in favor of the causal mechanism of this deviations.
- We cannot say that people are for sure altruistic or that they for sure like to follow a norm.
- Other factors could have caused this deviation from the selfish benchmark.
- Thus we need to design an experiment to **identify** the causal reason for this deviation.
  $\Longrightarrow$ Relies on experimental variations of key factors (treatments).

Outline of Lecture 2

Identification

Identification is how you link the *result* you see with the *conclusions* you draw from it.

- It is about generalizing the results to a theoretical construct.
- It is about *causality* and not simply correlations.

For *example*, say you observe (in the data) that people eat more ice creams when they wear shorts.

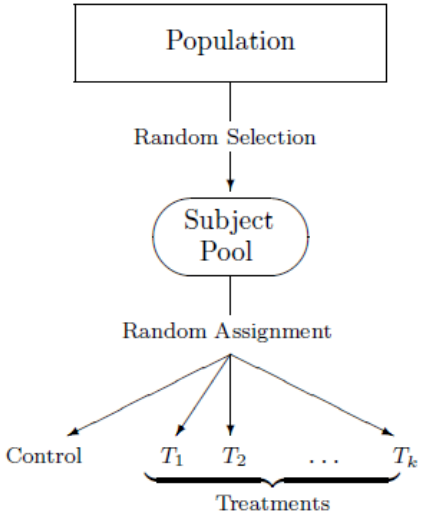
- No inference error because the whole population on average will eat more ice cream when they are wearing shorts compared to when they are wearing long trousers.
- *Identification error:* Drawing the conclusion that people wear shorts **because** they eat more ice cream.

Problem: Eating ice cream (and wearing shorts) are both correlated with higher temperatures. It is higher temperatures that causes people to wear shorts, not because they eat ice cream.

Identification Error

Omitted variables: There's a relationship between Y and X for reasons other than your theoretical conclusion.

Identification through Random Treatment Assignment



The Potential Outcome Framework

- We want to determine (identify) the causal effect of a binary treatment X_i on an outcome Y_i.
- Example:
  - Let Y_i represent health status.
  - The treatment is a new medicine:
    - $X_i = 1 \Rightarrow$ Takes the new medicine.
    - $X_i = 0 \Rightarrow$ Does not take the new medicine.
- Each individual has two potential outcomes:
  - $Y_i(1)$: Outcome if the individual takes the new medicine.
  - $Y_i(0)$: Outcome if the individual does not take the new medicine.
- The causal effect on individual i is:

$$Causal\ Effect_i = Y_i(1) - Y_i(0)$$

Observed Outcome and the Identification Problem

- The observed outcome Y_i can be written in terms of the potential outcomes:

$$Y_i = Y_i(1) \cdot X_i + Y_i(0) \cdot (1 - X_i)$$

- If the individual receives the treatment ($X_i = 1$): $Y_i = Y_i(1)$
- If the individual does not receive the treatment ($X_i = 0$): $Y_i = Y_i(0)$

The Identification Problem

We cannot identify the causal effect for an individual i because we observe either $Y_i(1)$ (outcome under treatment) or $Y_i(0)$ (outcome without treatment), but never both simultaneously.

$\longrightarrow$ **Between-subjects design**: Assign treatment to one group and compare outcomes to a similar untreated group (but: confounder of potential differences between groups).

$\longrightarrow$ **Within-subjects design**: Observe the same individuals over time, comparing behavior before and after treatment (but: confounder of order and time).

Outline of Lecture 2

Identification Between-Subjects

- Although we can never *truly* observe the causal effect for individual i, we might be able to estimate the average causal effect in a population with a between-subjects design.
- The average causal effect / average treatment effect is:

$$E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$$

- Suppose we set up an ideal randomized experiment:
  - We take a random sample of the population.
  - We randomly give half of the sample the treatment.
  - The other half does not get the treatment.

Identification Between-Subjects

- The potential outcomes can differ between individuals:

$$Y_i(1) \neq Y_j(1) \text{ and } Y_i(0) \neq Y_j(0) \text{ for } i \neq j$$

- However, if the treatment X_i is randomly assigned, the distribution of potential outcomes will be the same in the treatment group ($X_i = 1$) and in the control group ($X_i = 0$).
- With random assignment, the potential outcomes are independent of the treatment:

$$Y_i(1), Y_i(0) \perp X_i$$

- We thus have:

$$E[Y_i(1) \mid X_i = 1] = E[Y_i(1) \mid X_i = 0]$$
$$E[Y_i(0) \mid X_i = 1] = E[Y_i(0) \mid X_i = 0]$$

The Potential Outcome Framework & A Randomized Experiment

- In a randomized experiment, individuals are randomly assigned to a treatment and control group. We therefore have that:

$$E[Y_i(1)] = E[Y_i(1) \mid X_i = 1] = E[Y_i \mid X_i = 1]$$

$$E[Y_i(0)] = E[Y_i(0) \mid X_i = 0] = E[Y_i \mid X_i = 0]$$

- This implies that:

$$E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)] = E[Y_i \mid X_i = 1] - E[Y_i \mid X_i = 0]$$

- We can thus estimate the average causal effect of the treatment by taking the difference in mean outcomes of the individuals in the treated group and control group.

Example: Project STAR

- A large-scale and influential randomized experiment: Project STAR (Student-Teacher Achievement Ratio)
- Kindergarten students and their teachers were randomly assigned to one of three groups beginning in the 1985-1986 school year:
  - Small classes (13-17 students per teacher)
  - Regular-size classes (22-25 students)
  - Regular/aide classes (22-25 students) which also included a full-time teacher's aide
- Over all 4 years about 11,600 students from 80 schools participated in the experiment
- Project STAR was funded by the Tennessee legislature, at a total cost of approximately $12 million over four years.

Example: Project STAR

- Kindergarten students were randomly assigned to 3 groups
- To simplify, we combine the regular-size classes and the regular-size classes with an aide into 1 group
- This gives two groups:
  - A treatment group ($X_i = 1$): students assigned to a small class (13-17 students)
  - A control group ($X_i = 0$): students assigned to a regular class (22-25 students)
- We are interested in the causal effect of class size on student achievement.
- The outcome variable Y_i is the Stanford Achievement Test score at the end of kindergarten.

Example: Project STAR

- For each student i we have two potential outcomes:
  - $Y_i(1)$ is the test score in case student i would be in a small class.
  - $Y_i(0)$ is the test score in case student i would be in a regular class.
- The causal effect of class size on test score for pupil i is $Y_i(1) - Y_i(0)$.
- This is unobserved.
- Because students were randomly assigned to the treatment group (small class) and the control group (regular class), we can **estimate the mean causal effect** $E[Y_i(1) - Y_i(0)]$.
- We can compare mean test scores of the students in a small class ($E[Y_i|X_i = 1]$) with the mean test scores of students in a regular class ($E[Y_i|X_i = 0]$).

Example: Project STAR

- Mean test score students in regular class: $E[Y_i|X_i = 0] = 918.20$
- Mean test score students in small class: $E[Y_i|X_i = 1] = 931.94$
- Estimate of average causal effect: $E[Y_i|X_i = 1] - E[Y_i|X_i = 0] = 13.74$

```
. ttest testscore, by(small_class)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	4048	918.2013	1.135017	72.21422	915.9761	920.4266
1	1738	931.9419	1.831611	76.35863	928.3495	935.5343
combined	5786	922.3287	.9695111	73.7466	920.4281	924.2293
diff		-13.74055	2.107334		-17.87172	-9.609391

```
    diff = mean( 0) - mean( 1)                                    t =  -6.5204
Ho: diff = 0                     degrees of freedom =             5784

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) =  0.0000         Pr(|T| > |t|) =  0.0000          Pr(T > t) =  1.0000
```

Example: Project STAR

After **identifying** the effect (average causal effect), we have to ensure that we are drawing the correct **inference**.

- Rely on a representative sample of all kindergarten.
- Test whether your sample is representative or systematically different from the population in all kindergartens.
- Be aware of the specifics of the sample and careful in the conclusions that you draw (e.g., are the results specific for Tennessee?).
- Draw on former research showing similar behavior between subgroups of the population.

Statistical inference: We can test whether we find an effect by chance – even though there was random sampling. $\Longrightarrow$ Use statistical tools.

Example: Project STAR – Statistical Inference

Null Hypothesis: There is no difference in the means between the treatment and control groups.

$$H_0 : \mu_0 = \mu_1$$

Compute the t-statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{918.2 - 931.9}{\sqrt{\frac{72.2^2}{4048} + \frac{76.4^2}{1738}}} \approx -6.5204$$

```
. ttest testscore, by(small_class)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	4048	918.2013	1.135017	72.21422	915.9761	920.4266
1	1738	931.9419	1.831611	76.35863	928.3495	935.5343
combined	5786	922.3287	.9695111	73.7466	920.4281	924.2293
diff		-13.74055	2.107334		-17.87172	-9.609391

```
    diff = mean( 0) - mean( 1)                                    t =   -6.5204
Ho: diff = 0                     degrees of freedom =             5784
```

Example: Project STAR – Statistical Inference

Null Hypothesis: There is no difference in the means between the treatment and control groups.

$$H_0 : \mu_0 = \mu_1$$

Compute the t-statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{918.2 - 931.9}{\sqrt{\frac{72.2^2}{4048} + \frac{76.4^2}{1738}}} \approx -6.5204$$

p-value: Looking up the t-value in a table:

$$p < 0.00001$$

Conclusion: The probability of observing this t-value under H_0 is less than 0.001%. We **reject** H_0 and find strong evidence for a significant difference between group means.

Outline of Lecture 2

Identification with Within-Subject Designs

- Unlike between-subject designs, within-subject designs observe the same individuals in multiple treatment conditions.
- This eliminates unobservable individual heterogeneity by ensuring that differences in outcomes arise solely from the treatment.
- The treatment effect is estimated by comparing outcomes within the same individuals before and after treatment:

$$E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$$

- Crucial assumption: no change over time other than the treatment!

Advantages of Within-Subjects Design

- **Higher statistical power**: Fewer participants are needed since each subject provides multiple observations.
- **Reduced variance**: Individual-specific heterogeneity is controlled for, reducing noise in data.
- **Efficient data collection**: More observations per subject.
- **However:** More restrictive assumptions for identification. Econometrically not as clean a comparison as between-subject comparisons.

It's BAD Experiment Time!

Let's play a within-subject experiment!!

Reasons why it was a bad experiment:

- **I deceived you** – I promised to pay you chocolate balls and then didn't. This is an absolute NO GO and **will never happen in an economic experiment**.
- **Wealth effects:** Accumulated points from earlier games may have influenced later decisions.
- **Feedback effects:** Observing partner behavior after each game allowed you to adjust your strategy.
- **Learning effects:** You might not have completely understood the games in the first rounds. In the second round you had already played the game and might have understood it better for your second choices.

Example: Bertrand Game

- Firms compete by setting prices simultaneously.
- Consumers buy from the firm with the lowest price (assuming homogeneous goods and no capacity constraints).
- Nash Equilibrium: In the classic Bertrand model with identical products and constant marginal costs c, both firms price at $p = c$, leading to zero economic profits.

Our Experiment

In our experiment: $c_0 = 0$ and $c_1 = 40$.
Causal effect: $\mathbb{E}[p_i(c_1) - p_i(c_0)] = \mathbb{E}[p_i(c_1)] - \mathbb{E}[p_i(c_0)]$
Test with a paired t-test (as samples are dependent).
Does order matter here?

Example: Trust Game

- Two players: *Trustor* (Player 1) and *Trustee* (Player 2).
- Player 1 sends money x (usually tripled: $m = 3$), or keeps it.
- Player 2 either *returns* a share or *keeps all*.
- Nash Equilibrium: Player 2 keeps all, so Player 1 sends nothing $\rightarrow$ *No trust*.

Our Experiment

In our experiment: $m_0 = 3$ and $m_1 = 2$.
Causal effect: $\mathbb{E}[x_i(m_1) - x_i(m_0)] = \mathbb{E}[x_i(m_1)] - \mathbb{E}[x_i(m_0)]$
Test with a paired t-test (as samples are dependent).
Does order matter here?

Challenges and Potential Solutions

Crucial assumption: no change over time other than the treatment! Strictly speaking it can barely hold:

- **Order Effects**: Exposure to one treatment may influence responses to subsequent treatments.
  - → Solution: Randomize the order of treatments.
- **Time-Related Confounders**: Learning effects, feedback about matched partners behavior, fatigue, or boredom may bias results.
  - → Solution: Include control questions, thorough tutorial, or analyze early vs. late trials separately, give feedback only after all choices have been done.
- **Wealth Effect**: Players accumulate wealth over time.
  - → Solution: Randomly choose only one payoff-relevant round.

$\Longrightarrow$ Show empirically that order doesn't matter. If it does, analyze only the first choices in a between-subjects way.

Outline of Lecture 2

Internal and External Validity

Internal Validity

"A statistical analysis is **internally valid** *if the statistical inference about causal effects are valid for the population being studied."* (Stock and Watson, 2020, p.313)
$\Longrightarrow$ Related to identification.

External Validity

"The analysis is **externally valid** *if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings.* (Stock and Watson, 2020, p.313)
$\Longrightarrow$ Related to inference.

Threats to Internal Validity in a Randomized Experiment

- Analyzing data from an **ideal** randomized experiment will give an unbiased & consistent estimate of the causal effect of the treatment.
- In practice, setting up an ideal randomized experiment is not easy and often things do not go as planned.
- This is especially true for field experiments, wherease it is easier to obtain for laboratory experiments.
- **Threats** to internal validity:
  1. Failure to randomize
  2. Attrition
  3. Experimenter demand effects/ Hawthorne effect
  4. Small samples
  5. Failure to induce value

Failure to Randomize

- The treatment might not be assigned randomly but instead is based on characteristics or preferences of the subjects.
- We can "check" whether the treatment was randomly assigned by comparing observed characteristics between the treatment and control group.
- Table shows mean characteristics of students in project STAR:

Characteristic	**Small Class**	**Regular Class**	**Mean Difference**	**p-value**
Gender (boy=1)	0.514	0.513	0.001	0.969
Race (black=1)	0.312	0.331	-0.019	0.140
Eligible for free lunch	0.471	0.490	-0.019	0.162

- No significant difference in the observed characteristics between those assigned to the treatment group (small class) and the control group (regular class).

However: there may be differences in characteristics that we cannot observe. This might bias the results.

Attrition

- Attrition refers to subjects dropping out of the study after being randomly assigned to the treatment or control group.
- Not problematic if attrition is unrelated to the treatment.
- If attrition is related to the treatment, the estimate of the treatment effect will be biased.
- For example, if students of the smaller classes in the STAR project changed to other (better) schools before the Stanford Achievement Test.

Experimenter Demand Effect/ Hawthorne Effect

- Hawthorne effect: Human subjects might change their behavior, merely because they are part of an experiment.
- For example, teachers assigned to small classes might put in extra effort.
- They would like the researchers to find a positive effect of small class size.
- Teachers like to teach small classes.

Experimenter Demand Effect/ Hawthorne Effect

- In some experiments, a "double-blind" protocol can mitigate the Hawthorne effect.
- Subjects and experimenters know that they are in an experiment...
  - ...but neither knows which subjects are in the treatment group and which in the control group.
- In this case, the treatment & control group experience the same experimental effects...
  - ...and differences in outcomes can be attributed to the treatment.
- Double-blind experiments are not always feasible (especially in field experiments).

Experimenter Demand Effect in Laboratory and Online Experiments

- Experimental outcomes may not reflect real-world behavior due to subject awareness of being in an experiment.
- Subjects may alter behavior based on perceived experimenter expectations rather than true preferences.
- The experimental setting itself may motivate subjects to behave differently ("coming to play").
- Subjects may adjust earnings to influence the money that researchers have to pay, which is not a concern in real-world settings.

Small Samples

- Experiments with human subjects can be expensive.
- The sample size in experiments is therefore sometimes (too) small.
- Small samples do not produce biased estimates, but often produce imprecise estimates (large standard errors).
- In addition, large-sample approximations might not be justified and confidence intervals and hypothesis tests might not be valid.

Solution: Run a proper ex-ante power analysis and run the experiment only if you have enough funding.

Failure to Induce Value

- Introduced as a core internal-validity issue in (Smith, 1976): Induced value theory
- Solution: Pay participants for their actions as incentives drive decisions.

1. **Non-satiation:**
   - More must always be better at any point in the experiment.
   - Ensures participants care about consequences of their choices.
2. **Saliency:**
   - Decisions must be unambiguously linked with rewards.
   - Payoff differences are such that choices are worth it.
3. **Dominance:**
   - The whole experiment is attractive enough to compensate for the opportunity cost of participation.
   - Compensation must cover cognitive effort cost.

The Logic of Incentives

- Incentives ensure control over preferences through control of incentives.
- They can make the experiment incentive-compatible and internally valid.
- They decrease measurement error and make sure participants act as if they would in real life economic situations.
- Define relationships between choices and consequences (payoffs).

Threats to External Validity in a Randomized Experiment

Can we generalize the results based on the randomized experiment to other settings and populations?

- The population studied and the population of interest might differ.
- Often experiments use subjects that signed up for participation in the experiment (volunteers, students).
- These volunteers are often more motivated.
- Even if these volunteers are randomly assigned to treatment and control group...
- ...the estimated average treatment effect might not be informative for a general population.

Outline of Lecture 2

The Challenges of Designing an Experiment

- Designing a good experiment is difficult — there is rarely a perfect one.
- Every design involves trade-offs.
- The key is to design an experiment that best answers the research question.

Goals:

- Guarantee identification (ability to speak of a *causal* effect)
- Ensure statistical ability to detect an effect (decrease measurement error, decrease noise, large enough power, etc.)
- Ensure that we can infer something about the studied population.

The Many Choices

- Laboratory vs. online vs. field experiment
- Between- vs. within-subject designs
- Types of randomization
- One round vs. multiple rounds
- Framing vs. no framing
- etc.

Outline of Lecture 2

Laboratory vs. Field Experiments in Economics

Laboratory Experiments

- Conducted in controlled environments, allowing precise manipulation of variables.
- Aim to test theoretical predictions and isolate causal mechanisms.

Field Experiments

- Conducted in real-world settings, increasing external validity.
- Subjects often unaware they are part of an experiment.
- Aim to evaluate policy interventions, market behaviors, or naturally occurring economic interactions.

Laboratory vs. Field Experiments

Laboratory Experiments

+ High internal validity.
+ Easier to control confounding variables.
+ Easier to randomize well and less/ no attrition.
– Often use student participants or convenience samples.
– Conducted in an artificial environment.
– Lower external validity—results may not generalize well to real-world settings.

Field Experiments

+ Higher external validity—results are more generalizable.
+ Less experimenter-demand effects.
– More challenging to control confounding variables.
– More expensive and time-consuming.
– May have lower internal validity.
– Attrition may be a problem.

Online Experiments in Economics

Definition and Characteristics

- Conducted on platforms like Prolific and MTurk, reaching diverse and large participant pools.
- Participants complete tasks remotely for monetary compensation.
- Allows for rapid data collection of large samples.

Advantages

+ Greater demographic diversity compared to traditional lab experiments.
+ Lower costs and faster implementation.

Challenges

– Limited control over participants' environment, attention, and concentration.
– Potential for inattentive or non-serious responses $\longrightarrow$ data is usually noisier.
– Requires robust screening and attention checks to ensure data quality.
– Complex/ repeated/ interactive experiments are more difficult to conduct.

Do Self-Selected Student Samples Misrepresent Behavior in Lab Experiments?

- Comparative statics and correlations are similar across university samples, US-representative samples, and MTurk participants (Snowberg and Yariv, 2021).
- Self-selected students differ from the general student population only in some aspects (more risk averse, more willing to lie, less generous) (Snowberg and Yariv, 2021).
- Student and non-student samples exhibit similar distributions in key economic behaviors, including risk-taking and social preferences (Exadaktylos et al., 2013).
- University students are less generous than representative samples of Zurich and Norway (Falk, Meier, and Zehnder 2013; Cappelen et al. 2015)
- MTurk (online) participants behave similarly to university students on several "heuristic and biases" experiments and non-incentivized games, as well as (incentivized) repeated public goods and Prisoner's dilemma games (Paolacci, Chandler, and Ipeirotis 2010; Horton, Rand, and Zeckhauser 2011; Berinsky et al. 2012; Goodman, Cryder, and Cheema 2013; Arechar, Gächter, and Molleman 2018).

Do Self-Selected Student Samples Misrepresent Behavior in Lab Experiments?

Conclusion

- While selection effects exist, behavioral differences between student and general samples are not to large.
- Certain behaviors (e.g., competitiveness, risk preferences) show variation, but many core economic behaviors remain stable across samples.
- Comparative statics remain similar across samples.

Takeaway: Lab experiments provide valuable insights, but findings should be supplemented with diverse populations when external validity is a priority.

Outline of Lecture 2

Between- vs. Within-Subject Treatments

Between-Subject Design
Each participant experiences only one treatment.

Pros:

+ Avoids order effects and learning biases.
+ Mimics real-world decision-making scenarios.

Cons:

– Requires a larger sample size.
– Higher variance due to individual differences.

Within-Subject Design
Each participant experiences multiple treatments.

Pros:

+ Controls for individual heterogeneity.
+ More statistical power with fewer participants.

Cons:

– Risk of order effects and learning biases.
– Higher experimenter demand because of several choices under different conditions/ treatments.

Outline of Lecture 2

Bernoulli vs Complete Randomization

In between-subjects designs, randomization is essential for reducing bias and proper identification. There are different approaches to randomization, including:

- **Independent (Bernoulli) Randomization**: Each participant has an equal and *independent chance* of being assigned to a specific condition. This can lead to a different number of pariticpants in each condition.
- **Complete Randomization**: There is a predefined number of participants in each specific condition.

Stratification in Randomization

- Stratification means dividing the sample into subgroups (strata) before randomization.
- Ensures balanced representation across treatment groups.
- Common stratification variables include age, gender, or other relevant characteristics.
- **Benefits**:
  - Reduces variability between groups.
  - Controls for confounding variables.

Within-Session Randomization

Conducting randomization within a single experimental session offers several advantages:

- **Consistency**: It ensures that all participants undergo the same experimental conditions in a controlled environment, minimizing external variability.
- **Reduced Confounding**: Since all participants are randomized in a similar time frame, differences due to external factors (e.g., weather, day of the week) are less likely to influence results.

However: this may not always be feasible due to logistical constraints.

Common practice but not optimal: Session-level randomization. Can fail due to order effects, day(time) effects, news effects, selection effects, etc.

Outline of Lecture 2

One Round vs. Multiple Rounds

The choice depends on the research question:

- *One round*: Focus on initial responses, no reactions to each other or feedback.
  $\longrightarrow$ Each individual decision is a statistical independent observation.
- *Multiple rounds*: Capture learning through experience, observing others' behavior, and adapting responses.
  - Can reduce noise and make individual choices more precise.

  $\longrightarrow$ Individual decisions in a group are statistically dependent. The unit of observation is the group.
- Multiple rounds allow for convergence to Nash equilibria as players accumulate history and react to each other.

Multiple Rounds – Wealth Effects

In multi-round or multi-part experiments, participants accumulate wealth over time.
Problem:

- As participants accumulate wealth, their behavior may change in later rounds.

Solution:

- Pay only one randomly selected round (see Azrieli et al., 2018, for a theoretical argument).
- Ensures incentive compatibility in every round.
- **Downside:** Some participants may be unlucky and receive payment for a round where they earned less than in others.

Outline of Lecture 2

Framing vs. No Framing

Framing

- The way information is presented influences decision-making.
- Emphasizes certain aspects of information to guide behavior.
- Can lead to different decisions based on how the situation is framed.
- Example: "Sabotage", "Stealing", "Cheating", "Opponents", "Friends", etc.

No Framing

- Information is presented without emphasis on any specific aspect.
- Decisions are made based on the raw content.
- Provides a neutral presentation, avoiding potential biases introduced by framing.
- Examples: "Choice A", "Choice B", "Other players", "Other participants", etc.

The Norm of No Deception in Experimental Economics

- Strong norm against using deception, defined as actively misleading subjects.
- Key Issue: Trust between subjects and experimenters is essential for control over experimental settings.
- Deception risks undermining subject trust.
- Negative externality: deception spreads knowledge of deceptive practices, undermining future studies.

This is different to studies in Psychology, who do use deceptive methods.

Outline of Lecture 2

The Econometrics of Experimental Data

- Focus: testing treatment effects using experimental data.
- The beauty of experiments: causal effects can be identified without complex econometrics.
- If there is perfect randomization, then individual controls are unnecessary.
- The challenge: designing experiments so that fancy methods are not needed.

Descriptives

- Summarizing the data before analysis is essential.
- Key statistics: mean, median, standard deviation, and distributions.
- Visualizations (histograms, boxplots) help identify patterns and anomalies.
- Descriptive analysis provides intuition before formal econometric tests.

The Econometrics of Experimental Data

Which statistical test to use depends on the measuring scale.

- **Nominal Scale**: Categorical, unordered (e.g., gender, colors).
  $\longrightarrow$ Chi-square test of independence. Compares frequencies in the categories.
- **Ordinal Scale**: Categorical, ordered, but differences are not measurable (e.g., satisfaction levels, rankings).
  $\longrightarrow$ Mann-Whitney U test. Compares the medians.
- **Interval Scale**: Numeric, meaningful differences, but arbitrary zero (e.g., temperature in Celsius).
  $\longrightarrow$ If sample large, then t-test. Compares the means. Otherwise Mann-Whitney U.
- **Ratio Scale**: Numeric, meaningful differences and ratios, true zero exists (e.g., height, weight, price).
  $\longrightarrow$ If sample large, then t-test. Compares the means. Otherwise Mann-Whitney U.

Types of Samples in Statistical Tests

The type of sample determines the appropriate statistical test. Independent samples are from different groups that are not related to each other. Dependent samples are related (e.g., same subjects measured at different times).

- **One-sample test**: Used to compare the sample mean to a known population mean or hypothesized value.
  Examples: One-sample t-test, Wilcoxon signed-rank test.
- **Independent samples test** for between-subject comparisons: Used to compare the means of two independent groups. Assumes that the groups are not related.
  Examples: Independent t-test, Mann-Whitney U test.
- **Dependent samples test** for within-subject comparisons: Used when the samples are related, such as when the same participants are tested twice (paired data).
  Examples: Paired t-test, Wilcoxon signed-rank test.

Parametric vs. Non-Parametric Tests

A **parametric test** assumes that the data follows a certain distribution (usually normal).

- Used for interval or ratio scales with normal distribution assumptions.
- More powerful, requiring stronger assumptions.
- *Examples:* T-test, ANOVA, Pearson correlation.

A **non-parametric test** does not assume a specific distribution for the data, making it more flexible when the underlying distribution is unknown or non-normal.

- Used for nominal or ordinal scales, or when the distribution is unknown or not normal.
- Less restrictive models, ideal for small samples.
- *Examples:* Chi-square test, Mann-Whitney U test, Spearman's rank correlation.

Frequently Used Statistical Tests

	Level of measurement and parametric assumptions		
	Interval/ratio and normal	Ordinal or interval/ratio and not normal	Categorical
One sample			
	t-test	Wilcoxon test	Binomial test
	z-test	Sign test	
	Chi-square test for the variance	Kolmogorov–Smirnov test	Chi-squared test
Independent samples			
2-sample	*t*-test	Mann–Whitney test	Fisher exact test
	z-test	Kolmogorov–Smirnov test	Chi-squared test
	Welch's test		
	F-test	Siegel–Tukey test	
K-sample	One-way ANOVA	Kruskal-Wallis test	Chi-squared test
	Barlett's test	Levene's test	
Dependent samples			
2-sample	Paired *t*-test	Matched-pairs Wilcoxon test	McNemar test
K-sample	Repeated-measure ANOVA	Friedman test	Cochran's Q test

Source: Figure 7.5 Jacquemet and L'Haridon (2018).

Robustness Checks

Common robustness checks involve testing with different regression frameworks, including demographics and other elicited factors, as well as accounting for time and round effects.

Key considerations:

- Select the appropriate regression framework (e.g., linear regression, Logit regression, etc.).
- Correctly cluster standard errors, especially when subjects make multiple decisions.

Readings for Next Week

Lecture 3: Failures of Expected Utility Maximization

- Chapters 2 & 3 Cartwright (2024)

Thank you and see you next week!
Jonathan.Stabler@econ.muni.cz

References I

Azrieli, Y., Chambers, C. P. and Healy, P. J. (2018). Incentives in experiments: A theoretical analysis, *Journal of Political Economy* **126**(4): 1472–1503.

Cartwright, E. (2024). *Behavioral economics fourth edition*, Routledge.

Exadaktylos, F., Espín, A. M. and Branas-Garza, P. (2013). Experimental subjects are not different, *Scientific reports* **3**(1): 1213.

Jacquemet, N. and L'Haridon, O. (2018). *Experimental economics*, Cambridge University Press.

Smith, V. L. (1976). Experimental economics: Induced value theory, *The American Economic Review* **66**(2): 274–279.

Snowberg, E. and Yariv, L. (2021). Testing the waters: Behavior across participant pools, *American Economic Review* **111**(2): 687–719.

Stock, J. H. and Watson, M. W. (2020). *Introduction to econometrics*, Pearson.