

# Introduction to Research

Methodology of Statistical Research. Statistical Method and Logic of Statistical Inference.

Tomáš Oleš

Department of Economic Policy  
Faculty of Economics and Finance

February 1, 2025

## **Course Objectives:**

- Focus on practical application of empirical research in economics with an emphasis on public policy questions.
- Begin with randomized experiments and progress to basic regression analysis, introducing statistical software STATA.
- Emphasis on working with real datasets, replicating results from scientific studies, and mastering practical steps in data description and analysis.
- Familiarize students with key econometric concepts and methods necessary for understanding contemporary empirical research and conducting their own projects.
- Cover regression analysis principles and modern econometric techniques aimed at identifying causal relationships.

# Learning Outcomes: Applied Data Analysis for Public Policy

- Master modern methods of data analysis, visualization, hypothesis testing, and basics of linear regression.
- Gain proficiency in using STATA for data processing and analysis, applicable in personal research projects.
- Understand basic econometric concepts, estimation methods, and statistical hypothesis testing techniques.
- Differentiate between correlation and causation and learn experimental and quasi-experimental designs (e.g., RCT, Difference-in-Differences, instrumental variables, etc.).
- Critically evaluate scientific studies and analyses, apply data insights in public policy contexts, and argue effectively.

"We are what we repeatedly do. Excellence, then, is not an act, but a habit."

-Aristotle

# Course Completion Requirements

## Grading Breakdown:

- **20%** – Participation in seminars
- **50%** – Assignments
- **30%** – Final exam

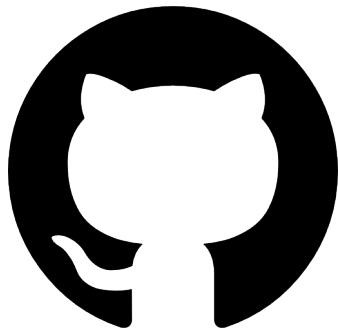
*Successful completion of the course requires consistent work throughout the semester and passing the final evaluation.*

# Required Literature

- Mehmetoglu, M., & Jakobsen, T. G. (2022). *Applied Statistics Using STATA: A Guide for the Social Sciences*.
- Lukáš Laffers (2024). *Pravdepodobnosť a štatistika 1*: Open Access e-text: <https://lukaslaffers.github.io/pas1/>
- Lukáš Laffers (2021). *Moderná aplikovaná regresia 1*: Open Access e-text: [https://static1.squarespace.com/static/52e69d46e4b05a145935f24d/t/64a0a2214abdb23994e932a2/1688248868079/MAR1\\_poznamkyMain.pdf](https://static1.squarespace.com/static/52e69d46e4b05a145935f24d/t/64a0a2214abdb23994e932a2/1688248868079/MAR1_poznamkyMain.pdf)
- IFP (2016). *Ideálny čas pre adresnejšie zdanenie fajčiarov*: Open Access e-text: <https://www.mfsr.sk/files/archiv/priloha-stranky/19972/81/Idealny-cas-adresnejsie-zdanenie-fajciarov.pdf>
- IFP (2019). *V nájme ďalej zájdeš: Podpora bývania na Slovensku*: Open Access e-text: [https://www.mfsr.sk/files/archiv/24/Podpora\\_byvania\\_analyza.pdf](https://www.mfsr.sk/files/archiv/24/Podpora_byvania_analyza.pdf)
- CORE Team (2018). *Doing Economics*. Open Access e-text: <https://core-econ.org/doing-economics>
- Cunningham, S. (2021). *Causal Inference: The Mixtape*. Yale University Press. (Chapters 4 and 5)

# Where Does ADAPP Live?

GitHub Repository: The central hub for all course materials, datasets, and code.



`https://github.com/tomasoles/applied\_data\_analysis\_for\_public\_policy/  
tree/main`

# Agenda

- Understand the methodological foundations of statistical research.
- Grasp the logic behind different types of statistical inference.
- Recognize the importance of sound theoretical frameworks.
- Gain insights into writing quantitative research papers and presenting data effectively.

# The Goal of Scientific Research and Statistics

- **Scientific Research:**

*The goal of scientific research is to make conclusions that go beyond the collected data. (King et al., 1994)*

- **Large-N Studies:**

- Enable generalizations about causal effects, provided causality is established.
- Depend on the availability of data (sample or full population).

- **Types of Statistics:**

- Descriptive: Describes distributions.
- Inferential: Examines relationships, enables predictions, and hypothesis testing.

- **Inferential Statistics:**

- Tool of the positivist tradition.
- Identifies patterns and regularities in the observable world.
- Roots in the systematic collection of data for induction (e.g., John Graunt, Sir William Petty, Hermann Conring).



# The History of Statistical Methods in Social Sciences

- **Key Contributions to Modern Statistics:**

- **Francis Galton:** Introduced the correlation coefficient, scatter plot, and regression analysis.
- **Karl Pearson:** Continued Galton's work and developed statistical methods further.
- **Émile Durkheim:** Placed statistics at the center of social sciences, linking variables such as suicide and religion.

- **Before Statistics in Social Sciences:**

- Research relied on philosophical reasoning and experiential facts (Ellwood, 1931).
- Example: An event in 17th-century Norway involving Scottish mercenaries illustrates this pre-statistical approach. In 1612, over 300 Scottish mercenaries crossed Norway to join Swedish forces during the Kalmar War. They encountered a lone Norwegian farmer in Gudbrandsdalen, leading to a dramatic event (story continues...)



# Statistical Evidence vs. Common Sense

- **The Sinclair Anecdote:**

- Captain George Sinclair concluded, without numerical evidence, that a frightened peasant was hiding in a linden tree.
- Modern social science would demand further evidence, such as 95% certainty, before trusting this conclusion.
- Quantitative researchers would rely on statistical tests, not just experience or common sense.

- **Outcome of the Incident:**

- Sinclair's intuition was correct, but the situation ended tragically:
  - About 500 Norwegian farmers ambushed the Scottish mercenaries.
  - George Sinclair was killed, and the surviving Scotsmen faced a grim fate.

- **Key Lesson:**

- The social sciences increasingly rely on data, statistical tests, and evidence rather than intuition alone.
- This reflects the shift from common sense to a data-driven approach in understanding phenomena.

# The Logic Behind Statistical Inference

- **Generalization in Social Science:**

- Statistical methods allow generalizations about the empirical world.
- Properly defining the population and sample is essential.

- **Common Errors to Consider:**

- Sampling error, interviewer variability, non-response, and questionnaire problems (Groves, 1989).
- Context and assumptions underlying statistical models must be considered (John, 2002).

- **Impact of Sample Size:**

- Sample size greatly influences the ability to generalize.
- Raises questions:
  - “Does population size matter for significance?”
  - “What if we examine the whole population?”

# The Logic Behind Statistical Inference

- **Generalization in Social Science:**

- Statistical methods allow generalizations about the empirical world.
- Properly defining the population and sample is essential.

- **Common Errors to Consider:**

- Sampling error, interviewer variability, non-response, and questionnaire problems (Groves, 1989).
- Context and assumptions underlying statistical models must be considered (John, 2002).

- **Impact of Sample Size:**

- Sample size greatly influences the ability to generalize.
- Raises questions:
  - “Does population size matter for significance?”
  - “What if we examine the whole population?”

- **Central Limit Theorem:**

- As sample size  $N$  increases, the sampling distribution of the mean becomes approximately normal.
- The sampling distribution will fall around the variable's population mean.

- **Sampling Assumptions:**

- Units must be sampled randomly or with a known probability of selection.
- Stratified sampling divides the population into districts to closely examine subgroups.

- **Sample Size:**

- Large samples ( $N = 1000-1200$ ) make it easier to obtain significant results compared to small samples ( $N = 25$ ).

# Causation, P-Values, and Regression

- **P-Values:**

- Denote the probability of being mistaken when rejecting the null hypothesis.
- The closer the p-value is to 0, the more certain we can be about the hypothesis.

- **Correlation vs. Causation:**

- Correlations observed in regression analysis do not imply causation.
- Observed relationships must be interpreted using theories about human action (Elster, 1989).

- **Experimental Method:**

- Provides the best means to assess causal relationships by manipulating environments.
- Ensures that discovered relationships are not influenced by context (Moses and Knutsen, 2012). We will talk about this at the end of the course :).

# Population Size and Required Sample Size

- The size of the sample matters for making inferences, not the population size.
- A sample of 1000 is equally effective for small and large populations.
- The exception: If the sample exceeds a few percent of the total population, confidence intervals shrink.

Population	Sample
10	10
50	44
100	80
200	132
500	217
1000	278
3000	341
100,000+	385

Table: Sample size needed for a 95% confidence interval

# Why Use Significance Levels When Examining the Whole Population?

- **Investigating the Whole Population:**

- Social scientists may analyze an entire population (e.g., all fast food restaurants in a city).
- Unlike sample theory, this follows **stochastic model theory**, which generalizes from observed data to the underlying process.

- **Stochastic vs. Sample Theory:**

- Sample theory generalizes from a sample to a population.
- Stochastic models recognize that results vary due to randomness, even under constant conditions.

- **Why Significance Levels Matter:**

- Even when studying a whole population, observed relationships may result from random processes.
- Confidence intervals and significance levels help distinguish real associations from random chance.
- A lack of statistical significance suggests the observed association is as likely due to chance as to an actual relationship (Gold, 1969; Henkel, 1976).



# What is Science? ... and Scientific Claim?



# General Laws and Theories

- **Challenges in Establishing Causal Inference:**

- Hume: No finite amount of experimentation can prove that  $X$  causes  $Y$ .
- Popper: A proposition is scientific only if it is **falsifiable**.
- Social scientists should aim for conclusions that align with both theory and common sense Mayo (1980).

- **The Role of Theories:**

- Theories explain social behavior and must:
  - Define constructs.
  - Describe causal relationships.
  - Apply across different settings and times (Smith and Mackie, 2000).

- **Statistical Laws vs. Universal Laws:**

- Hempel (1959):
  - Statistical laws provide probabilistic explanations.
  - Universal laws predict a specific outcome whenever conditions are met.

- **Limitations of Statistical Research:**

- Patterns exist in nature but measurement errors affect accuracy.
- Proxies in surveys may not perfectly capture what we measure.
- Strong theoretical reasoning must support statistical findings.

# Writing a Quantitative Research Paper

- **Importance of Quantitative Research:**

- A large portion of social science research is quantitative.
- Statistical skills improve your chances of publishing in academic journals.

- **Key Components:**

- **Problem Statement:** Formulate a research question or testable hypothesis.
- **Research Methods:** Describe your sample, population, and variables.
- **Descriptive Statistics:** Include  $N$ , mean, standard deviation, skewness, and kurtosis.
- **Results Section:** Present regression tables and summarize findings.

Table: Regression Model of Welfare Attitudes

Variable	$\beta$	Std. E	p-value
Constant	4.664	0.195	0.000
Eastern European	0.391	0.292	0.203
Asian	0.631	0.364	0.107
$R^2$	0.376		
$N$	19		

# Questions

1. What is the main advantage of large-N studies compared to small-N studies?
2. In what way does population size matter when it comes to statistical inference?
3. What is the purpose of a sensitivity analysis?

# References I

- Elster, J. (1989). *Nuts and Bolts for the Social Sciences*. Cambridge University Press, Cambridge.
- Gold, D. (1969). Statistical tests and substantive significance. *American Sociologist*, 4(1):42–46.
- Groves, R. (1989). *Survey Errors and Survey Costs*. Wiley, New York.
- Hempel, C. (1959). The logic of functional analysis. In Gross, L., editor, *Symposium on Sociological Theory*, pages 271–307. Harper & Row, New York.
- Henkel, R. (1976). *Tests of Significance*. Sage, Beverly Hills, CA.
- John, P. (2002). Quantitative methods. In Marsh, D. and Stoker, G., editors, *Theory and Methods in Political Science*. Palgrave, New York, 2 edition.
- King, G., Keohane, R., and Verba, S. (1994). *Designing Social Inquiry: Scientific Inferences in Qualitative and Quantitative Research*. Princeton University Press, Princeton, NJ.
- Laffers, L. (2021). *Draft poznámok k predmetu Moderná Aplikovaná regresia 1*. UMB Banská Bystrica.
- Mayo, D. (1980). The philosophical relevance of statistics. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1980(1):97–109.
- Mehmetoglu, M. and Jakobsen, T. G. (2022). *Applied Statistics using Stata: a Guide for the Social Sciences*. Sage.

## References II

- Moses, J. and Knutsen, T. (2012). *Ways of Knowing: Competing Methodologies in Social and Political Research*. Palgrave, Basingstoke, 2 edition.
- Smith, E. and Mackie, D. (2000). *Social Psychology*. Psychology Press, Philadelphia, 2 edition.