UNIVERSITY OF MANNHEIM
School of Social Sciences

Quantitative Methods in Political Science: Introduction

Thomas Gschwend | Oliver Rittmann | Viktoriia Semenova | David M. Grundmanns

Week 1 - 8 September 2021

Schedule

- Round of Introduction
- Overview
- Data Analysis: Why Care?
- Some Basics of Data Analysis

Overview

Roadmap

- Understand and model stochastic processes
- Understand statistical inference
- Implement it mathematically and learn how to estimate it
  - OLS
  - Maximum Likelihood
- Implement it using statistical software
  - Learn how to use R
  - Basic programming skills

Data Analysis: Why Care?

What careful data analysis can reveal...

- Consider the following table on graduate school admissions at the UC Berkeley in 1973:[1]

	Male	Female
Admit	1198	557
Reject	1493	1278

- Why is the message of this 2x2 contingency table hard to decipher?

[1] Source: Bickel et al. 1975.

What careful data analysis can reveal…

- Let's convert the raw numbers into percentages.
- Based on this information, what are your conclusions now?

	Male	Female
Admit	45%	30%
Reject	55%	70%
Total	100%	100%

- Is graduate admission at UC Berkeley biased towards men?

What careful data analysis can reveal…

- Some crucial information is missing: admission decisions are made by each department, not by the university.
- Let's disaggregate the data by department:

- What's going on?

What careful data analysis can reveal...

- The apparent correlation between gender and admission rates at the university level is what we call a *spurious correlation*.
- Different admission rates at the university level do not result from discrimination against women, but are an artifact of women applying to the more competitive departments with smaller intake. We call this *self-selection*.
- There is an `omitted variable`: department choice. When we ignore this information, women seem to be discriminated against, while in reality they just apply to those departments that are harder to get into.
- This phenomenon is known as *Simpson's Paradox*: An association or comparison that holds for groups at the aggregate, i.e., university level, can reverse direction when the data are disaggregated to the individual, i.e., departmental level.

What careful data analysis can reveal…

- The essence of statistical analysis is to allow for systematic *controlled comparisons*, because we are interested in examining effects holding other factors constant.
- In multivariate analysis, we speak of statistical control: "holding constant", "controlling for", "accounting for", "ceteris paribus", …
- What is the (causal) effect of *X* on *Y*? Randomized experiments are often considered the gold standard for *causal inference*. To come as close as possible to this ideal with observational studies, we seek to control statistically for confounding factors.
- Without controlling for such confounding factors, our statistical models are misspecified, and inferences are biased.
- There is a substantial difference between *correlation* and *causation*. To identify causal effects instead of only correlational relationships we need good theories of social inquiry.
- There is no statistical method that can determine the causal story from the data alone.

Correlation vs. Causation

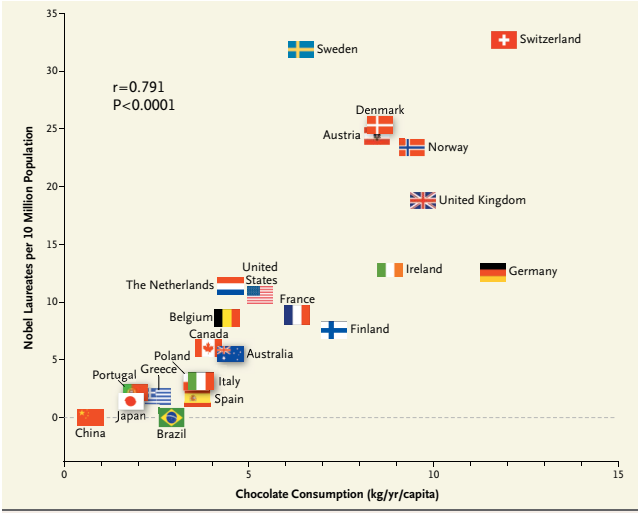


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Some Basics of Data Analysis

Key Concepts and Terms

- Unit of analysis: The observation described by a set of data. For example, voters, parties, bills, elections, voting decisions, legislative output. Very often our data have multiple levels of analysis (e.g., individuals, regions, countries), calling for different statistical techniques.
- Variables: Any characteristic related to the unit of analysis. A variable can take on different values for different observations.
- Types of variables: e.g., nominal (e.g., gender), ordinal (e.g., school grades), interval (e.g., GDP), ratio (e.g., duration).
- Data set: Set of variables for a given set of observations. Should come with a codebook.
- Hypothesis: Statement about the nature of the social and political world, often expressed as statements about relationships between variables (e.g., "The lower *X*, the higher *Y*").
- Measurement: Refers to the way in which variables are quantified (e.g., economic wealth measured as GDP).

Types of Data in Political Science

- Cross-section data: Sample of voters, governments, countries, or other units, taken at a *given point in time*. Observations are typically assumed to be independent.
- Time series data: Observations on units *over time*, e.g., number of conflicts in country X. Because past events can influence future events and lags in behavior are prevalent in social sciences, time is an important dimension in such a data set. Observations are not independent across time (serial correlation).

Types of Data in Political Science

- Pooled time series cross-section data: Data consist of comparable time series data observed on a variety of units. For instance, units are countries, and for each country we observe annual data on a variety of political and economic variables. Typically, we have few units, but long time series. Pooling the data increases the number of observations and makes it possible to control for exogenous shocks. Observations are usually not independent.
- Panel data: A large number of the *same* cross-sectional units, e.g., survey respondents, are observed repeatedly over a number of "waves" (interviews). With panel data, the time series is usually very short. Common in studies of political behavior. For example, German Socio-Economic Panel (SOEP) or the GIP (German Internet Panel) in Mannheim.

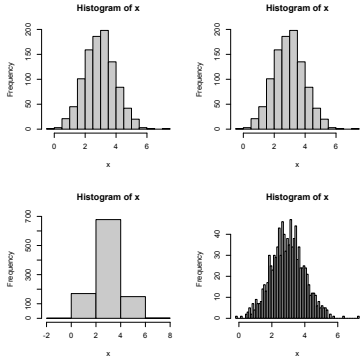
Examining and Visualizing Data

Statistical graphs are central to effective data analysis in the early stages of an investigation, for statistical modeling, and for the presentation of your results.

- Examine the characteristics of each variable's distribution separately. (Hint: Probability theory will become relevant to us!)
- The distribution of a variable tells us what values it takes, and how often it does.
- Summary statistics summarize the key characteristics of the variable/ the data.
- Summary statistics can be represented graphically, numerically, or both.

Visualizing Data: Histogram

A *histogram* shows the distribution of the measurements of a variable. A histogram is a bar graph in which the height of the bar shows how many observations fall in particular subintervals (bins), plotted along the horizontal axis.

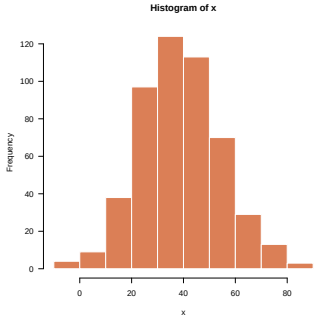


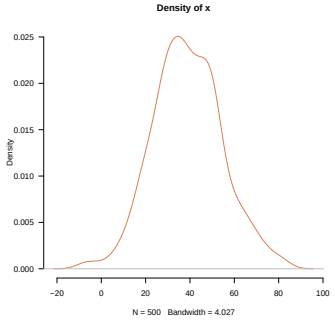
- In constructing histograms, we want to have enough bins to preserve some detail, but not too many, so that the display is dominated by sampling variation and captures too many idiosyncratic variation.
- The form of the histogram depends on the arbitrary width of the bins.

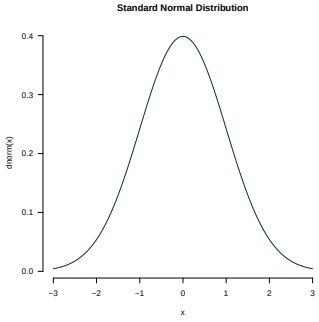
Visualizing Data: Density Plots

We can also plot the (empirical) frequency distribution as a density plot. Density plots address the deficiencies of histograms by averaging and smoothing.

Density plots are also preferred when we examine probability distributions for continuous random variables. The density of the probability is described by a mathematical formula $f(X = x)$, called the probability density function (PDF) for the random variable X.



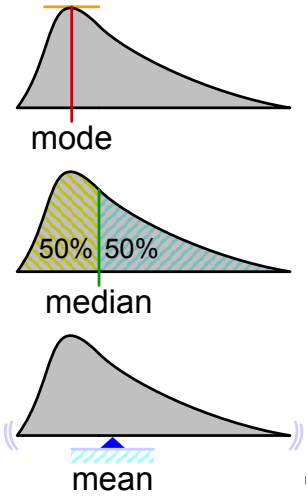




Describing Data: Measures of Central Tendency and Variability

- Central tendency: mean, median, mode
- Variability: variance, standard deviation, range, IQR
- Visualization: boxplots

Describing Data: Central Tendency



- Mode: Most frequently occurring value of X. Some distributions can have more than one mode.
- Median: Value of X that falls in the middle position when the observations are ordered from smallest to largest. Median = 50th percentile = 2nd quartile.
- Mean: Most common measure of central tendency. Arithmetic mean:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Describing Data: Comparing Measures of Central Tendency

- In a perfectly symmetric distribution, e.g., normal distribution: mode $=$ median $=$ mean
- Not true when distribution is non-symmetric.
  - right-skewed distribution (positive skew): median $<$ mean
  - left-skewed distribution (negative skew): median $>$ mean
- Mean is sensitive to outliers, while the median is more robust.

Describing Data: Variability

- Sample Variance: Average of the squared deviations from the mean:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Why do we average by dividing by n-1? Sum of deviations is always zero. Thus, the last deviation can be found once we know the other n-1. So we are not averaging n unrelated numbers. Only n-1 squared deviations vary freely, these are called *degrees of freedom* of the variance.

- Standard Deviation: Square-root of sample variance:

$$s = \sqrt{s^2}$$

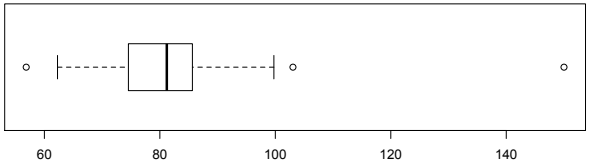
- Range: Difference between largest and smallest measurement:

$$RANGE = x_{Max} - x_{Min}$$

- Interquartile Range (IQR): Difference between upper and lower quartiles (range of the middle 50% of the distribution)

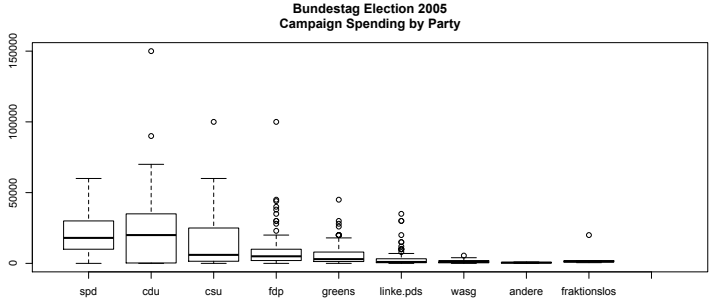
$$IQR = x_{Q3} - x_{Q1}$$

Describing Data: Boxplots



- Shows summary statistics of a variable's distribution. Useful when we require a compact (visual) representation of a distribution.
- The boxplot shows: Q1, Median, Q3 and "whiskers" from the Q1/Q3 to the smallest and largest observation that are not outliers. Outliers are defined as observations outside the fence: Q1-1.5(IQR) and Q3+1.5(IQR). Dots mark any observation outside the whiskers.

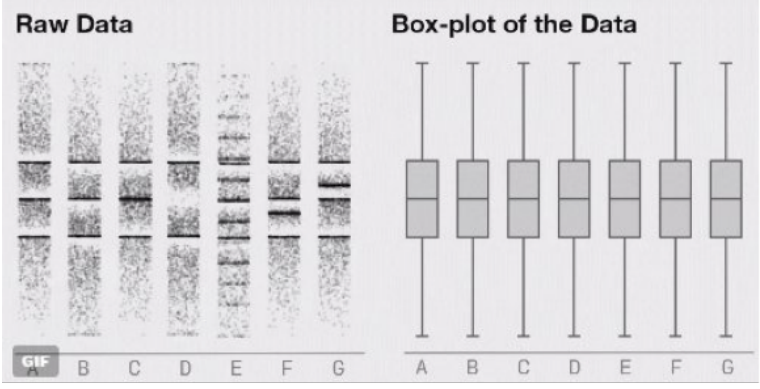
Describing Data: Parallel Boxplots



Parallel boxplots display the relationship between a quantitative response variable and a discrete (categorical or quantitative) explanatory variable.

Example: Campaign Spending of Bundestag Candidates (in Euro) by Party.

Describing Data: Even Simple Plots Are Data Reductions!



Describing Data: The Scatterplot for Bivariate Data

- The scatterplot is a direct representation of observations on two quantitative variables. It is simple and the most useful of all statistical graphs.
- Variable plotted on horizontal axis: X, variable plotted on vertical axis: Y.
- We can describe the relationship between two variables using the patterns shown in the scatterplot.

Questions we should keep in mind when investigating a scatterplot:

- What type of pattern do we see?
- How strong is the pattern?
- Are there any unusual observations? Labels?