

Quantitative Research Methods II

Vera E. Troeger

Office: 0.67

E-mail: v.e.troeger@warwick.ac.uk

Office Hours: by appointment

Quantitative Data Analysis

Descriptive statistics: description of central variables by statistical measures such as median, mean, standard deviation and variance

Inferential statistics: test for the relationship between two variables (at least one independent variable and one dependent variable)

For the application of quantitative data analysis it is crucial that the selected method is appropriate for the data structure:

- DV:
 - Dimensionality: spatial and dynamic
 - continuous or discrete
 - Binary, ordinal categories, count
 - Distribution: normal, logistic, poisson, negative binomial
- Critical points
 - Measurement level of the DV and IV
 - Expected and actual distribution of the variables
 - Number of observations and variance

Measurement Level

The appropriate method largely depends on the measurement level, type, and distribution of the dependent variable!

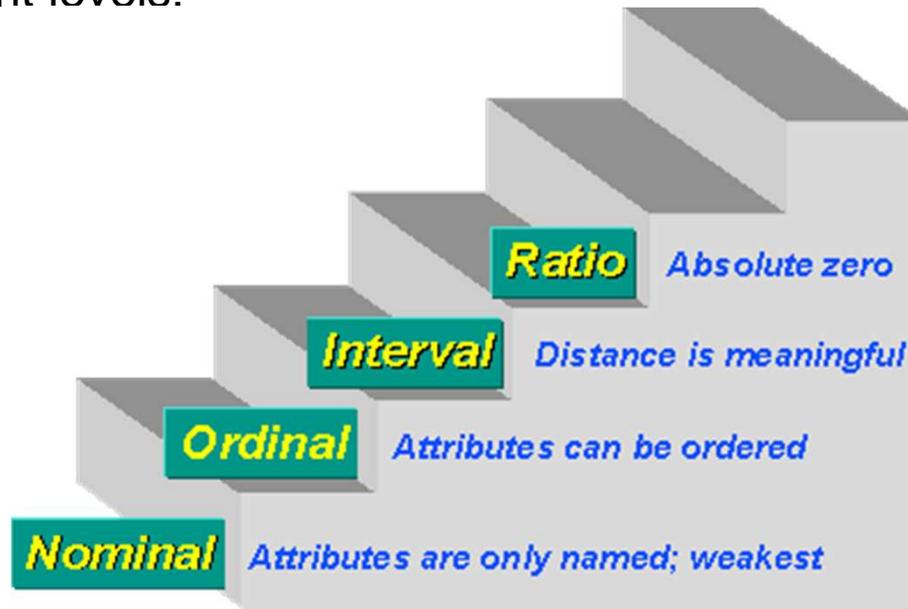
Measurement levels of variables:

The level of measurement refers to the relationship among the values that are assigned to the attributes for a variable.

- Nominal: the numerical values just "name" the attribute uniquely. No ordering of the cases is implied. For example, party affiliation is measured nominally, e.g. republican=1, democrat=2, independent=3: 2 is not more than one and certainly not double. (qualitative variable)

- Ordinal: the attributes can be rank-ordered. distances between attributes do not have any meaning. For example, on a survey one might code Educational Attainment as 0=less than H.S.; 1=some H.S.; 2=H.S. degree; 3=some college; 4=college degree; 5=post college. In this measure, higher numbers mean *more* education. But is distance from 0 to 1 same as 3 to 4? The interval between values is not interpretable in an ordinal measure. Averaging data doesn't make sense.
- Interval: distance between attributes *does* have meaning. E.g. temperature (in Fahrenheit), the distance from 30-40 is same as distance from 70-80. The interval between values is interpretable. It makes sense to compute an average of an interval variable. But: in interval measurement ratios don't make any sense - 80 degrees is not twice as hot as 40 degrees.
- Ratio: there is always an absolute zero that is meaningful. This means that one can construct a meaningful fraction (or ratio). Weight is a ratio variable. In applied social research most "count" variables are ratio: number of wars. But also other continuous variables like gdp or government consumption.

Measurement levels:



- It's important to recognize that there is a hierarchy implied in the level of measurement idea. At lower levels of measurement, assumptions tend to be less restrictive and data analyses tend to be less sensitive. At each level up the hierarchy, the current level includes all of the qualities of the one below it and adds something new. In general, it is desirable to have a higher level of measurement (e.g., interval or ratio) rather than a lower one (nominal or ordinal).
- Knowing the level of measurement helps you decide how to interpret the data from a variable and what statistical analysis is appropriate on the values that were assigned.

Variable types:

- Discrete vs. Continuous variables: A discrete variable is one that cannot take on all values within the limits of the variable. For example, responses to a five-point rating scale can only take on the values 1, 2, 3, 4, and 5. The variable cannot have the value 1.7. A variable such as a person's height can take on any value. Variables that can take on any value and therefore are not discrete are called continuous. – for statistical analysis it is important whether the dependent variable is discrete or continuous.
- Count variables: discrete – specific distribution, positive values, number of wars/ terrorist attacks, numbers of acqui communautaire chapters closed
- Binary variables: discrete, either 1 or 0, yes/no, Gender, parliamentary/presidential,
- Truncated variables: only observations are used that are larger or smaller than a certain value: analysis of the determinants of poverty – only poor people are analyzed
- Censored variables: values above or below a certain threshold cannot be observed: income categories
- Categorical variables: answering categories in surveys
- Nominal variables with more than 2 categories: party affiliation

The appropriate statistical model heavily depends on the type of the dependent variable: probit/logit models for binary variables, poisson/negative binomial models for count variables etc.

Distribution of variables

- Frequency distribution/ density: measures the frequency with which a certain value occurs in a sample
- Probability distribution/ density: measures the probability with which a certain value occurs in a population, the sum of probabilities equals 1
- Distributions are uniquely characterized by the determining parameters and their moments
- Moments are: mean, variance, skewness, kurtosis etc.
- We always distinguish between the “true value” and the sampling value of a moment
- 1st moment: central tendency of a distribution, most common is the *mean* (in a sample, also called expected value of a variable):

$$\mu = E(x) = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

- 2nd moment: “width” or “variability” around the central value, most common is the *variance* or its square root the standard deviation:

$$\sigma^2 = Var(x_1 \dots x_n) = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2; \quad \sigma = \sqrt{Var(x_1 \dots x_n)}$$

Descriptive Statistics

- Variables can be describe by some useful descriptive measures
- Mean: most common measure of central tendency of a distribution

$$\mu = E(x) = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

- Median: depicts the border between two halves: ordered sample – value that divides sample in halves: half or the observations are smaller and half are larger than the median: median is resistant towards outliers and is therefore better suited as measure for central tendency than the mean in case the distribution is not normal (or symmetric)

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & , n \text{ uneven} \\ 1/2 \left(x_{(n/2)} + x_{(n/2+1)} \right) & , n \text{ even} \end{cases}$$

- Modus/ Mode: the value with the highest frequency in a distribution, the value that has the highest probability of occurrence; only nominal level needed (ordinal for median, interval for mean)
- Minimal value, Maximal value
- Range: max value – min value
- Standard Deviation, Variance:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \quad \sigma = \sqrt{\sigma^2}$$

Example:

Variable: 0, 1, 10, 5, 0, 3, 7, 2, 1, 5, 6, 7, 4, 7, 1, 7, 10, 4, 9, 8, 7

Mean: 4.952

Median: 0, 0, 1, 1, 1, 2, 3, 4, 4, 5, 5, 6, 7, 7, 7, 7, 7, 8, 9, 10, 10

Mode: 7

Measures of association between two variables

- Co-variance: association between two variables, e.g. years of schooling (x) and earnings (y): positive value for proportional relationship (if schooling increases earnings do as well) and negative value for inverse proportional relationship (if schooling increases, earnings decrease):

$$\text{COV}_{xy} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

- Correlation: measures essentially the same as co-variance, but values range between -1 and 1; better measure than co-variance since it is a standardized measure and does not depend on the units the variables are measured in, e.g. if hourly earnings are measured in pence instead of pounds the co-variance will be affected but the correlation coefficient won't change.

$$\rho_{xy} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N \sqrt{(x_i - \bar{x})^2} \sum_{i=1}^N \sqrt{(y_i - \bar{y})^2}} = \frac{\text{COV}_{xy}}{\sigma_x \sigma_y}$$

- Both measures of association are flawed since they cannot take into account other variables that might influence the relationship. No notion of causality between x and y involved;

Datasets:

- Datasets contain dependent, independent and intervening variables for answering a specific research question/ testing specific theoretical propositions
- All variables in the dataset have the same dimensionality (observations for the same cases, units and time points)
- Variables in a dataset can have different measurement levels, types and distributions:
- Example: study of tax competition in OECD countries: dataset contains the DV: tax rates in OECD countries over time, IV: economic variables: government debt, gdp, unemployment, FDI, capital formation etc.; political variables: colour of the government party, election dates, capital restrictions, corporatism etc.

Data-types:

Dimensionality of the data

- Cross-sectional data: observations for N units at one point in time
- Time series data: observations for 1 unit at different points in time
- Panel data: observations for N units at T points in time: N is significantly larger than T – mostly used for micro data – units are individuals
- Time series cross section (TSCS) data: = panel data, but mostly used for macro data – aggregated (country) data
- Cross section time series (CSTS) data: observations for N units at T points in time: $T > N$

The names Panel, TSCS and CSTS are used interchangeably, however the distinction is useful since asymptotic characteristics of estimators are derived always for the larger dimension.

Micro Data: Individual Data

- Survey data: Eurobarometer, National Election Study (US), British Election Study, socio-economic panel (Germany and other countries)
- Individual Income (LIS), firm data
- Research questions from the fields of Political Behaviour, British Politics, Public Opinion and Polling

Macro Data: Aggregated Data on different levels

- Economic indicators: Inflation, Unemployment, GDP, growth, population (density) and demographic data, government spending, public debt, tax rates, government revenue, interest rates, exchange rates, income distribution, FDI, foreign aid, trade (exports/ imports), no of employees in different sectors etc.
- Political indicators: electoral system (majority, proportional), political system (parliamentary, presidential, federal), political institutions: CBI, exchange rate system (fixed, floating), federal courts, capital controls, number of veto players, regime type (democracy, autocracy), union density, labour market regulations, wage negotiation system (corporatism), human and civil rights, economic and financial openness, political particularism etc.