Research Methodology for PhDs

GBSB GLOBAL
BUSINESS SCHOOL

Session 8-1 Topics

- Quantitative research methods
  - -surveys, polls
  - -statistical analysis,

Quantitative Research

Quantitative research is a research strategy that focuses on quantifying the collection and analysis of data. It is formed from a deductive approach where emphasis is placed on the testing of theory, shaped by empiricist and positivist philosophies

Quantitative research methods are used to observe events that affect a particular group of individuals, which is the sample population

Surveys, Questionnaires

A survey is a systematic method for gathering information from (a sample of) entities for the purposes of constructing quantitative descriptors of the attributes of the larger population of which the entities are members.

Surveys are conducted to gather information that reflects population's attitudes, behaviors, opinions and beliefs that cannot be observed directly.

Surveys, Questionnaires

Questionnaire and survey measures are probably the most widely used research tools within the social sciences (Fife-Schaw, 1995). Their low cost, minimal resource requirements and potentially large sample-capturing abilities make them an attractive research method for academics and practitioners alike. However. the

Surveys – Qustionnaire Design

Questionnaire - a document containing questions and other types of items designed to solicit information appropriate for analysis.

- The format of a questionnaire can influence the quality of data collected.
- A clear format for contingency questions is necessary to ensure that the respondents answer all the questions in the questionnaire.
- The order of items and wording in a questionnaire can influence the responses given.
- Clear instructions are important for getting appropriate responses in a questionnaire.
- Questionnaires should be pretested before being administered to the study sample.

Survey Tips

- The form and meaning of questions should be appropriate to the project.
- The questions must be clear and precise. – Negative terms should be avoided – Double-barreled questions (multiple questions enclosed within one) should be avoided.
- Questions should be relevant to the respondent.
- Respondents must be competent and willing to answer the questions.
- The order and wording of questions should be set in a manner to avoid biased responses.

Survey/Poll Design Tips

Decide what the aim of your survey is: You firstly need to decide what the purpose, or aim, of your survey is. Having a clear idea about your survey's purpose will help you construct your survey and help obtain the information that you need and are interested in. One way to help decide the aim of your survey is to look at what other surveys have, or have not, done before. You can then use this information to identify something new that you would like you address in your study, and make a prediction about what you expect the outcomes of the survey will be.

Decide who your target population is: Based on your survey aim, you will next need to determine who will be completing your survey as well as consider how many participants from your target population will need to complete it. Knowing who your target population is and having a sufficient sample size, will ensure that you have enough accurate and relevant data to answer your survey aim.

Decide on your method: You will then need to choose the right survey design, timeframe and method that will allow you to answer your aim and research question most effectively. The survey design will be dependent on how and when the variables of interest need to be measured. You may consider quantitative or qualitative methods, and within this a longitudinal or cross-sectional design.

(c)https://www.supersurvey.com/Research

Survey/Poll Design Tips

Develop your survey questions: You then will need to write your survey questions. Each question you include needs to have a purpose and target one facet of what you are interested in. It is important when writing survey questions to avoid being repetitive because participants may become disengaged. It is a good idea to include a mix of closed and open-ended questions to understand focused responses and the motivations underlying those responses respectively.

Administer your survey: Once you have developed your survey, you can then send it to individuals in your target population for them to complete. Advertising by social media is a popular way to spread the word about your survey and encourage participants to complete it (if online) or contact you as the researcher (if interviews).

Collate and analyse the data: Once you have reached the target number of responses, you can collate and analyse your obtained responses. The analyses that you use will depend on the type of data that you have. If you have quantitative data this is typically analysed using statistics, whereas qualitative data is typically analysed by themes.

Draw conclusions: Using the results you obtain from your analyses, you can then interpret the data and draw your conclusions about whether the aim of your survey was addressed. You may use these conclusions to form the basis of new research.

(c)https://www.supersurvey.com/Research

Qualitative vs Quantitative Surveys

Aspect	Quantitative Survey Design	Qualitative Survey Design
Scale of Research	Large-scale research	Smaller-scale research
Type of Questions	Closed questions (e.g., multiple-choice, dichotomous response)	Open-ended questions in an interview format
Nature of Data	Numerical data analysed using statistics	Data analysed and reported in respondents' language (e.g., quotes)
Purpose	To obtain a general snapshot of trends in a population	To conduct in-depth analyses of motivations underlying responses
Time Frame Consideration	Time frame is crucial for design choice	Time frame is crucial for design choice
Study Designs	**Longitudinal Survey Study:** Administered at least twice to examine changes over time. **Cross-sectional Survey Study:** Administered once to assess responses at a specific point in time.	**Longitudinal Survey Study:** Administered at least twice to examine changes over time. **Cross-sectional Survey Study:** Administered once to assess responses at a specific point in time.
Methods of Survey Administration	Electronic methods (e.g., online questionnaires) preferred for their accessibility and speed. Phone or face-to-face interviews can be quicker but may require more time for analysis due to the need for transcription.	Similar considerations for methods of administration, with a note that qualitative methods may involve more detailed interviews that require significant analysis.

Computerized Methods

CAPI – computer-assisted personal interviewing, in which the computer displays the questions on screen, the interviewer reads them to the respondent and then enters the respondent's answer.

ACASI – audio computer-assisted self-interviewing, in which the respondent operates a computer, the computer displays the question on its screen and plays recordings of the questions to the respondent, who then enters his/her answers. CATI – computer-assisted telephone interviewing, which is the telephone counterpart to CAPI.

IVR – interactive voice response, the telephone counterpart to ACASI, in which the computer plays recordings of the questions to respondents over the telephone who then respond by using the keypad of the telephone or saying their answers aloud.

Web – internet surveys (e.g. Qualtrics, SurveyMonkey, Google/Office Forms), in which a computer administers the questions online.

Examples of Survey Types

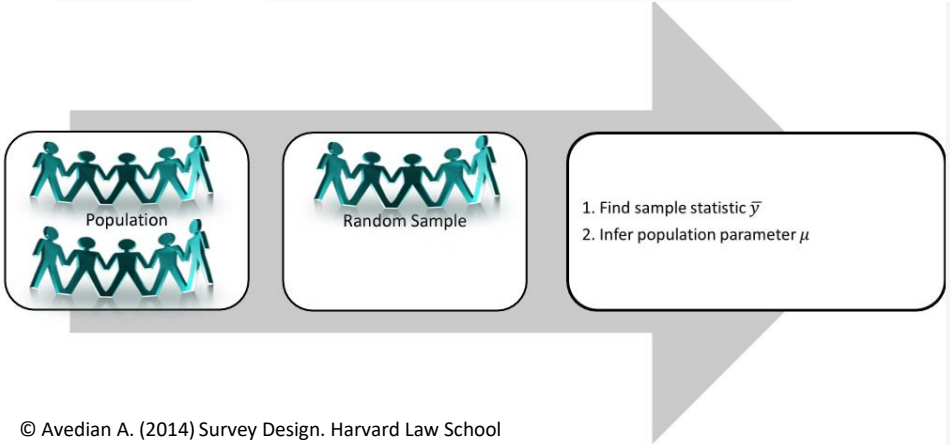
Cross-Sectional Surveys:	Longitudinal Surveys:	Descriptive Surveys:	Analytical Surveys:	Exploratory Surveys:	Explanatory Surveys:
•Conducted at a single point to analyze current trends or opinions within a specific population.	•Follow the same subjects over time to observe changes, making them perfect for studying trends and outcomes.	•Aim to describe the characteristics of a population or phenomenon, focusing on attitudes, beliefs, or behaviors.	•Seek to understand patterns or trends in the data, employing complex statistical methods.	•Used for new topics to gather preliminary information that helps define problems and suggest hypotheses.	•Designed to explain phenomena, often building on previous research to understand the causes of events or behaviors.

Defining Samples

Population: Is an entire collection of people, firms, states or things, that we are interested in, which we wish to describe, explain or predict.

Population distribution is usually unknown; we make inferences about its characteristics such as the parameter.

• Sample: A sample that is representative of the population that we actually observe and is used to infer about the population. Sample value we find from surveys is called statistic.



© Avedian A. (2014) Survey Design. Harvard Law School

Data Cleansing & Preparation

1. Duplicate data needs to be removed. The same data may be received from multiple sources. When merging the data sets, data must be de-duped.

2. Missing values need to be filled in, or those rows should be removed from analysis. Missing values can be filled in with average or modal or default values.

3. Data elements may need to be transformed from one unit to another. For example, total costs of health care and the total number of patients may need to be reduced to cost/patient to allow comparability of that value.

4. Continuous values may need to be binned into a few buckets to help with some analyses. For example, work experience could be binned as low, medium, and high.

5. Data elements may need to be adjusted to make them comparable over time. For example, currency values may need to be adjusted for inflation; they would need to be converted to the same base year for comparability. They may need to be converted to a common currency.

Data Cleansing & Preparation (Cntd)

6. Outlier data elements need to be removed after careful review, to avoid the skewing of results. For example, one big donor could skew the analysis of alumni donors in an educational setting.

7. Any biases in the selection of data should be corrected to ensure the data is representative of the phenomena under analysis. If the data includes many more members of one gender than is typical of the population of interest, then adjustments need to be applied to the data.

8. Data should be brought to the same granularity to ensure comparability. Sales data may be available daily, but the salesperson compensation data may only be available monthly. To relate these variables, the data must be brought to the lowest common denominator, in this case, monthly.

9. Data may need to be selected to increase information density. Some data may not show much variability, because it was not properly recorded or for any other reasons. This data may dull the effects of other differences in the data and should be removed to improve the information density of the data.

Privacy and Ethics

- keep confidential private information about survey participants,
- minimize the possibility of causing psychological discomfort or harm to respondents,
- when possible use paper-based, self-administered questionnaires (SAQ), instead of face-to-face survey to elicit information of a sensitive nature.

Math Statistics Fundamentals

- Average, Median, Mode
- Dispersion, Deviation

Average, Median, Expected Value (Mean)

Average - the numerical result obtained by dividing the sum of two or more quantities by the number of quantities; an arithmetic mean (2+4+7+9)/4=5.5

Median - is the value separating the higher half from the lower half of a data sample Median for {1,1,3,9} is **2**

Expected value (also known as EV, expectation, average, or mean value) is a long-run average value of random variables. It also indicates the probability-weighted average of all possible values**.**

Dispersion

Dispersion - (*or variability, scatter, or spread*) is the extent to which a distribution is stretched or squeezed.

Dispersion is a way of describing how spread out a set of data is. When a data set has a large value, the values in the set are widely scattered; when it is small the items in the set are tightly clustered.

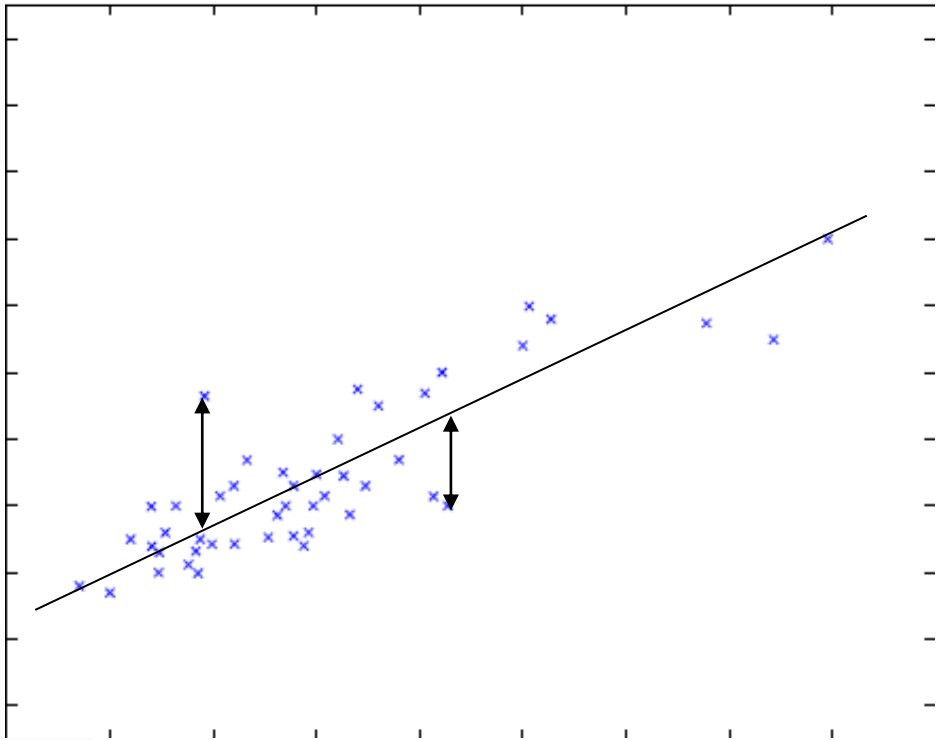
$$D(X) = M\left[(X - M(X))^2\right]$$

Deviation (Standard deviation) - is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the Dispersion

$$\sigma(X) = \sqrt{D(X)}$$

Dispersion Example

House cost



House square

Session 8-2 Topics

- Fundamental Quantitative research methods
  - Correlation Analysis
  - Regression functions

Correlation

a relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone

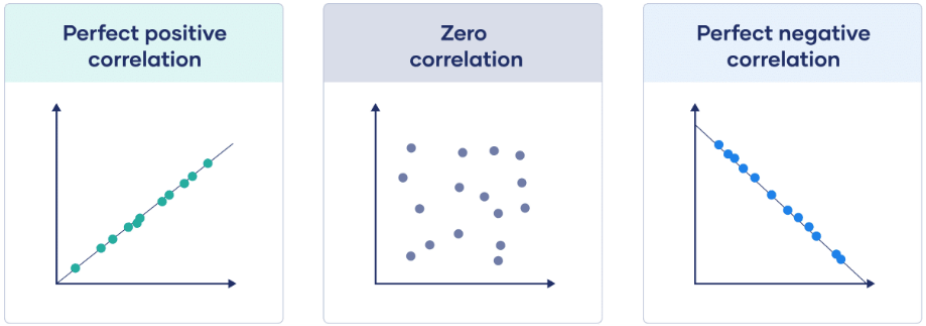
Correlation

Correlation is a statistical relationship between two or more independent data sets.

Sales, Euro	Discount, Euro	Sales, pieces
1160716	47379	367
1083892	40471	357
1307516	55301	487
1101605	46982	390
998808	39210	319
1176649	48949	370
1361669	89354	323
1533370	98453	374
1786288	89106	394
1772433	82260	476
1747858	91276	492
3146339	167060	680

Correlation Types

Correlation coefficient value	Correlation type	Meaning
1	Perfect positive correlation	When one variable changes, the other variables change in the same direction.
0	Zero correlation	There is no relationship between the variables.
-1	Perfect negative correlation	When one variable changes, the other variables change in the opposite direction.



(c)https://tagvault.org/blog/types-of-correlation/
(c)https://www.scribbr.com/statistics/correlation-coefficient/

Pearson Correlation coefficient

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sum_{j=1}^{n}\left(Y_j - \overline{Y}\right)^2}},$$

n – number of elements in a data set

$\overline{X}$ $\overline{Y}$ - mean (average) for X, Y datasets

-1<r<1

If (R>0.5 or R<-0.5), correlation exists, or use $\mathbf{R^2}$>0.25

r>0,5 – positive correlation

r < -0,5 – negative correlation

R coefficient works worse in case of data anomaly, fluctuations

Coefficient of determination	Explanation
r^2	The correlation coefficient multiplied by itself

Pearson Correlation coefficient (Cntd)

Strength of Correlation (Absolute Value of r)	Alternative Interpretation
0 – 0.3	Weak correlation
0.3 – 0.7	Moderate correlation
Above 0.7	Strong correlation

Assumptions your data must meet if you want to use Pearson's r:

- Both variables are on an interval or ratio level of measurement
- Data from both variables follow normal distributions
- Your data have no outliers
- Your data is from a random or representative sample
- You expect a linear relationship between the two variables

Madrid Campus: C/Numancia 6, 28039 Madrid
Barcelona Campus: Carrer d'Aragó, 179 08011 Barcelona
info@global-business-school.org
+34 930 086 588

Other Correlation Tools

Correlation coefficient	Type of relationship	Levels of measurement	Data distribution
Pearson's r	Linear	Two quantitative (interval or ratio) variables	Normal distribution
Spearman's rho	Non-linear	Two ordinal, interval or ratio variables	Any distribution
Point-biserial	Linear	One dichotomous (binary) variable and one quantitative (interval or ratio) variable	Normal distribution
Cramér's V (Cramér's φ)	Non-linear	Two nominal variables	Any distribution
Kendall's tau	Non-linear	Two ordinal, interval or ratio variables	Any distribution

©https://www.scribbr.com/statistics/correlation-coefficient/

Spearman's RHO

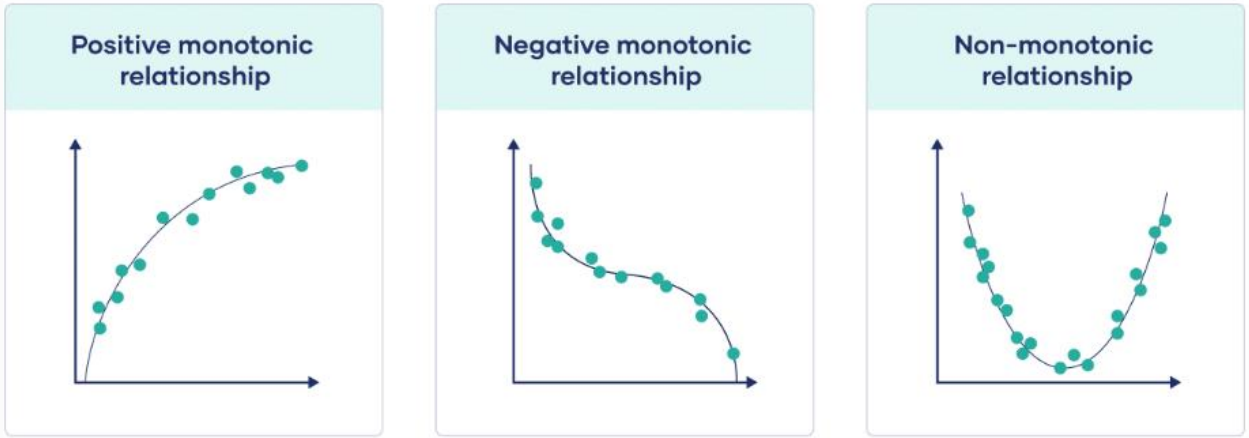
Spearman's rank correlation coefficient, is a rank correlation coefficient because it uses the rankings of data from each variable (e.g., from lowest to highest) rather than the raw data itself.

You should use Spearman's rho when your data fail to meet the assumptions of Pearson's r.

This happens when at least one of your variables is on an ordinal level of measurement or when the data from one or both variables do not follow normal distributions.

©https://www.scribbr.com/statistics/correlation-coefficient/

Spearman's RHO Application Areas



©https://www.scribbr.com/statistics/correlation-coefficient/

Spearman's RHO

Spearman's rank correlation coefficient formula	Explanation
$r_s = 1 - \frac{6 \sum d_i^2}{(n^3 - n)}$	• r_s= strength of the rank correlation between variables • d_i = the difference between the x-variable rank and the y-variable rank for each pair of data • $\sum d^2_i$ = sum of the squared differences between x- and y-variable ranks • n = sample size

©https://www.scribbr.com/statistics/correlation-coefficient/

Tools for Statistical analysis

- Microsoft® Excel, and special Add-in “Analysis Package”
- Google Spreadsheets
- IBM SPSS or PSPP
- QlickView
- Tableau

Madrid Campus: C/Numancia 6, 28039 Madrid
Barcelona Campus: Carrer d'Aragó, 179 08011 Barcelona
info@global-business-school.org
+34 930 086 588

Regression

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables

Regression models describe the relationship between variables by fitting a line to the observed data.

Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line.

Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Types of Regression

Linear

- (y=ax+b)

Non-Linear

- (i.e. y=b+a*exp(x), or y=a+b*ln(x), or y=a+b*sin(x), etc.

Multiple

- (y=a+bx1+cx2+…zxn, or y=a+b*ln(x1)+…z*ln(xn)

Newton – Rafton Algorithm

Repeat {select function type}

Repeat{

Change a= Θj , and b= Θi with step α -

- Repeat (j) {
- Θj = 0
  - For i=0 to N

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

  - If (J(Θj))<J(j) then J(j)=J(Θj)
  - Next I
- Inc(j)

}}

J = $\sum_i (y_i - f_i(x))^2 \to \min_x$

Regression Application Examples

Causality Assessment:

- While correlation does not imply causation, regression analysis provides a framework for assessing causality by considering the direction and strength of the relationship between variables.
- It allows researchers to control for other factors and assess the impact of a specific independent variable on the dependent variable.
- This helps in determining the causal effect and identifying significant factors that influence outcomes.

Model Building and Variable Selection:

- Regression analysis aids in model building by determining the most appropriate functional form of the relationship between variables.
- It helps researchers select relevant independent variables and eliminate irrelevant ones, reducing complexity and improving model accuracy.
- This process is crucial for creating robust and interpretable models.

Hypothesis Testing:

- Regression analysis provides a statistical framework for hypothesis testing.
- Researchers can test the significance of individual coefficients, assess the overall model fit, and determine if the relationship between variables is statistically significant.
- This allows for rigorous analysis and validation of research hypotheses.

Regression Pro vs Con

Advantages of Regression Analysis	Disadvantages of Regression Analysis
Provides a quantitative measure of the relationship between variables	Assumes a linear relationship between variables, which may not always hold true
Helps in predicting and forecasting outcomes based on historical data	Requires a large sample size to produce reliable results
Identifies and measures the significance of independent variables on the dependent variable	Assumes no multicollinearity, meaning that independent variables should not be highly correlated with each other
Provides estimates of the coefficients that represent the strength and direction of the relationship between variables	Assumes the absence of outliers or influential data points
Allows for hypothesis testing to determine the statistical significance of the relationship	Can be sensitive to the inclusion or exclusion of certain variables, leading to different results
Can handle both continuous and categorical variables	Assumes the independence of observations, which may not hold true in some cases
Offers a visual representation of the relationship through the use of scatter plots and regression lines	May not capture complex non-linear relationships between variables without appropriate transformations
Provides insights into the marginal effects of independent variables on the dependent variable	Requires the assumption of homoscedasticity, meaning that the variance of errors is constant across all levels of the independent variables

Thank You!
Read the Recommended Readings
You're welcome with your discussions and questions in VLE!

Please note, that since the recordings are done, some Readings may become unavailable. Inform us immediately in VLE, so we can offer substitutions