Survey Data Analysis

May 24, 2010

Dr. Abdoulaye Diop (adiop@qu.edu.qa) - Qatar University

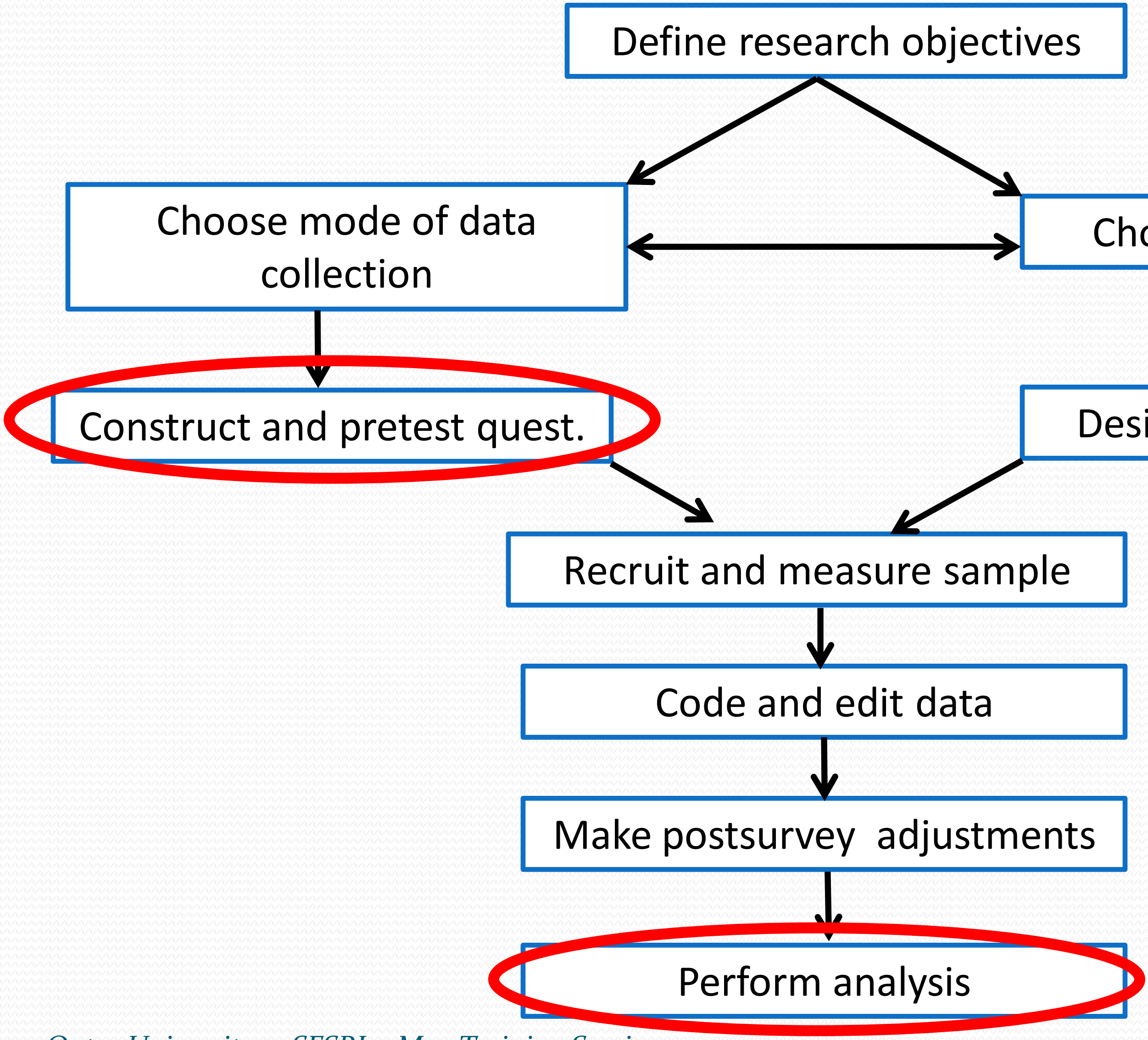
Dr. Kenneth M.Coleman (Ken.Coleman@marketstrategies.com) - University of Michigan

SESSION I

Survey Data Analysis

Compiled by Abdoulaye Diop,
SESRI, Qatar University, and
Kenneth M. Coleman,
University of Michigan

The Survey Research Process



Define research objectives

Choose mode of data collection

Choose sampling frame

Construct and pretest quest.

Design and select sample

Recruit and measure sample

Code and edit data

Make postsurvey adjustments

Perform analysis

Observation Units and Variables

Variable = measurable characteristic of an observation unit, which varies across different units

- Individual group (e.g. family, household, couple)
- Institution, organization or community (e.g. school,
- Enterprise, municipality)
- Text (e.g. newspaper article, research)
- Event or activity (rallies, strike)

Terminology and Notation

- Any numeric value describing a characteristics of a population is called a parameter, often represented by Greek letters (e.g., μ = population mean).

- Any numeric value describing a characteristics of a sample is called a statistic, often represented by English words or letters with symbols (e.g., $\overline{x}$ = sample mean of variable χ).

Survey Data Analysis

- **Data Preparation**
  (Cleaning and organizing the data for analysis)
- **Descriptive Statistics**
  (Describing the Sample Data)
- **Inferential Statistics**
  (Testing Hypotheses and Models)

Survey Data Analysis

- **Data Preparation**

  (Cleaning and organizing the data for analysis)

  - ✓ Data entry, data extraction (Data collection modes)
  - ✓ Develop and document a database structure (SPSS, STATA)
  - ✓ Checking data for accuracy
  - ✓ Transformation of the data / Recoding of variables
  - ✓ Evaluate missing values, Don’t knows & Refusals
  - ✓ Data weighting (probability of selection & post-stratification & non-response adjustments)

Survey Data Analysis

- **Descriptive Statistics**
  (Describing the Sample Data)

  - ✓ Describing the sample data by providing simple summaries about the sample and measures.

  - ✓ Presenting quantitative descriptions in a manageable form.

Survey Data Analysis

- **Inferential Statistics**
  (Testing Hypotheses and Models)
  - ✓ Investigate questions, models, hypotheses.
  - ✓ Infer from sample to population.

Levels of Measurement

There are 4 levels of measurement most often used in statistics:

- Nominal
- Ordinal
- Interval
- Ratio

(NOIR)

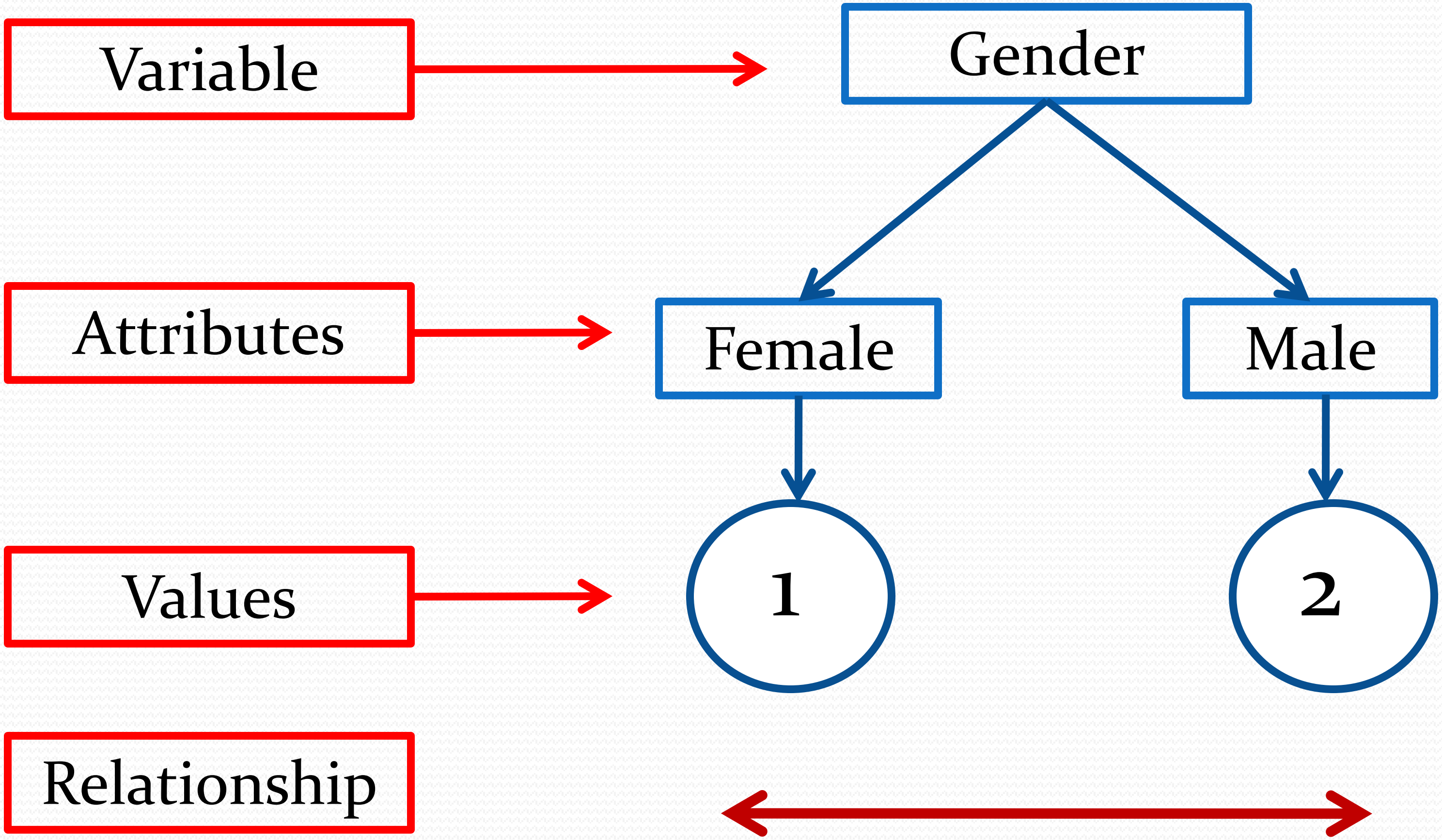
Levels of Measurement

Knowledge of the levels of measurement helps analysts decide on :

1. How to interpret data from that variable.

2. What statistical analysis is appropriate on the values that were assigned.

Levels of Measurement



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

Nominal Measurement

Classification of measurements into a set of categories

- Categories should be mutually exclusive and exhaustive.
- The numbers produced by nominal measurement are ***frequencies of occurrence*** in the categories (e.g., 22 females, 12 Males, etc).
- Nominal measurement applies to ***categorical (discrete) variables.***
- Nominal data is ***also*** termed qualitative data.

Ordinal Measurement

Rank ordering of elements on a continuum

- Ordinal measurement does not measure the amount of the variable or precise measures of the differences among the different ranks- e.g., first place, second place, last place in a race.

- Ordinal data can tell you that the person in 1st place finished ***before*** the person in 3rd place, but ***not by how much.***

Interval Measurement

All the properties of ordinal and nominal variables PLUS

- the number assigned reflects the amount of the variable.
- there are precisely defined intervals between and among the observations.
- the zero point is defined arbitrarily and does not represent an absence of the property being measured.

(e.g.: temperature, IQ measure)

Ratio Measurement

All the properties of the previous scales PLUS

- the number assigned reflects the amount of the variable.
- the size of the measurement unit remains constant.
- and the zero point represents an absence of the property being measured.

e.g.: Length, time

How to Determine the Level of Measurement

Questions	No	Yes
A) Do the categories of the variable have a rank order?	variable is nominal;	Go to B)
B) Are the "distances" or "intervals" between the categories meaningful?	variable is ordinal;	Go to C)
C) Does the scale have a meaningful zero?	Variable is interval;	Variable is ratio level measure.

Levels of Measurement

- Nominal: Observations can be named, and are distinguishable by names, but cannot be ordered by magnitude or rank.
- Ordinal: Observations can be ordered by relative magnitude, but precise differences between observations cannot be specified.
- Interval: Observations can be made with sufficient accuracy that precise distances between observations can be estimated.
- Ratio: Interval level measurement in the special case where an absolute zero point can be identified [a VERY rare occurrence in social sciences].

Types of Variables

DISCRETE/CONTINUOUS

- **Discrete variables**: can only assume certain values and there are usually “gaps” between values. Typically these variables can be ‘counted’.

  e.g.: the number of students in a classroom

- **Continuous variables:** can assume any value within a specific range.

  e.g.: the time it takes to travel from Villagio to Qatar University (QU)

Types of Variables

INDEPENDENT/DEPENDENT

- **Independent variables**: variables we use to explain a certain outcome. (in regression analysis these variables are often called predictor variables)

  e.g.: a person's number of years of schooling and his/her income

- **Dependent variables**: variables we are trying to explain.

  e.g.: income in the previous example

Univariate Analysis

Examination across cases of one variable at a time

(3 major characteristics of a single variable)

- The distribution
- Measures of central tendency
- Measures of dispersion

Distribution

Summary of the frequency of individual values for a variable:

- Listing the number or % of Male and Female for the Gender variable
- With large number of possible values group row scores into categories (age: 18-24; 25-34; 35-44; 45-54, 55-64; 65+)

Frequency Table (SPSS)

EDU14 The school prepares your children for university education

		Frequency	Percent	Valid Percent	Cumul %
Valid	1 Strongly agree	6	1.4	42.9	42.9
	2 Somewhat agree	4	.9	28.6	71.4
	3 Somewhat disagree	3	.7	21.4	92.9
	8 DON'T KNOW	**1**	**.2**	**7.1**	100.0
	Total	14	3.2		100.0
	Missing System	428	96.8		
Total		442	100.0		

- Frequency = simple count of the cases with a certain variable value
- Percent = percentage of the cases with a certain variable value
- Valid Percent = percentage of the cases with a certain variable value (excluding missing values)
- Cumulative Percent = percentage of the cases with the given or a smaller value, e.g. strongly agree (1) + somewhat agree (2) = 71.4%

Frequency Table (SPSS)

EDU14 The school prepares your children for university education

		Frequency	Percent	Valid Percent	Cumul %
Valid	1 Strongly agree	6	1.4	46.2	46.2
	2 Somewhat agree	4	.9	30.8	**76.9**
	3 Somewhat disagree	3	.7	23.1	100.0
	Total	13	2.9	100.0	
Missing	**8 DON'T KNOW**	**1**	**.2**		
	System	428	96.8		
	Total	429	97.1		
Total		442	100.0		

Measures of Central Tendency

Various ways to describe the central, most common or middle value in a distribution or set of data.

- Mode : Most frequently occurring scores.
  (the most common response)

- Median : The 50th percentile, the second quartile.
  (the middle response)

- Mean : Arithmetic means, average.
  (sum of all scores divided by n)

Dispersion

Dispersion refers to the spread of the values around the central tendency

- Variance (the mean of squared deviations from the mean)
- Standard Deviation (measure of spread around the mean – square root of variance)- the square root of the variance , characterizes the typical deviation from the mean
- Range [maximum value-minimal value]
- Interquartile Range (distance between first and third quartile)

Graphical Methods

Categorical/Discrete Variables

- Pie charts
- Bar graphs

Numeric Variables

- Histograms
- Line charts
- Box plots/box-and-whisker plots

Basic Choices in Data Display

- Most all statistical analysis program generate alternative graphical formats, and generally in color.
- Think about which type of graph works best to present the analytical point; and what your audience or readership may most easily understand.
- As will be argued later, too much “visual busy-ness” can detract from the point.

H. Weisberg, J. Krosnick & B. Bowen, An Introduction to Survey Research, Polling & Data Analysis, 3rd edition, Sage, 1996, pp. 196-197.

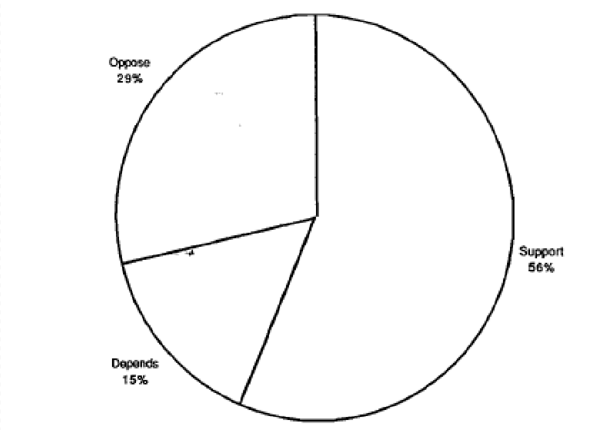


Figure 9.1. Pie Chart of Attitudes on Government Health Insurance

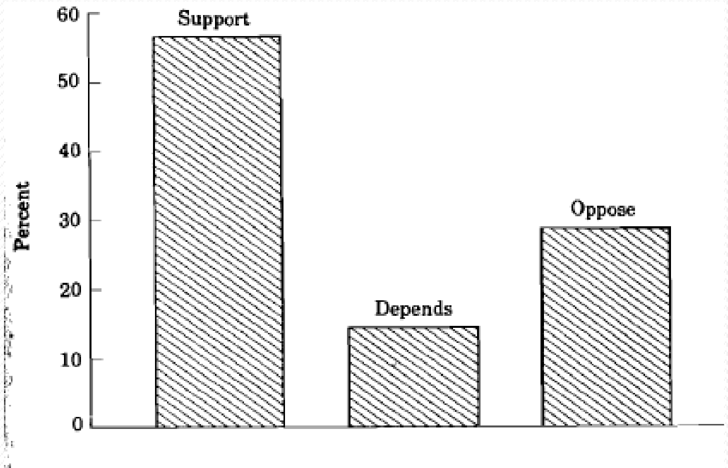


Figure 9.2. Bar Chart of Attitudes on Government Health Insurance

Objectives of Univariate, Bivariate and Multivariate Analysis

- Univariate analysis → Description. Description can "test hypotheses" on occasions when social expectations about characteristics are widespread, but remain empirically unexamined. In those cases, empirically founded description can tend to confirm or disconfirm social expectations.
- Bivariate analysis → Examining simple relationships; testing simple hypotheses with "no controls" for other possible interpretations.
- Multivariate analysis → Testing more complex sequences of multiple causation. Using tools that allow for assessment of indirect and direct empirical linkages, plus spurious relationships (apparent, but indirect or non-existent causal links). Multivariate analysis is very much about "controlling for" alternative possible interpretations – often via examining bivariate relationships while statistically "holding constant" other effects.

Levels of Measurement Determine the Statistical Tools to Which We Can Take Recourse

Nominal by Nominal: Measures of Statistical Significance: Chi Square. Measures of Strength of Association: Phi (for 2 x 2 tables), Cramer's V (2 x n tables), Gamma, and Lambda.

Nominal by Ordinal: Most of the above.

Ordinal by Ordinal: Measures of Strength of Association: Spearman's *rs* (*rho),* Kendall's rank-order τ (tau – better with many tied rankings).

Interval by Interval: Measures of Strength of Association: Pearson's r (correlation coefficient). Multivariate analytical techniques: Multiple regression, discriminant analysis, factor analysis, latent structure analysis, etc.

Measures of association can be used with nominal variables, as in the illustration below

Preferred Youth Sport	United Kingdom	United States	N
Youth Who Play Soccer/Football	88%	12%	100
Youth Who Play Baseball	6%	44%	50
Youth Who Play Neither	6%	44%	50
n	100	100	200

Lambda is a measure of association appropriate to two nominal variables. In this case Lambda = .32. The maximum value = 1.0, so this would represent only a moderate association (.32) between country (a nominal variable) and preferred youth sport (another nominal variable), although statistically significant.

Ordinal Data: The Likert Scale

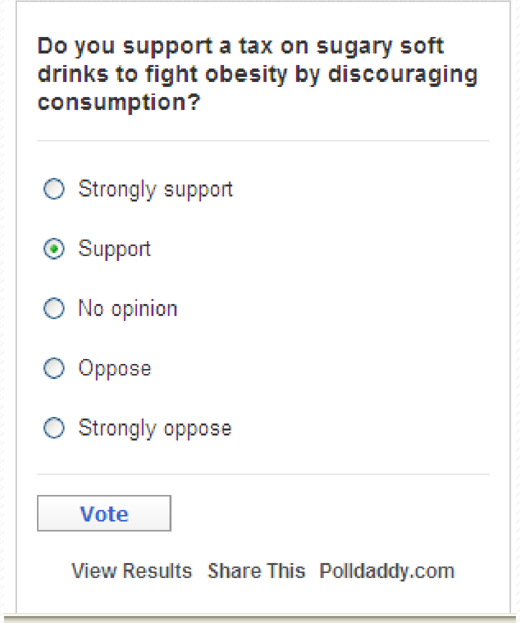
- The Likert Scale, discussed in the January training session, is a classic ordinal scale – in that one knows that certain answers exceed or fall short of others in intensity of affect (emotional charge) and the direction of those differences.
- However, one cannot specify precise differences of magnitude.

- Typical Likert Scale:
- Please tell me how you feel about the following statement: "I think the best sport in the world is competitive swimming."
  - Disagree Strongly
  - Disagree Somewhat
  - Neither Agree nor Disagree
  - Agree Somewhat
  - Agree Strongly

Uses of Ordinal Data: Description and Enhancement to Approximate Interval Scales

- Likert-style ordinal data presented in bar-graph form can often provide an instructive reading on important public policy questions.
- It is not uncommon for survey researchers to combine a number of Likert items, with numerical values attached (for example, from 1 to 5), to form an index that sufficiently approximates interval level measurement that it can be treated as such for the purposes of multivariate analyses. Typically, such a procedure implies that each question is weighted equally.

A Typical Likert Scale Item



Do you support a tax on sugary soft drinks to fight obesity by discouraging consumption?

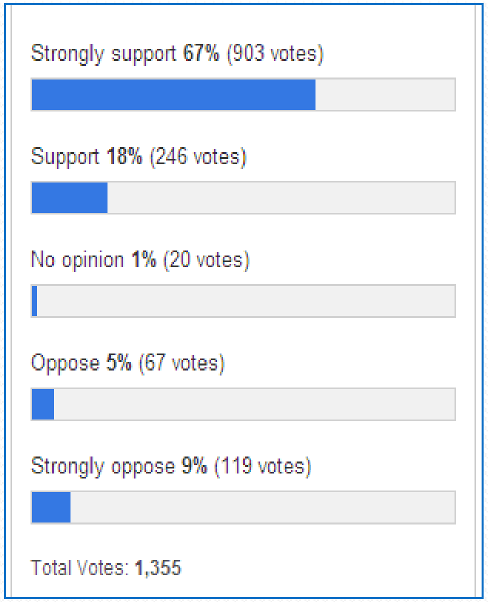
○ Strongly support

● Support

○ No opinion

○ Oppose

○ Strongly oppose



Strongly support **67%** (903 votes)

Support **18%** (246 votes)

No opinion **1%** (20 votes)

Oppose **5%** (67 votes)

Strongly oppose **9%** (119 votes)

Total Votes: **1,355**

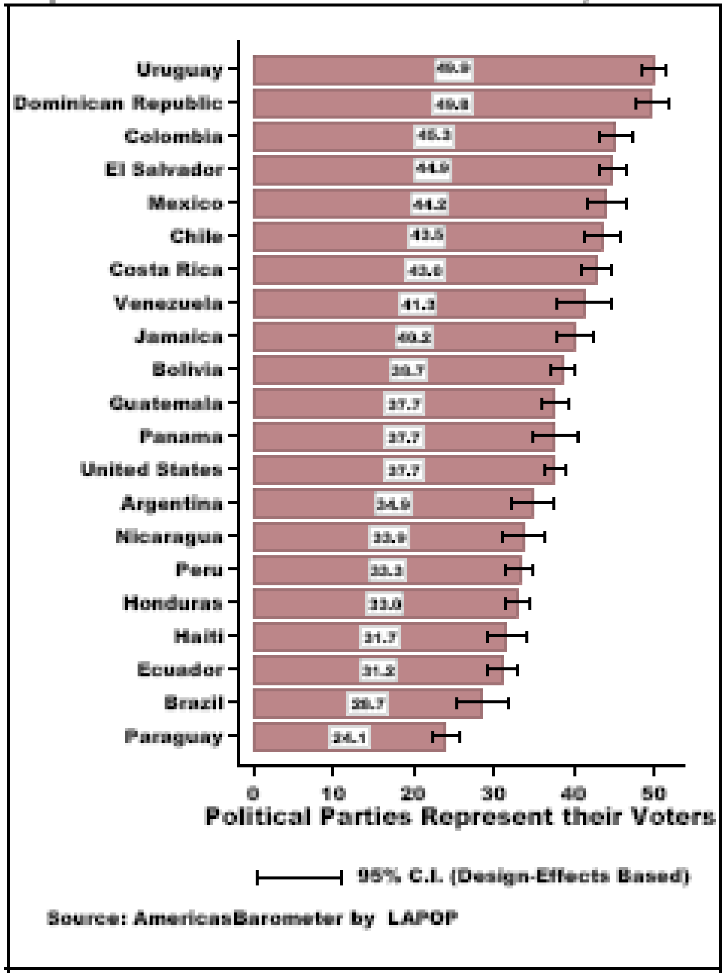
Note: While these are data from 2010, they should not be taken as representative of the views of the US public because they represent self-selected volunteers, who are not a representative sample of the total population of adults.

Confidence Intervals To Interpret Bivariate Data

- If a dependent variable can be measured at an interval level, use of graphic representations of confidence intervals around average scores can help to interpret visually whether nominal level independent variables, such as countries, actually do vary significantly.
- If confidence intervals <u>do</u> <u>not</u> <u>overlap</u>, then the mean values for any two countries differ significantly.
- Most statistical analysis programs today facilitate reporting of confidence intervals.

<u>Source</u>: Latin American Public Opinion Project, Vanderbilt University. AmericasBarometer Insight Series, No. 36. <u>Confidence intervals</u> = <u>+</u>2 σ (standard deviations)

Figure 1.
Average Agreement that Political Parties Represent their Voters in the Americas, 2008



Source: AmericasBarometer by LAPOP

Presenting Multivariate Results in Tabular Form: The Case of Multiple Regression

- The table at the right presents the results of a multiple regression equation in conventional form, where statistical significance is often indicated by a beta coefficient (*b*) 2x as large as the standard error.

- This particular regression equation, predicting a preference for private – rather than public provision of services – includes variables at the individual level of analysis, but also includes one's country as a contextual variable, comparing Costa Ricans and Chilean with the excluded baseline category of Mexicans.

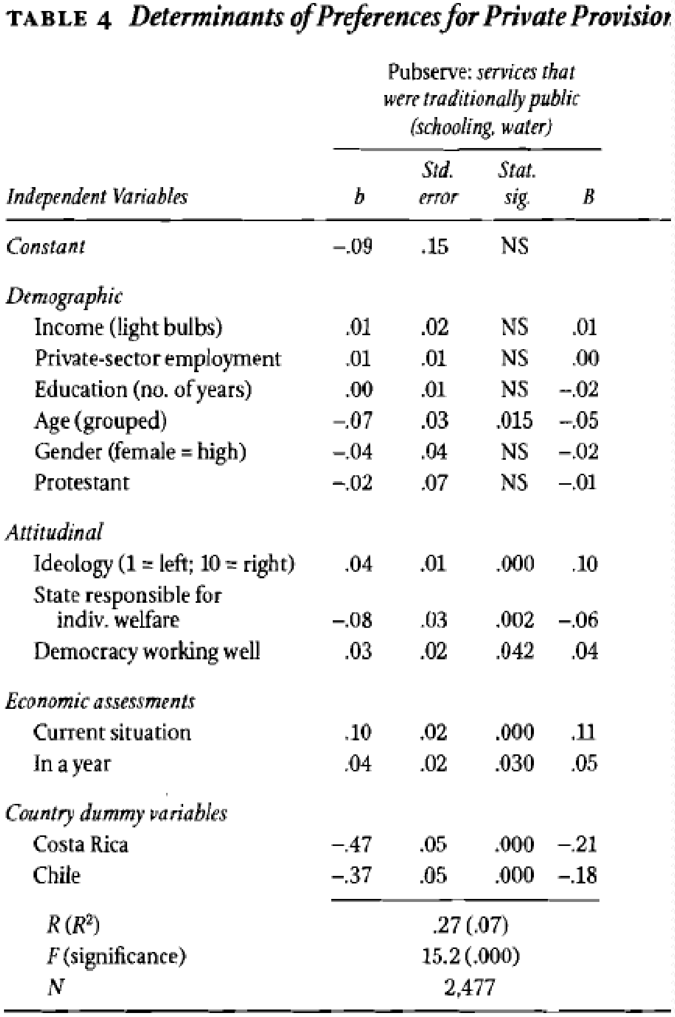


TABLE 4 *Determinants of Preferences for Private Provision*

	Pubserve: *services that were traditionally public (schooling, water)*			
Independent Variables	*b*	*Std. error*	*Stat. sig.*	*B*
Constant	–.09	.15	NS	
Demographic				
Income (light bulbs)	.01	.02	NS	.01
Private-sector employment	.01	.01	NS	.00
Education (no. of years)	.00	.01	NS	–.02
Age (grouped)	–.07	.03	.015	–.05
Gender (female = high)	–.04	.04	NS	–.02
Protestant	–.02	.07	NS	–.01
Attitudinal				
Ideology (1 = left; 10 = right)	.04	.01	.000	.10
State responsible for indiv. welfare	–.08	.03	.002	–.06
Democracy working well	.03	.02	.042	.04
Economic assessments				
Current situation	.10	.02	.000	.11
In a year	.04	.02	.030	.05
Country dummy variables				
Costa Rica	–.47	.05	.000	–.21
Chile	–.37	.05	.000	–.18
R (R^2)	.27 (.07)			
F (significance)	15.2 (.000)			
N	2,477			

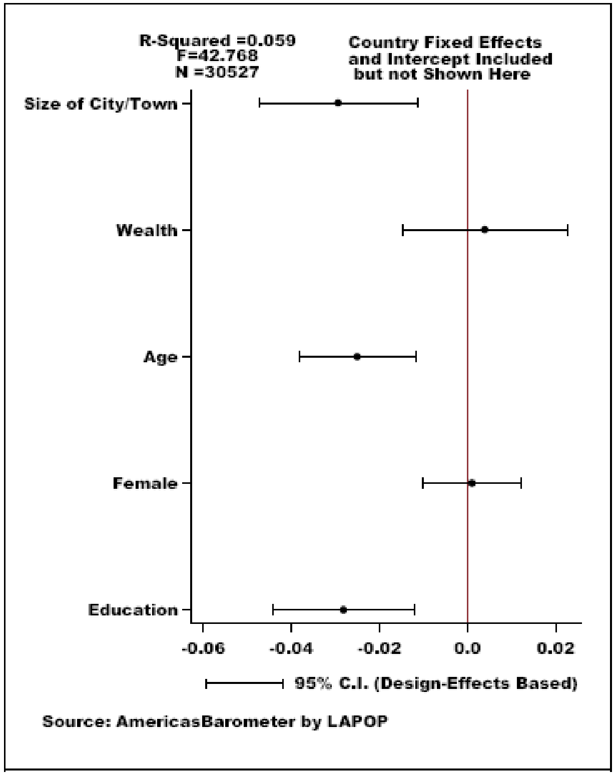
Source: Coleman, "Politics and Markets in Latin America," in Camp, R.A., Citizen Views of Democracy in Latin America. U Pittsburgh Press, 2001, p. 195.

Presenting Multivariate Analyses: The Case of Multiple Regression Results in Graphic Form

- The graphic to the right uses confidence intervals to indicate whether a regression coefficient (standardized β) is statistically significant.
- Coefficients that <u>do</u> <u>not</u> intersect with the Y axis in the center are statistically significant. This is easier for non-statistically sophisticated audiences to understand.
- In this illustration, young people, those from small towns and those with less education are significantly less likely to see political parties as representing the views of voters.

<u>Source</u>: Latin American Public Opinion Project, op cit., following example of Kastellec, J.P. and Leoni, E.L. "Using Graphs Instead of Tables in Political Science". *Perspectives on Politics* (2007), **5**:**4**:755-771.

Figure 2.
Socio-economic and Demographic Determinants of Support for the Belief that Political Parties Represent their Voters in Latin America, 2008



Source: AmericasBarometer by LAPOP

Visual Clarity

- Edward Tufte has dedicated much effort to improving graphic presentations in the social sciences. His general rules of thumb include these:
  - Graphical excellence I consists of complex ideas communicated with clarity, precision and efficiency
  - Graphical excellence is that which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
  - Graphical excellence is nearly always multivariate

In the prior example, we can see a couple of these principles at work:

- One can assess statistical significance easily and visually – rather than having to compare two numbers at once – or to recall that <.05 = statistical significance
- One can compare the effects of five independent variables on a dependent variable and judge their relative effects all at once.
- There is little wasted ink in the graph – for example, there are no grid lines, only X & Y axes and minimal identification of concepts. For the statistically inclined , details are posted in the upper left hand corner or at the bottom, away from the visual center of the graph.

Source: Tufte, E.R., The Visual Display of Quantitative Information, 1983, p. 31.

Measurement Exercise

- There is often a challenge in asking about sensitive topics, such as asking about family income as an indicator of social status. Sometimes people will not answer such questions. In the January training session, it was also mentioned that it is a good idea to have a variety of indicators or measures of most concepts .
  - Example: Asking only one indicator of social status, such as family income, my generate a variety of challenges, such as a general unwillingness to share information; or some respondents (females; those 21 year old, if in sampling frame) may not know family income). **So one would have missing data in those cases.**
  - Supplemental solution used in Mexico: Estimated number of light bulbs in the house, which is highly correlated with family income: 0-9; 10-19; 20-29; 30-39; 40-49; 50+ light bulbs in the house.
  - But would that work in Qatar or Bahrain? What would be a better surrogate indicator of social status here?

Additional References:

Fowler, F.J. (2004) Survey Research Methods, Fourth Edition. Thousand Oaks, CA: Sage.

Groves, R. M., Fowler, F.J., Couper, M. P., Lepkowski, J.M., Singer, E., & Tourangeau, R. (2009). Survey Methodology, Second Edition. New York: John Willey.

Research Methods Knowledge Base: http://www.socialresearchmethods.net/kb/index.php

Thank You!