Data Analysis

Session 1, June 22nd, 2010

Microsoft Research India Summer School on Computing for Socioeconomic Development

Discussion led by Aishwarya Ratan

Topics

- Types of analysis
- Analysing qualitative data
- Working with quantitative data
- Inference from data analysis
  - Validity
    - Internal
    - External
  - Causality

Types of data analysis exercises

- Seeing patterns in data (descriptive analysis)
- Testing a hypothesis
- Impact evaluation
  - Did intervention A cause outcome B?

What we should avoid…..

"If you can't prove what you want to prove, demonstrate something else and pretend that they are the same thing. In the daze that follows the collision of statistics with the human mind, hardly anybody will notice the difference."

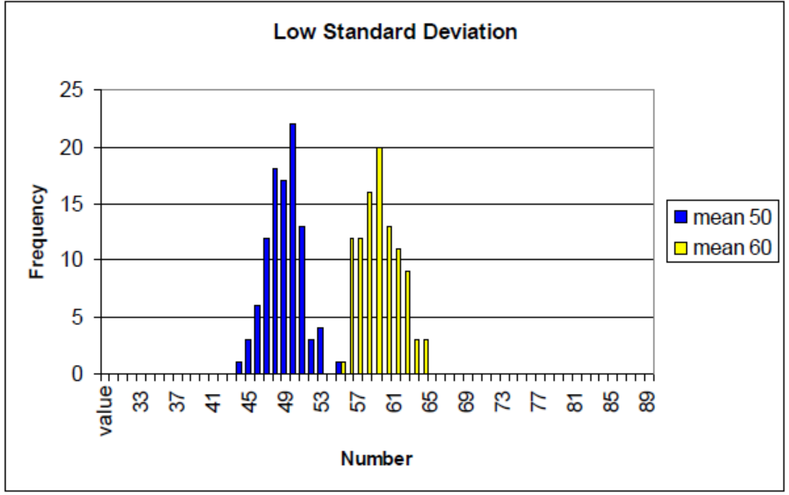
- *How to Lie with Statistics (Huff and Geis, 1993)*

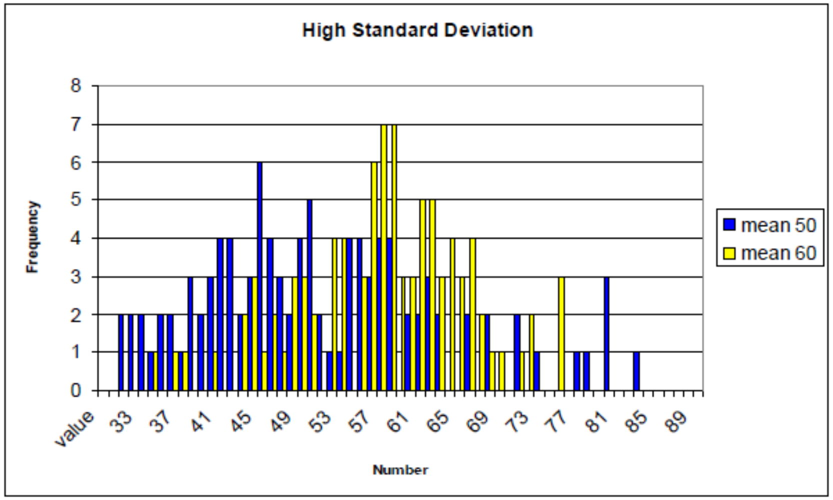
Working with quantitative data

- Plotting distributions
- Correlations
- Central tendencies within group, sub-groups
- Hypothesis testing
- Significance testing – are the measured differences due to chance or do they reflect a systematic pattern?
  - t-tests, f-tests

Key descriptive statistics (normal distribution)

- Central tendency (mean, median, mode)
- Variance
- Standard deviation





Figures source: Duflo, Esther, Rachel Glennerster, and Abhijit Banerjee, *RES.14-001 Abdul Latif Jameel Poverty Action Lab Executive Training: Evaluating Social Programs, Spring 2009*. (Massachusetts Institute of Technology: MIT OpenCourseWare), http://ocw.mit.edu (Accessed 20 Jun, 2010). License: Creative Commons BY-NC-SA

Hypothesis testing for difference of means and statistical significance

- H_0: X_1 = X2
- H_a: X1 ≠ X2
- If the $t_{stat} < t_{crit}$ for a given p value (significance level), then we cannot reject H_0 (i.e. the difference observed is likely due to chance)
  - Rule-of-thumb : For p=0.05, t_{stat}>~2 is typically considered statistically significant at a 95% confidence level

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

Key criteria when doing quantitative analysis

- How reliable is my data for this sample?
- How was my sample chosen from the population?
- What is the distribution of the variable of interest in the population?
- Can I distinguish the trend from noise?

Sample hypotheses to test using the data you have collected

1. Individuals with higher education receive and send more SMSs than those with lower levels of education
2. Ownership of consumer durables (fridge, TV, etc.) increases with household income
3. Households with lower household income have higher dependency ratios
4. Individuals with higher education have higher income
5. Those who attended private schools earn more individually than those who attended government schools
   - Is the difference statistically significant at a 95% confidence level

Are these patterns true for this sample? (internal validity)

- Possibly not

- Why?
  - Measurement error
  - Inconsistent administration of questions
  - Varying interpretations
  - Veracity of information (self-reporting, no triangulation, little trust in interviewer, setting of interview)

Can we say these patterns hold for poor households in Bangalore? (external validity)

- No
- Heavy selection bias
  - Sample of respondents with time, in visible public spaces, confident and willing to talk in one of 3 selected languages – is this representative of poor households in Bangalore?
- Would a stratified sample have resolved this problem?
  - We decide to meet X Kannada households, Y Tamil households, etc.

Could such an analysis be used to say ‘going to a private school must be encouraged because it leads to higher income’? (Causal Inference)

- Definitely Not

- All the problems from before +
- How do we know that it is the private school education that <u>caused</u> the higher earnings?
  - Are the two groups comparable on all factors except type of education?
  - There are other factors that lead both to private school education and higher income (e.g. parents’ wealth)
  - Quality of private schooling is highly variable; some worse, some better than govt schools

Key issues with data analysis

- Researchers almost always work with a sample
- Internal validity requires ensuring that what you are measuring is true for this sample
- External validity
  - If you want to use the sample to understand the population, the sample must be representative – a random sample is the best way to ensure this
- Causal inference
  - Making a claim that A causes B requires a valid counterfactual – this is very difficult to do using non-experimental research methods

References/ Resources

- Shadish, Cook and Campbell. (2001) “Experimental and Quasi-Experimental Designs for Generalized Causal Inference“, Wadsworth Publishing.

- “A Brief Course in Business Statistics.” William Mendenhall, Robert J. Beaver, Barbara M. Beaver. South-Western College Pub; 2nd edition, 2000.

Field Survey – Sample of occupations (85 respondents)

Auto Driver

Auto Rickshaw Driver

Auto Rickshaw Driver

Auto-Driver

Auto-rickshaw driver

Auto-rickshaw driver

Autowallah

Bhel-puri Vendor

Business - Toys for young kids

Cloth shop retailer

Cobbler

Cobbler

Commodity seller

Cook at a restaurant

Daily Labor

Driver

Dry Fruit Merchant

Fancy item seller

Flower Seller

Flower Seller

Flower Vendor

Flower seller

Floweriest

Footwear shop owner

Fruit Mechant

Fruit Seller

Fruit Seller

Fruit seller

Fuit Seller (all kinds)

House Maid

Housekeeping Services

Irons clothes

Jewlery Trinket Vedor

Juice Vendor

Juice shop employee

Laborer

Lassi Kart Vendor

Lathe Machine Turner

Magazine Reseller

Maid

Mechanic

Mechanic

Mobile Canteen Owner

Office Assistant

Office boy

Orange Fruit Juice shop owner

Owner of a pan shop

Owns Call Tax

Paan shop wallah

Paani Porri Vendor

Pan & Cigarette Vendor

Panipuri Seller

Peanut Vendor

Plastic vendor

Pressing Clothes/Part time office assistant

Railway Porter

Rikshaw driver

STREET VENDOR-CLOTHES

STREET VENDOR-SEASONAL

Security Guard

Security Guard

Security Guard

Security Guard

Seller

Selling bedsheets/pillow covers on platform

Selling fruits/veg on platform

Shoe seller

Small-scale real-estate agent

Spices Vendor

Stationary+Agriculture

Street vendor

Student/flower seller

Sugar Cane Juice

Sugarcane Juice Vendor

Sugarcane juice vendor, salesman and works in bakery store

Sweeper

Tailor

Taxi Driver

Tea stall in platform

Truck driver

Vegetable vendor

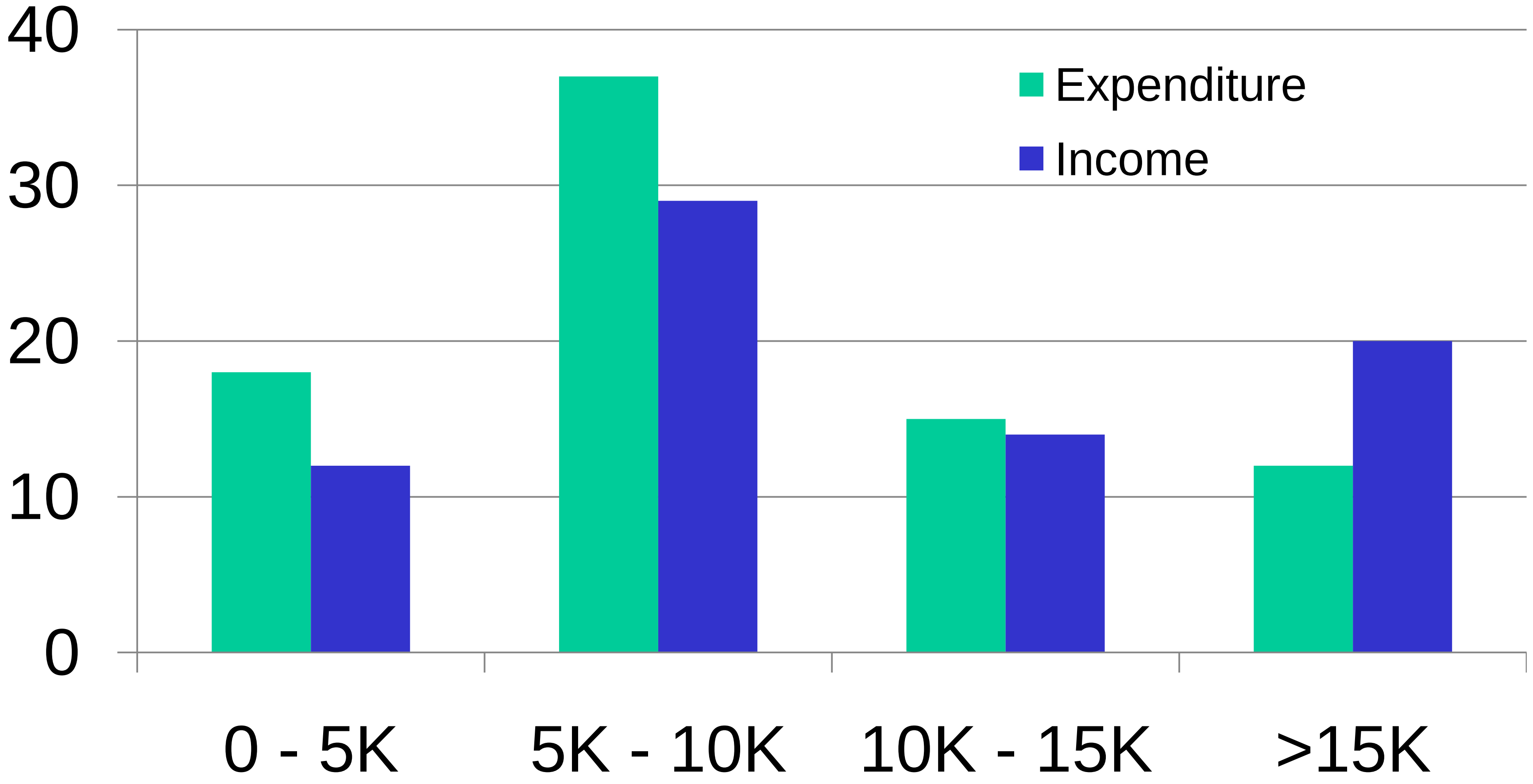
Waiter

Washerman

Works in a saloon

Compiled by Bill Thies

Field survey – HH Income and Expenditure



Compiled by Bill Thies