Application of big data analytics in social sciences
Lecture 2. Exploratory Data Analysis

Hao Luo, PhD

Department of Social Work and Social Administration
The University of Hong Kong

Sep 13, 2021

Disclaimer: contents are based on Agresti and Finlay (2009)

Descriptive Statistics

Types of Descriptive Methods

- Tabular Methods Frequency Distribution Table
- Graphical Methods
- Numerical Methods

Different methods answer different questions about data. Naturally, different questions have different answers. In general, we cannot look at data from all possible angles using only one method. So it's best to use more than one method when we're summarizing a data set, even if the different methods produce some overlap of information.

Variables and their measurement

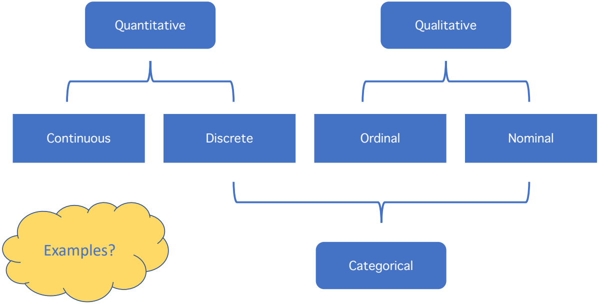
Definition

A variable is a characteristic that can vary in value among subjects in a sample or population.

Variables and their measurement

Definition

A variable is a characteristic that can vary in value among subjects in a sample or population.



Frequency Distribution Table

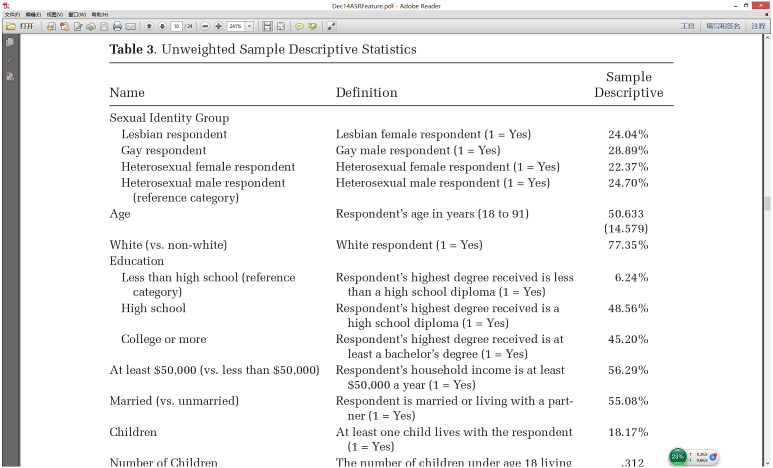

Dec14ASRFeature.pdf - Adobe Reader
文件(F) 编辑(E) 视图(V) 窗口(W) 帮助(H)
打开
10 / 24
241%
工具
填写和签名
注释
23%
0.2K/s
0.6K/s

Table 3. Unweighted Sample Descriptive Statistics

Name	Definition	Sample Descriptive
Sexual Identity Group		
Lesbian respondent	Lesbian female respondent (1 = Yes)	24.04%
Gay respondent	Gay male respondent (1 = Yes)	28.89%
Heterosexual female respondent	Heterosexual female respondent (1 = Yes)	22.37%
Heterosexual male respondent (reference category)	Heterosexual male respondent (1 = Yes)	24.70%
Age	Respondent's age in years (18 to 91)	50.633 (14.579)
White (vs. non-white)	White respondent (1 = Yes)	77.35%
Education		
Less than high school (reference category)	Respondent's highest degree received is less than a high school diploma (1 = Yes)	6.24%
High school	Respondent's highest degree received is a high school diploma (1 = Yes)	48.56%
College or more	Respondent's highest degree received is at least a bachelor's degree (1 = Yes)	45.20%
At least $50,000 (vs. less than $50,000)	Respondent's household income is at least $50,000 a year (1 = Yes)	56.29%
Married (vs. unmarried)	Respondent is married or living with a partner (1 = Yes)	55.08%
Children	At least one child lives with the respondent (1 = Yes)	18.17%
Number of Children	The number of children under age 18 living	.312

Frequency Distribution Table III

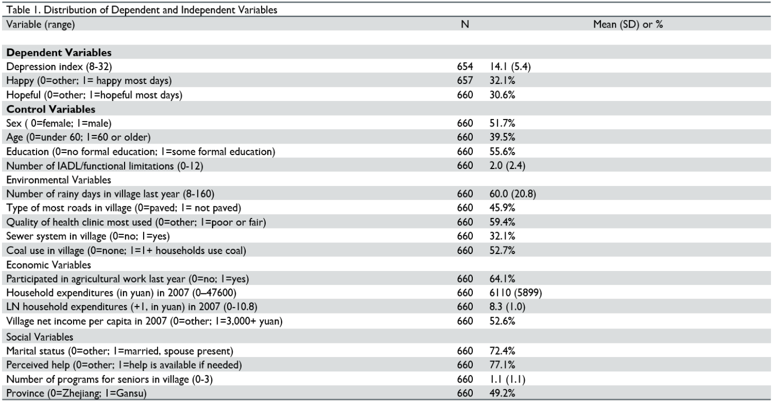
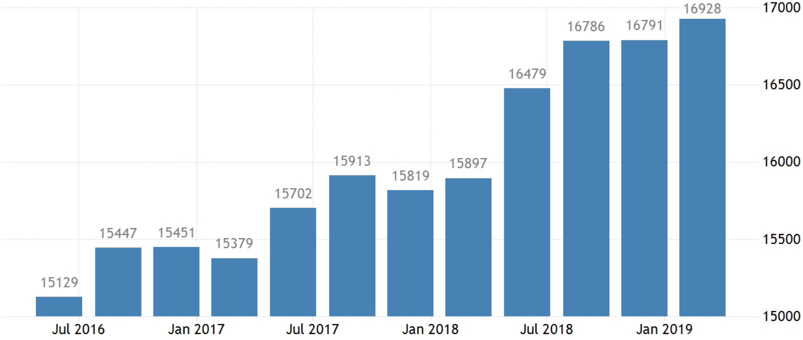


Table 1. Distribution of Dependent and Independent Variables

Variable (range)	N	Mean (SD) or %
Dependent Variables		
Depression index (8-32)	654	14.1 (5.4)
Happy (0=other; 1= happy most days)	657	32.1%
Hopeful (0=other; 1=hopeful most days)	660	30.6%
Control Variables		
Sex (0=female; 1=male)	660	51.7%
Age (0=under 60; 1=60 or older)	660	39.5%
Education (0=no formal education; 1=some formal education)	660	55.6%
Number of IADL/functional limitations (0-12)	660	2.0 (2.4)
Environmental Variables		
Number of rainy days in village last year (8-160)	660	60.0 (20.8)
Type of most roads in village (0=paved; 1= not paved)	660	45.9%
Quality of health clinic most used (0=other; 1=poor or fair)	660	59.4%
Sewer system in village (0=no; 1=yes)	660	32.1%
Coal use in village (0=none; 1=1+ households use coal)	660	52.7%
Economic Variables		
Participated in agricultural work last year (0=no; 1=yes)	660	64.1%
Household expenditures (in yuan) in 2007 (0–47600)	660	6110 (5899)
LN household expenditures (+1, in yuan) in 2007 (0-10.8)	660	8.3 (1.0)
Village net income per capita in 2007 (0=other; 1=3,000+ yuan)	660	52.6%
Social Variables		
Marital status (0=other; 1=married, spouse present)	660	72.4%
Perceived help (0=other; 1=help is available if needed)	660	77.1%
Number of programs for seniors in village (0-3)	660	1.1 (1.1)
Province (0=Zhejiang; 1=Gansu)	660	49.2%

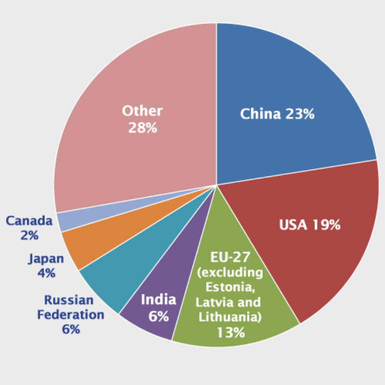
Yeatts, D. E., Pei, X., Cready, C. M., Shen, Y., Luo, H., & Tan, J. (2013). Village characteristics and health of rural Chinese older adults: Examining the CHARLS Pilot Study of a rich and poor province. *Social Science & Medicine, 98*, 71-78.

Qualitative variable: bar chart



SOURCE: TRADINGECONOMICS.COM | CENSUS AND STATISTICS DEPARTMENT, HONG KONG

Qualitative variable: pie chart



2008 Global CO2 Emissions from Fossil Fuel Combustion and some Industrial Processes (million metric tons of CO2). Credit: EPA; source: National CO2 Emissions from Fossil-Fuel Burning, Cement Manufacture, and Gas Flaring: 1751-2008.

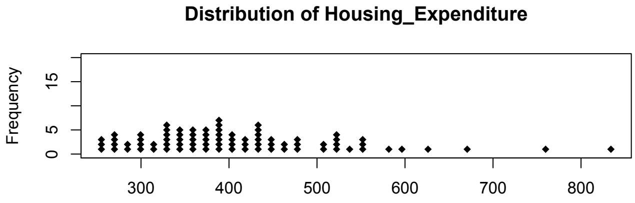
Examining Graphs

1. center
2. spread
3. shape
    - Symmetric distribution: one half is approximately a mirror image of the other half;
    - Left-skewed distribution: has a longer left tail than right tail;
    - Right-skewed distribution: has a longer right tail than left tail.

Patterns and Deviation from Patterns

- Clusters and gaps:
- Outliers: is an observation that is surprisingly different from the rest of the data.

Quantitative variable: dotplots



- One of the easiest plots to make;
- Most effective for smaller data sets. If the data set is too large, then the dotplot will be very cluttered.

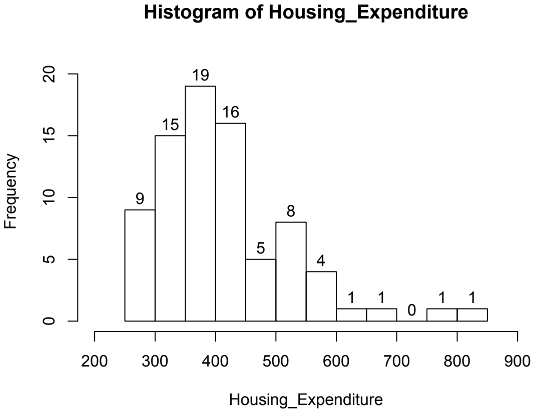
Quantitative variable: stemplots (Stem-and-leaf plots)

```
2 | 6667888
3 | 00011122333333455556666788888999999
4 | 0011122234444455567888
5 | 00233335568
6 | 037
7 | 6
8 | 3
```

- It shows every value;
- When turned on its side, it resembles a histogram;
- Inconvenient for very large data sets.

Histograms

A histogram can be drawn using frequencies, relative frequencies, or percentages.



- It looks like a stemplot on its side. But it is useful for displaying large data sets.
- It might fail to show the pattern of very small data set.
- The pattern of data within each group is lost due to grouping.

Measures of Central Tendency

Measures of central tendency determine the central point of a variable or the point around which all the measurements are scattered. Two main measures: *mean* and *median*.

¾ùÖµ Mean (average): a data set's center of gravity, the point at which the whole group of data balances.

- *Population mean* (μ):

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \tag{1}$$

- *Sample mean* ($\bar{X}$ or $\bar{Y}$):

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{2}$$

where $n =$ the number of measurements in the sample

It is affected by extreme of outlier measurements.

Measures of Central Tendency

Median (M): The median is the point that divides the measurements in half. That is, half of the values are at or below the median, and half are at or above the median.

- Suppose there are n measurements in a data set.
- Arrange the measurements in increasing order (i.e., from smallest to largest)
- Compute

$$I = \frac{n+1}{2} \tag{3}$$

- Then the median $M =$ the value of the Ith measurement.

It is not affected by outliers.

Measures of Variation

Measure of variation (or “measures of spread”) summarize the spread of a data set. They describe how measurements differ from each other and/or from their mean. The three most commonly used measures of variation are:

- Range
- Interquartile range
- Standard deviation

Standard deviation

- *Population standard deviation (σ):*

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}} \tag{4}$$

That is, we square the difference between each point and the mean, add those squares, divide by the number of points, and take the square root.

- *Sample standard deviation (s):*

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \tag{5}$$

- It takes every measurements into account
- It is affected by outliers

Standard deviation

- Standard deviation is measured in the same units as are data values, whereas variance is measured in squared units of the data values.
- It can be used as a unit for measuring the distance between any measurement and the mean of the data set.
- A standard deviation (or variance) of 0 indicates that all of the measurements are identical.
- It is the positive square root of variance. Variance will always be a positive number.
- A larger standard deviation indicates a larger spread among the measurements. The larger the standard deviation, the wider the graph.

Measures of Position

Quartiles, percentiles, and standardized scores (z-scores) are the most commonly used measures of position. These measures are used to describe the position of a value with respect to the rest of the values of variable.

Standardized scores, commonly known as z-scores, are independent of the units in which the data values are measured. Therefore, they are useful when comparing observations measured on different scales.

$$z - score = \frac{\text{measurement} - \text{mean}}{\text{standard deviation}}$$

It gives the distance between the measurement and the mean in terms of the number of standard deviations.

Example

The mean and the standard deviation of the daily high temperatures in degrees Fahrenheit for two cites are given below:

City	Mean	Standard deviation
South Bend	80	12
North Bend	84	4

Yesterday, both cities reported a high temperature of 95 degrees. Which city had the more unusually high temperature?

Box plot (box-and-whisker plot)

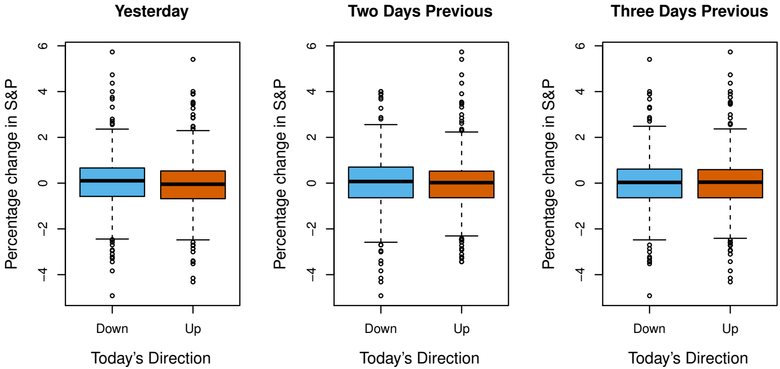


Figure: Left: Boxplots of the previous day¡¯s percentage change in the S&P index for the days for which the market increased or decreased, obtained from the Smarket data. Center and Right: Same as left panel, but the percentage changes for 2 and 3 days previous are shown.