

UNIVARIATE ANALYSIS OF NUMERICAL DATA

1. Univariate analysis explores each variable in a data set separately:

- It looks at the range of values
- The central tendency of the values
- It describes the pattern of response to the variable
- It describes each variable on its own

2. Univariate analysis

- Categorical variable (e.g. housing)
- Numerical variable (e.g. age)

3. Univariate analysis - Numeric

- Statistical analysis
 - Point estimation
 - Count, min, max, average, median, mode.
 - Dispersion
 - Range, standard deviation, variance, co-variance
 - Skewness, kurtosis
 - Missing value
 - Outliers
 - Binning
- Visualization
 - Histogram, box plot and etc...

Univariate Analysis – Numeric

Age					
Count	900	Average	35.25	St Dev	11.20
Min	19	Median	33	Variance	125.37
Max	75	Mode	27	Covariance	32%
Range	55	Skewness	1.09		
Missing	0	Kurtosis	0.88		

Univariate Analysis – Challenges

Variable	
Categorical	Numeric
Missing values	Missing values
Invalid values	Outliers
Numerization	Binning

4. Missing data

- Data entry error
- Data processing error
- Certain data may not be available at the time of entry
- How to handle missing data
 - Fill in the missing values manually
 - Ignore the records with missing data
 - Fill in it automatically
 - A global constant (e.g. “?”)
 - The variable mean

5. Outliers

- Data points inconsistent with the majority of data
- Different outliers
 - Valid: CEO's salary
 - Noisy: one's age = 200, widely deviated points
- Removal methods
 - Box plot
 - Clustering
 - Curve-fitting

6. Binning

- Binning is a process of transferring continuous variables into categorical counterparts
- Binning methods
 - Equal-width
 - Equal-frequency
 - Entropy-based methods
- Variable values (e.g. age)
 - 0, 4, 12, 16, 16, 18, 24, 26, 28
- Equal-width binning
 - Bin 1: 0, 4 [-, 10] bin
 - Bin 2: 12, 16, 16, 18 [10, 20] bin
 - Bin 3: 24, 26, 28 [20, +] bin
- Equal-frequency
 - Bin 1: 0, 4, 12 [-, 14] bin
 - Bin 2: 16, 16, 18 [14, 21] bin
 - Bin 3: 24, 26, 28 [21, +] bin

7. Numerization

- Numerization is the process of transferring categorical variable into numerical counterparts.
- Numerization methods
 - Binary method
 - Ordinal method
- Variable values (e.g. housing)
 - For free, own, rent
- Binary method
 - For free: 1, 0, 0
 - Own: 0, 1, 0
 - Rent: 0, 0, 1
- Ordinal method
 - Own: 5
 - For free: 3
 - Rent: 1

8. Quantification

- Introduction
 - To conduct quantitative analysis, responses to open-ended questions in survey research and the raw data collected using qualitative methods must be coded numerically.
 - Most responses to survey research questions already are recorded in numerical format
 - In mailed and face-to-face surveys, responses are keypunched into a data file.
 - In telephone and internet surveys, responses are automatically recorded in numerically format.
- Developing code categories

- Coding qualitative data can use an existing scheme or one developed by examining the data.
- Coding qualitative data into numerical categories sometimes can be a straightforward process
 - Coding occupation, for example, can rely upon numerical categories defined by the Bureau of the census.
- Coding most forms of qualitative data, however, requires much effort
- This coding typically requires using an iterative procedure of trial and error
- Consider, for example, coding responses to the question, “What is the biggest problem in attending college today?”
- The researcher must develop a set of codes that are;
 - Exhaustive of the full range of responses
 - Mutually exclusive (mostly) of one another.
- In coding responses to the question, “What is the biggest problem in attending college today?” the researcher might begin, for example, with a list of 5 categories, then realize that 8 would be better, then realize that it would be better to combine and use a total of 7 categories
- Each time the researcher makes a change in the coding scheme, it is necessary to restart the coding process to code all responses using the same scheme

9. Distribution

- Data analysis begins by examining distributions
- One might begin, for example, by examining the distribution of responses to a question about formal education, where responses are recorded within six categories

- A frequency distribution will show the number and percent of responses in each category of a variable

10. Central tendency

- A common measure of central tendency is the average or mean of the responses
- The median is the values in the middle case when all responses are rank-ordered
- The mode is the most common responses
- When data are highly skewed, meaning heavily balanced toward one end of the distribution, the median or mode might be better represent the most common or centered response.
- Consider this distribution of respondent ages:
 - 18, 19, 19, 19, 20, 20, 21, 22, 85
- The mean equals 27. But this number does not adequately represent the common respondent because the one person who is 85 skews the distribution toward the high end.
- The median equals 20
- This measure of central tendency gives a more accurate portrayal of the middle of the distribution

11. Dispersion

- Dispersion refers to the way the values are distributed around some central value, typically the mean.
- The range is the distance separating the lowest and highest values (e.g. the range of the ages listed previously equals 18-85)
- The standard deviation is an index of the amount of variability in a set of data
- The standard deviation represent dispersion with respect to the normal (bell shape) curve

- Assuming a set of numbers is normally distributed, then each standard deviation equals a certain distance from the mean.
- Each standard deviation (+1, +2, etc) is the same distance from each other on the bell-shaped curve, but represents a declining percentage of responses because of the shape of the curve.
- For example, the first standard deviation account 34.1% of the values below and above the mean
 - The figure 34.1% is derived from probability theory and the shape of the curve.
- Thus approximately 68% of all responses fall within one standard deviation of the mean
- The second standard deviation accounts for the next 13.6% of the responses from the mean (27.2% of all responses) and so on.
- Dispersion measures
 - Spread around the mean
 - Variance – too abstract, a step towards standard deviation
 - Standard deviation (from mean) – more intuitive
 - Standard deviation
 - Average distance between mean and each value in data set
 - Translates variance into same scale as mean and all the values
 - High values are generally bad
- If the responses are distributed approximately normal and the range of responses is low – meaning that most responses fall closely to the mean – then the standard deviation will be small
 - The standard deviation of professional golfer's score on a gold course will be low
 - The standard deviation of amateur golfer's scores on a golf course will be high

13. Continuous and Discrete Variables

- Continuous variables have responses that form a steady progression (e.g. age, income)
- Discrete (i.e. categorical) variables have responses that are considered to be separate from one another (i.e. sex, religious)
- Sometimes, it is matter of debate within the community of scholars about whether a measured variable is continuous or discrete
- This issue is important because the statistical procedures appropriate for continuous-level data, especially as related to the measurement of the dependent variable.
- Example: suppose one measures amount of formal education within five categories (less than hs, hs, 2 years vocational/college, college, post college)
- Is this measure continuous or discrete?
- In practice, five categories seem to be cut off point for considering a variable as continuous
- Using a seven-point response scale will give the researcher greater chance of deeming a variable to be continuous.

14. Subgroup comparison

- Collapsing response categories
 - Sometimes the researcher might want to analyse a variable by using fewer response categories than were used to measure it
 - In these instances, the researcher might want to collapse one or more categories into a single category
 - The researcher might want to collapse categories to simplify the presentation of the results or because few observations exist within some categories
- Collapsing response example

<u>Response</u>	<u>Frequency</u>
Strongly disagree	2
Disagree	22
Neither agree nor disagree	45
Agree	31
Strongly agree	1

One might want to collapse the extreme responses and work with just three categories

<u>Response</u>	<u>Frequency</u>
Disagree	24
Neither agree nor disagree	45
Agree	32

- Handling “Don’t Know”
 - When asking about knowledge of factual information (“Does you teenager drink alcohol?”) or opinion on a topic the subject might not know much about (“Do school officials do enough to discourage teenagers from drinking alcohol?”), it is wise to include a “don’t know” categories as a possible responses.
 - Analyzing “don’t know” responses, however can be a difficult task
 - The research-on-research literature regarding this issues is complex and without clear-cut guidelines for decision making
 - The decisions about whether to use “don’t know” response categories and how to code and analyse them tends to be idiosyncratic to the research and the researcher.