Types of Variables, Summary Statistics, and Graphical Displays

Data Science – Fall 2016

1

Class Survey Data

- Combined for two classes (Data Science & Business Stat)

2

Types of Variables

- Storage/Computing
  - How is it used in computation?
  - How much efficiency is there in storing the data?
- Analysis
  - How will the data be used for analysis?
    - Categorical, Ordinal, or Numeric
    - Text Analysis
- For the Class Data – what type of variable is each one?

3

Explanatory/Response Variables

- Explanatory (or independent) variables typically precede response variables chronologically. They are often thought of as causal agents (and often erroneously).
- The response (or dependent) variable(s) is/are the outcome that is measured. It is what is thought of as caused by the explanatory variable.
- In some data mining contexts (like cluster analysis), explanatory and response variables are not set or not predetermined.

4

Hans Rosling Video

- 200 countries, 200 years, 4 minutes
- How many variables were included?
- What types of variables are they?
- Which one(s) is/are the response variable(s)?

5

Summary Measures - Center

- Center (Location)
  - Mean – sensitive to outliers, typical measure
  - Median – resistant to outliers
  - Mode – not useful for numeric data
- Spread
  - Range – very sensitive to outliers
  - IQR – used with median
  - Standard Deviation – used with mean

6

Summary Measures – Dispersion/Spread

- Range = Max - Min
- IQR
  - Q_1 is the 25th percentile, meaning 25% are less than or equal to this value
  - Q_3 is the 75th percentile, meaning 75% are less than or equal to this value
  - IQR = Q_3-Q_1
  - The IQR is resistant to outliers (like the median)
- Standard deviation

7

Variance

- Average (or unbiased average) of squared deviations of values from the mean (in squared units)

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

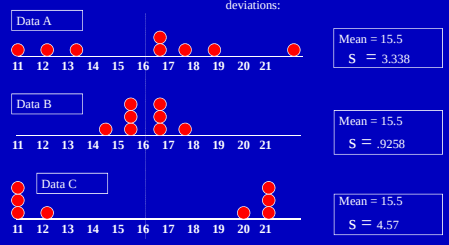
Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Comparing Standard Deviations

Same mean, but different standard deviations:



Summary Measures – Categorical/Ordinal Data

- To summarize categorical data, use frequency and relative frequency (%)

Gender	Frequency	Relative Frequency (%)
M	14	56%
F	10	40%
Other	1	4%

Freq/Total: 10/25 = 40% or 0.40

- To summarize ordinal data, you can ALSO use cumulative frequency and relative frequency

Likelihood of Voting	Frequency	Relative Frequency (%)	Cum Freq	Cum Rel Freq
Definitely Not	2	4%	2	4%
Likely Not	12	24%	14	28%
Likely	16	32%	30	60%
Definitely	20	40%	50	100%

All lower 2+12

All lower 14+16

Summary Measures – Categorical/Ordinal Data Joint Frequency Distributions

- Used to examine data that is characterized by two variables that are both categorical (or ordinal). Numeric variables may be collapsed into ordinal categories and then used in a joint frequency distribution.
- The number of rows and columns corresponds to the number of categories for each variable

Joint Frequency Example

- A company offers 3 extended warranty plans for purchase
- The plans are sold both in the store and online

Extended Warranty Sales			
	1 yr	5 yrs	10 yrs
In Store	69	60	35
Online	89	27	11

Marginal Frequencies

- Marginal Frequencies are the row and column totals

Extended Warranty Sales				
	1 yr	5 yrs	10 yrs	TOTAL
In Store	69	60	35	164
Online	89	27	11	127
TOTAL	158	87	46	291

Marginal (Column) Frequencies

Joint Probability Distribution

- The joint probability distribution is the probability of being in each cell
- It is useful when talking about % of total (sales)

Extended Warranty Sales				
	1 yr	5 yrs	10 yrs	TOTAL
In Store	0.24	0.21	0.12	0.56
Online	0.31	0.09	0.04	0.44
TOTAL	0.54	0.30	0.16	1.00

Cell Value / Total

69 / 291 = 0.24

Conditional Probabilities

- The conditional probability is the number in that cell divided by the row or column total
- It is useful in comparing distributions across different populations

Extended Warranty Sales				
	1 yr	5 yrs	10 yrs	TOTAL
In Store	0.42	0.37	0.21	1.00
Online	0.70	0.21	0.09	1.00
TOTAL	0.54	0.30	0.16	1.00

Cell Value / Row (or Col) Total

69 / 164 = 0.42

Online sales had shorter EW Sales

Types of Graphs – Univariate (One Variable)

- **Categorical Data**
  - **Bar Chart**
  - **Pie Chart**
- **Quantitative Data**
  - **Stem and Leaf**
  - **Histogram**
  - **Dotplot**

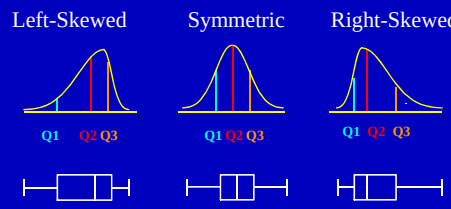
Types of Graphs – Bivariate (Two Variables)

- **Categorical Data**
  - **Multi-columnBar Chart**
  - **Multiple Pie Charts**
- **Quantitative Data**
  - **Back-to-back Stem and Leaf**
  - **Scatterplot or Line Chart**
  - **Histograms**

Describing Distributions

- For numeric variables, we summarize the distribution by…(SOCS+IT)
  - Shape (# of modes, skew)
  - Outliers
  - Center (Mean or median)
  - Spread (range for one group, dispersion/spread if comparing groups)
  - Anything Interesting
  - Take-home message

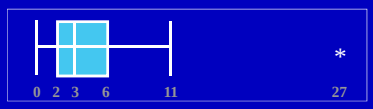
Distribution Shape and Box and Whisker Plot



Box-and-Whisker Plot Example

- A boxplot is a graphical display of the five number summary (min, Q1, median, Q3, max)
- Below is a Box-and-Whisker plot for the following data:

Min	Q_1		Q_2		Q_3		Max
0	2	2	2	3	3	4	5	6	11	27



- This data are right skewed, as the plot depicts

Graphical Summaries

- See Handout
- For each graph identify…
  - The variable(s) being measured
  - The type of variable(s)
  - The type of graph being displayed
  - Summarize what you see on each graph (in one or two sentences)

Summarizing Survey Data

- Let's look together at summarizing student survey data. What variables might we want to compare

23

For Next Class

- Bias & Multivariate Thinking
- See Learning Outcome sheet online
- Do pre-assessment
- Extra Office Hours – tonight, 5-7pm, SAC310

24

Population Estimation

- Separate into two groups…

- What is your best guess for the population of the Philippines?
- Record and review these
- This is an example of bias!