BAN 602: Quantitative Fundamentals

Lecture 1

Agenda

1. Introduction - DIKW
2. Data Types
3. Big Picture of Chapter 2 (knowledge/concept mapping)
    - Descriptive Statistics 1 - Tabular and Graphic
4. Simpson’s Paradox
5. Big Picture of Chapter 3
    - Descriptive Statistics 2 - Numerical Measures
6. Example of numerical description of data

Data

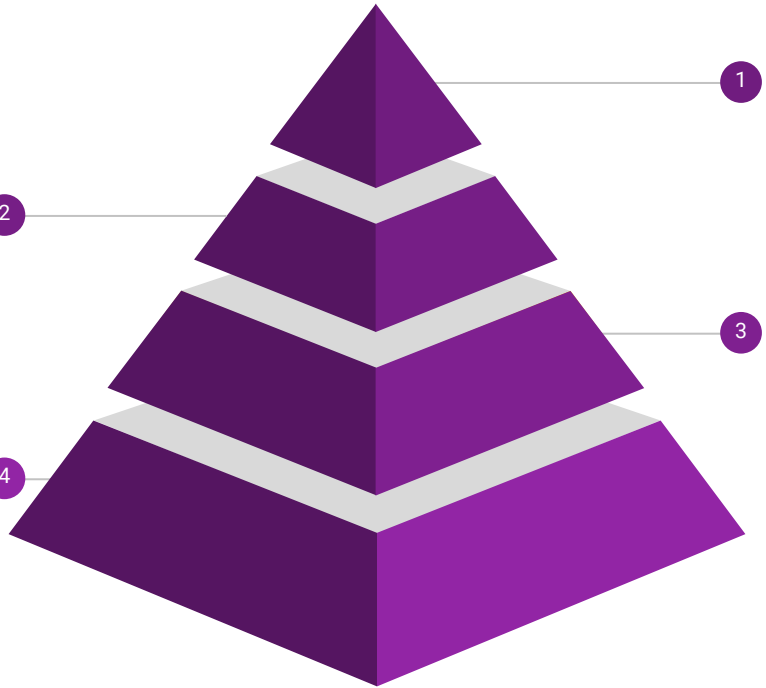
What is data?

What is information?

What is knowledge?

Why Do We Need DIK?

DIKW Pyramid



Wisdom
If-Then: Inference and Prediction

Information
What, When, Where, Who

Knowledge
How, Why

Data
Basic Fact

Data Types

Data (Cross-sectional vs. time-series)			
Categorical		Quantitative	
Nominal (Count Data) ● Example (binary): Gender	Ordinal ● Example: unlikely – likely on a scale of 1-5	Discrete: ● Example: Resulting numbers from rolling a die	Continuous ● Example: weight, length, time, $.

What is the difference between ordinal (categorical) data and discrete (quantitative) data?

Big Picture of Chapter 2

Descriptive Statistics: Tabular & Graphic Description of Data

	Categorical	Quantitative
One Variable	Absolute and relative frequencies and frequency • Table • Graph: bar chart, pie chart	Absolute and relative frequencies and frequency • Table (binning) • Graph: dot plot, histogram, stem-and-leaf plot
Two Variables	• **Cross-tabulation** • Graph: side-by-side and stacked bar chart	• **Cross-tabulation** (binning) • Graph: scatter

Example of Cross-Tabulation: Frequency Table of Two Variables

	Promoted	Not Promoted	Total
M	60	60	120
F	30	50	80
Total	90	110	200

We must be extremely careful when we study the relationship between two variables with a cross-tabulation of aggregated data. (Simpson's Paradox)

Simpson's Paradox

	Promoted	Not Promoted	Total
M	60	60	120
F	30	50	80
Total	90	110	200

Can we draw the conclusion (of gender discrimination) that male officers are more likely to be promoted than female officers? Why or why not?

Relative Frequency Table of Two Variables

(aka: **Joint Probability Table**)

	Promoted	Not Promoted	Total
M	0.3 (joint prob)	0.3 (joint prob)	0.6 (marginal)
F	0.15 (joint prob)	0.25 (joint prob)	0.4 (marginal)
Total	0.45 (marginal)	0.55 (marginal)	1

A. Randomly select an officer from HPD. 60% chance that this is a male officer.
B. 60% of officers @ HPD are male.
C. There is 30% chance a male officer is promoted.
D. Of all officers at HPD, 30% are male and promoted.
E. Randomly select an officer. 30% chance a promoted male officer is selected.

Relative Frequency Table of Two Variables
(aka: **Joint Probability Table**)

	Promoted	Not Promoted	Total
M	0.3 (joint prob)	0.3 (joint prob)	0.6 (marginal)
F	0.15 (joint prob)	0.25 (joint prob)	0.4 (marginal)
Total	0.45 (marginal)	0.55 (marginal)	1

Question 1 of today's exercise: use your own language to briefly explain the meaning of the probabilities highlighted.

Hidden Variable - Frequency

Male	Promoted	Not Promoted	
age<=40	50	30	80
age>40	10	30	40
	60	60	120

Female	Promoted	Not Promoted	
age<=40	12	6	18
age>40	18	44	62
	30	50	80

Hidden Variable - Relative Frequency (Conditional Probability Table)

Male	Promoted	Not Promoted	
age<=40	0.625	0.375	1
age>40	0.25	0.75	1

Female	Promoted	Not Promoted	
age<=40	0.67	0.33	1
age>40	0.29	0.71	1

Question 2 of today's exercise: use your own language to briefly explain the meaning of the probabilities highlighted.

Hidden Variable - Relative Frequency (Conditional Probability Table)

Male	Promoted	Not Promoted	
age<=40	0.625	0.375	1
age>40	0.25	0.75	1

Female	Promoted	Not Promoted	
age<=40	0.67	0.33	1
age>40	0.29	0.71	1

What conclusion(s) can we draw?

Simpson’s Paradox

- Conclusions drawn from two or more separate crosstabulations that can be reversed when the data are aggregated into a single cross-tabulation.
- The relationship between two variables may be explained by a third hidden variable.
- **Question 3 of today’s exercise: use your own language to briefly explain Simpson’s Paradox.**

Big Picture of Chapter 3

Descriptive Statistics: Numeric Description of Data

Measures of Location	Measures of Variability	Measures of Distribution Shape	Measures of Association

• Mean (Weighted; Geometric)
• Median
• Mode
• Relative location:
  ○ Percentile and Quartiles
  ○ **Z-score** (Chebyshev's; Empirical rule)
	• Range
	• MAD
	• Interquartile range
	• Variance and **standard deviation**
	• Coefficient of variation
		• Distribution shape: skewness	• Covariance
			• Correlation coefficient
			• Trend line and its slope

Some Tools and Applications

- Numeric Measure Summary Technique/Tool: Box Plots – min, Q1, Q2, Q3, max and mean
  - Summary of one variable
  - Summary and comparison of multiple variables: side-by-side box plots
- Detecting Outliers (Application of Relative Location)
  - 3 sigma or more away from the mean
  - 1.5 IQR below Q1 or above Q3

Example of Numeric Description of Data
(Prepare for Next Week's Quiz!)

	X	Y	Zx	Zy
	1	6	-1.4142	1.2649
	2	5	0	0.6325
	2	4	0	0
	2	3	0	-0.6325
	3	2	1.4142	-1.2649
mean	2	4	0	0
var	0.5	2.5	1	1
sd	0.7071	1.5811	1	1

	X &d Y	Zx & Zy
cov	-1	-0.8944
r	-0.8944	-0.8944
b1	-2	-0.8944
b0	8	0

1. Any outlier?
2. What does r tell us?
3. What do b1's tell us?
4. **Observations that can be generalized? (Q4 of Exercise)**