GPCO 453: Quantitative Methods I

Sec 03: Exploratory Data Analysis

Shane Xinyang Xuan[1]
ShaneXuan.com

October 23, 2017

[1]Department of Political Science, UC San Diego, 9500 Gilman Drive #0521.

Contact Information

Shane Xinyang Xuan

xxuan@ucsd.edu

The teaching staff is a team!

Professor Garg	Tu	1300-1500 (RBC 1303)
Shane Xuan	M	1100-1200 (SSB 332)
	M	1530-1630 (SSB 332)
Joanna Valle-luna	Tu	1700-1800 (RBC 3131)
	Th	1300-1400 (RBC 3131)
Daniel Rust	F	1100-1230 (RBC 3213)

Roadmap

In this section, we cover the basics for exploratory data analysis:

- Data structure

Roadmap

In this section, we cover the basics for exploratory data analysis:

- Data structure
- Unit of analysis

Roadmap

In this section, we cover the basics for exploratory data analysis:

- Data structure
- Unit of analysis
- Variable type

Roadmap

In this section, we cover the basics for exploratory data analysis:

- Data structure
- Unit of analysis
- Variable type
- Dispersion

Roadmap

In this section, we cover the basics for exploratory data analysis:

- Data structure
- Unit of analysis
- Variable type
- Dispersion
- Cross tabulation

Roadmap

In this section, we cover the basics for exploratory data analysis:

- Data structure
- Unit of analysis
- Variable type
- Dispersion
- Cross tabulation
- Primer on marginal probability and conditional probability

Roadmap

In this section, we cover the basics for exploratory data analysis:

- Data structure
- Unit of analysis
- Variable type
- Dispersion
- Cross tabulation
- Primer on marginal probability and conditional probability
- Geometric mean

Roadmap

In this section, we cover the basics for exploratory data analysis:

- Data structure
- Unit of analysis
- Variable type
- Dispersion
- Cross tabulation
- Primer on marginal probability and conditional probability
- Geometric mean
- Variance and standard deviation

Roadmap

In this section, we cover the basics for exploratory data analysis:

- Data structure
- Unit of analysis
- Variable type
- Dispersion
- Cross tabulation
- Primer on marginal probability and conditional probability
- Geometric mean
- Variance and standard deviation
- Percentiles

Data Structure

- Time-series data track the same sample at different points in time
  - Marry-2002
  - Marry-2003

    $\vdots$
  - Marry-2008

Data Structure

- Time-series data track the same sample at different points in time
  - Marry-2002
  - Marry-2003

    $\vdots$
  - Marry-2008
- Cross sectional data observe different subjects at the same point of time
  - Marry-2002
  - Jake-2002

    $\vdots$
  - Dan-2002

Variable Types

- Nominal (categorical)

  i.e. Hillary, Donald, Gary, Jill

- Ordinal (can rank)

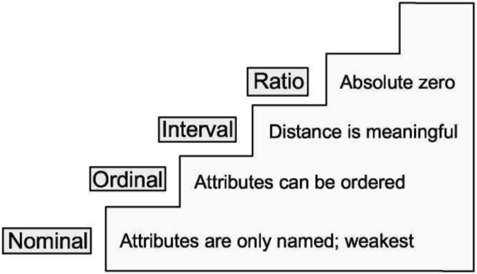
  i.e. strongly agree > agree > neutral > disagree > strongly disagree

- Interval (different by how much?)

  i.e. grade in school, happiness index, election fraud index

Variable Types

Figure: Hierarchy of measurement levels (Trochim & Donnelly 2006)



Variable Types: Examples

Table: Variable Types

Variable	**Type**
Celsius	Interval
Kelvin	Ratio
GDP	Ratio
Country	Nominal
Gender	Nominal
Age	Ratio
Distance	Ratio
Happiness index	Interval

The Unit of Analysis

- Unit of Analysis is the “case” of the data set

The Unit of Analysis

- Unit of Analysis is the “case” of the data set
  - a collection of information about schools

The Unit of Analysis

- Unit of Analysis is the "case" of the data set
  - a collection of information about schools
  - a collection of information about classes

The Unit of Analysis

- Unit of Analysis is the "case" of the data set
  - a collection of information about schools
  - a collection of information about classes
  - a collection of information about people

The Unit of Analysis

- Unit of Analysis is the “case” of the data set
  - a collection of information about schools
  - a collection of information about classes
  - a collection of information about people
  - a collection of information about countries

The Unit of Analysis

- Unit of Analysis is the “case” of the data set
  - a collection of information about schools
  - a collection of information about classes
  - a collection of information about people
  - a collection of information about countries
  - a collection of information about states

The Unit of Analysis

- Unit of Analysis is the "case" of the data set
  - a collection of information about schools
  - a collection of information about classes
  - a collection of information about people
  - a collection of information about countries
  - a collection of information about states
- One way to think: What is my unit of analysis $\rightarrow$ what items do I want to compare?

Dispersion

Positive Skew: Mean $>$ Median

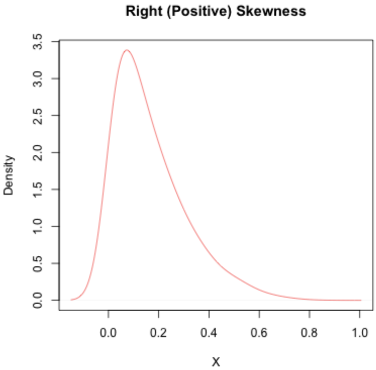
Dispersion

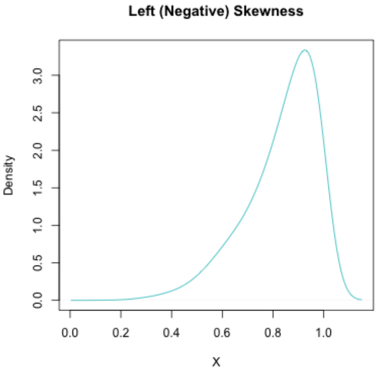
Positive Skew: Mean $>$ Median

Negative Skew: Mean $<$ Median

Positive Skew: Mean $>$ Median

Negative Skew: Mean $<$ Median





Conditional Probability

- Students taking the GMAT were asked about their undergraduate major and intent to pursue MBA as a full time or part time student:

	Business	Engineering	Other	Total
Full time	352	197	251	800
Part time	150	161	194	505
Total	502	358	445	1305

Conditional Probability

- Students taking the GMAT were asked about their undergraduate major and intent to pursue MBA as a full time or part time student:

	Business	Engineering	Other	Total
Full time	352	197	251	800
Part time	150	161	194	505
Total	502	358	445	1305

- Develop a joint probability table

Conditional Probability

- Students taking the GMAT were asked about their undergraduate major and intent to pursue MBA as a full time or part time student:

	Business	Engineering	Other	Total
Full time	352	197	251	800
Part time	150	161	194	505
Total	502	358	445	1305

- Develop a joint probability table

	Business	Engineering	Other	Total
Full time	.269	.151	.192	.613
Part time	.115	.124	.148	.387
Total	.385	.274	.341	1

Conditional Probability

	Business	Engineering	Other	Total
Full time	.269	.151	.192	.613
Part time	.115	.124	.148	.387
Total	.385	.274	.341	1

Conditional Probability

	Business	Engineering	Other	Total
Full time	.269	.151	.192	.613
Part time	.115	.124	.148	.387
Total	.385	.274	.341	1

- If a student intends to attend classes full time, what is the probability that he was an undergraduate engineering major?

Conditional Probability

	Business	Engineering	Other	Total
Full time	.269	.151	.192	.613
Part time	.115	.124	.148	.387
Total	.385	.274	.341	1

- If a student intends to attend classes full time, what is the probability that he was an undergraduate engineering major?
  $\frac{197}{800} \approx .2463$

Conditional Probability

	Business	Engineering	Other	Total
Full time	.269	.151	.192	.613
Part time	.115	.124	.148	.387
Total	.385	.274	.341	1

- If a student intends to attend classes full time, what is the probability that he was an undergraduate engineering major?
  $\frac{197}{800} \approx .2463$
- If a student was an undergraduate business business major, what is the probability that he intends to be full time?

Conditional Probability

	Business	Engineering	Other	Total
Full time	.269	.151	.192	.613
Part time	.115	.124	.148	.387
Total	.385	.274	.341	1

- If a student intends to attend classes full time, what is the probability that he was an undergraduate engineering major?
  $\frac{197}{800} \approx .2463$
- If a student was an undergraduate business business major, what is the probability that he intends to be full time?
  $\frac{352}{502} \approx .7012$

Conditional Probability

	Business	Engineering	Other	Total
Full time	.269	.151	.192	.613
Part time	.115	.124	.148	.387
Total	.385	.274	.341	1

- If a student intends to attend classes full time, what is the probability that he was an undergraduate engineering major?
  $\frac{197}{800} \approx .2463$
- If a student was an undergraduate business business major, what is the probability that he intends to be full time?
  $\frac{352}{502} \approx .7012$
- Let F denote the event that the student intends to be full time, and B be the event that the student was a business major. Are F and B independent?

Conditional Probability

	Business	Engineering	Other	Total
Full time	.269	.151	.192	.613
Part time	.115	.124	.148	.387
Total	.385	.274	.341	1

- If a student intends to attend classes full time, what is the probability that he was an undergraduate engineering major?
  $\frac{197}{800} \approx .2463$
- If a student was an undergraduate business business major, what is the probability that he intends to be full time?
  $\frac{352}{502} \approx .7012$
- Let F denote the event that the student intends to be full time, and B be the event that the student was a business major. Are F and B independent?

  Since $\Pr(F|B) \neq \Pr(F)$, we know F and B are not independent.

Geometric Mean

- The geometric mean is a type of average, and it is commonly used for growth rates (i.e. population growth, or interest rates)

$$\left(\prod_{i}^{n} x_i\right)^{1/n} = \sqrt[n]{x_1 x_2 \cdots x_n} \tag{1}$$

Geometric Mean

- The geometric mean is a type of average, and it is commonly used for growth rates (i.e. population growth, or interest rates)

$$\left(\prod_{i}^{n} x_i\right)^{1/n} = \sqrt[n]{x_1 x_2 \cdots x_n} \tag{1}$$

- You have a stock ($PV = 90000$) that increases by 50% the first year after you bought it, 20% the second year, and 90% the third year. How much is the stock worth after Year 3?

Geometric Mean

- The geometric mean is a type of average, and it is commonly used for growth rates (i.e. population growth, or interest rates)

$$\left(\prod_{i}^{n} x_i\right)^{1/n} = \sqrt[n]{x_1 x_2 \cdots x_n} \tag{1}$$

- You have a stock ($PV = 90000$) that increases by 50% the first year after you bought it, 20% the second year, and 90% the third year. How much is the stock worth after Year 3?
- One way to calculate is $(90000)(1.5)(1.2)(1.9)$

Geometric Mean

- The geometric mean is a type of average, and it is commonly used for growth rates (i.e. population growth, or interest rates)

$$\left(\prod_{i}^{n} x_i\right)^{1/n} = \sqrt[n]{x_1 x_2 \cdots x_n} \tag{1}$$

- You have a stock ($PV = 90000$) that increases by 50% the first year after you bought it, 20% the second year, and 90% the third year. How much is the stock worth after Year 3?
- One way to calculate is $(90000)(1.5)(1.2)(1.9)$
- Another way to calculate is to use the geometric mean:

$$(90000)\overbrace{\left(\underbrace{\left[\sqrt[3]{(1.5)(1.2)(1.9)}\right]}_{\text{geometric mean}}\right)^{3}}^{\text{compounding by 3 years}} \tag{2}$$

Variance and Standard Deviation

- Variance for a sample is defined as

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}$$

Standard deviation is defined as

$$\begin{aligned} \sigma &\equiv \sqrt{\sigma^2} \\ &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}} \end{aligned}$$

Variance and Standard Deviation

- Example

x_i	$x_i - \overline{X}$	$(x_i - \overline{X})^2$
1		
2		
3		
4		
5		

Find the mean

$$\overline{X} = \frac{1+2+3+4+5}{5} = 3$$

Variance and Standard Deviation

- Example

x_i	$x_i - \overline{X}$	$(x_i - \overline{X})^2$
1	-2	
2	-1	
3	0	
4	1	
5	2	

Calculate the 2^{nd} column

$$x_1 - \overline{X} = 1 - 3 = -2$$
$$x_2 - \overline{X} = 2 - 3 = -1$$
$$\vdots$$
$$x_5 - \overline{X} = 5 - 3 = 2$$

Variance and Standard Deviation

- Example

x_i	$x_i - \overline{X}$	$(x_i - \overline{X})^2$
1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4

Square the 2^{nd} column

$$(x_1 - \overline{X})^2 = (-2)^2 = 4$$
$$(x_2 - \overline{X})^2 = (-1)^2 = 1$$
$$\vdots$$
$$(x_5 - \overline{X})^2 = 2^2 = 4$$

Variance and Standard Deviation

- Example

x_i	$x_i - \overline{X}$	$(x_i - \overline{X})^2$
1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4

Let me remind you of the formula

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n-1}$$

$$= \frac{4+1+0+1+4}{5-1}$$

$$= 2.5$$

$$\sigma = \sqrt{2.5}$$

Percentiles

- Location of the p-th percentile is

$$L_p = \frac{p}{100}(n+1) \tag{3}$$

Percentiles

- Location of the p-th percentile is

$$L_p = \frac{p}{100}(n+1) \tag{3}$$

- We arrange the following numbers in ascending order:

	3710	3755	3850	3880	3880	3890	3920	3940	3950	4050	4130	4325
Position	1	2	3	4	5	6	7	8	9	10	11	12

Percentiles

- Location of the p-th percentile is

$$L_p = \frac{p}{100}(n+1) \tag{3}$$

- We arrange the following numbers in ascending order:

	3710	3755	3850	3880	3880	3890	3920	3940	3950	4050	4130	4325
Position	1	2	3	4	5	6	7	8	9	10	11	12

- The location of the 80th percentile is

$$L_{80} = \left(\frac{80}{100}\right)(12+1) = 10.4 \tag{4}$$

Percentiles

- Location of the p-th percentile is

$$L_p = \frac{p}{100}(n+1) \tag{3}$$

- We arrange the following numbers in ascending order:

	3710	3755	3850	3880	3880	3890	3920	3940	3950	4050	4130	4325
Position	1	2	3	4	5	6	7	8	9	10	11	12

- The location of the 80th percentile is

$$L_{80} = \left(\frac{80}{100}\right)(12+1) = 10.4 \tag{4}$$

- The 80th percentile is the value in position 10 (4050) plus 0.4 times the difference between the value in position 11 (4130) and the value in position 10 (4050):

$$4050 + 0.4(4130 - 4050) = 4082 \tag{5}$$