E.D.A.

Greg C Elvers, Ph.D.

1

Exploratory Data Analysis

- One of the most important steps in analyzing data is to look at the raw data
- This allows you to:
  - find observations that may be incorrect
  - quickly tell if the data are "reasonable" (i.e., if they conform to expectations)
  - see trends in the data
- The process of looking at the data is often called *exploratory data analysis (E.D.A.)*

2

E.D.A.

- Usually, the data set is so large that just looking at the data is meaningless
- The data need to be organized and summarized before you can interpret them
- Exploratory data analysis does just that

3

Steps for E.D.A.

- The first step in most exploratory data analysis procedures is to organize the data by sorting it
- The sorted data is then presented graphically in one (or more) of several manners:
  - Stem and leaf plots
  - Frequency distributions
  - Tukey box plots

4

Stem and Leaf Plots

- Each quantitative observation is broken into two parts: the *stem* and the *leaf*
- The stem are all the digits of the number except for the least significant digit
- The leaf is the least significant digit

Obs.	Stem	Leaf
14	1	4
132	13	2
41.2*	41	2
	4	1
1234*	123	4
	12	3

*Depending on the range of numbers in the distribution, either stem and leaf could be used

Stem and Leaf Plots

- For each observation, determine its stem and its leaf
- Sort the stems, removing any duplicates
- List the leaves, one by one, to the right of its stem

59 57 75 90 100 95 74 84 84 91
73 88 78 69 64 74 53 86 64 72

Stem | Leaf
5 | 379
6 | 449
7 | 234458
8 | 4468
9 | 015
10 | 0

Create a Stem and Leaf Plot

- Create a stem and leaf plot from the following IQs:

82	80	97	111	121
116	96	105	105	112
95	109	100	92	86
96	76	108	87	94
104	88	120	91	85

Frequency Distributions

- A *frequency distribution* is a table that lists how often each number (or range of numbers) in the data occurs

82	80	97	111	121
116	96	105	105	112
95	109	100	92	86
96	76	108	87	94
104	88	120	91	85

Class	Frequency
70-79	1
80-89	6
90-99	7
100-109	6
110-119	3
120-129	2

Frequency Distributions

- The *class* is a range of numbers that represent a category
  - All members of the category have the same characteristics
- Frequency distributions allow you to quickly look at a large set of data to determine the general characteristics of the data

9

Cumulative Frequency Distributions

- The *cumulative frequency distribution* is derived from the frequency distribution by listing the number of scores that are less than or equal to the class.
- The cumulative frequency distribution is useful for calculating the *percentile rank*

Class	Freq.	C. Freq.
70-79	1	1
80-89	6	7
90-99	7	14
100-109	6	20
110-119	3	23
120-129	2	25

10

Percentile Rank

- The *percentile rank* is the percentage of observations that are at or below a given score
  - In the previous example, what percent of scores are less than or equal to your IQ (116)?
- To calculate the percentile rank, first create the cumulative frequency distribution
- Then, apply the formula given on the next slide

11

Percentile Rank

- $$PR = \frac{\text{cum } f_{ll} + \left[\frac{(X_i - X_{ll})}{w}\right] f_i}{N} X\, 100$$
- cum f_{ll} = cumulative frequency of the class below X
- X_i = score to be converted to percentile rank
- X_{ll} = score at the lower real limit of the class containing X
- w = width of the class
- f_i = number of cases within the class containing x
- N = number of scores in the distribution

12

Cumulative Frequency$_{ll}$, X_i

- E.g., the cumulative frequency of the class below 116 is 20
- cum f_{ll} = 20
- The score to be converted, X_i is 116 in this example

Class	Freq.	C. Freq.
70-79	1	1
80-89	6	7
90-99	7	14
100-109	6	20
110-119	3	23
120-129	2	25

Lower Real Limit

- Because the classes are continuous, we need to find the true limit of the class
- The unit of measure is one, so the lower real limit of the class containing X_i is: 110 - (1 / 2) = 109.5

Class	Freq.	C. Freq.
70-79	1	1
80-89	6	7
90-99	7	14
100-109	6	20
110-119	3	23
120-129	2	25

Width, Frequency, and N

- The width of the class is 10 (the difference of the true limits, e.g. 79.5 - 69.5 = 10)
- The number of observations within the class containing X_i is 3 = f_i
- N, the number of scores is 25

Class	Freq.	C. Freq.
70-79	1	1
80-89	6	7
90-99	7	14
100-109	6	20
110-119	3	23
120-129	2	25

Calculating the Percentile Rank

$$PR = \frac{\text{cum } f_{ll} + \left[\frac{(X_i - X_{ll})}{w}\right] X\ f_i}{N} X\ 100 = \frac{20 + \left[\frac{(116 - 109.5)}{10}\right] X\ 3}{25} X\ 100 = 87.8$$

- cum f_{ll} = 20
- X_i = 116
- X_{ll} = 109.5
- f_i = 3
- w = 10
- N = 25

Score Corresponding to a Percentile Rank (PR)

- Create the cumulative frequency distribution
- Use the following formula where
  - cum f_{PR} = cumulative frequency (percentile rank X number of observations / 100)
  - cum f_{ll} = cumulative frequency of the class below the class cum f_{PR} containing PR
  - X_{ll} = score at lower real limit of class containing PR
  - w = width of class
  - f_i = number of cases within the class containing PR

$$X_{PR} = X_{ll} + \frac{w(\text{cum } f_{PR} - \text{cum } f_{ll})}{f_i}$$

17

What Score Corresponds to a Percentile Rank of 87.8?

- cum f_{PR} = the percentile rank times the number of scores divided by 100
- 87.8 X 25 / 100 = 21.95

Class	Freq.	C. Freq.
70-79	1	1
80-89	6	7
90-99	7	14
100-109	6	20
110-119	3	23
120-129	2	25

18

Cumulative Frequency_{ll}

- Convert the cumulative frequencies to percentages (divide each by the number of observations, e.g. 25)

Class	Freq.	C. Freq.	% C. Freq
70-79	1	1	0 - 4
80-89	6	7	5 - 28
90-99	7	14	29 - 56
100-109	6	20	57 - 80
110-119	3	23	81 - 92
120-129	2	25	93 - 100

19

Cumulative Frequency_{ll}

- The cumulative frequency below the class containing 87.8% of the scores is 20
- cum f_{ll} = 20

Class	Freq.	C. Freq.	% C. Freq
70-79	1	1	0 - 4
80-89	6	7	5 - 28
90-99	7	14	29 - 56
100-109	6	20	57 - 80
110-119	3	23	81 - 92
120-129	2	25	93 - 100

20

Lower Real Limit and Width

- The lower true limit of the class containing 87.8 is:
  110 - (1 / 2) = 109.5
- X_{ll} = 109.5
- The width of the class is 10 (see previous width)

Class	Freq.	C. Freq.	% C. Freq
70-79	1	1	0 - 4
80-89	6	7	5 - 28
90-99	7	14	29 - 56
100-109	6	20	57 - 80
110-119	3	23	81 - 92
120-129	2	25	93 - 100

Cumulative Frequency$_{ll}$

- The number of observations in the class containing 87.8% of the scores is 3

Class	Freq.	C. Freq.	% C. Freq
70-79	1	1	0 - 4
80-89	6	7	5 - 28
90-99	7	14	29 - 56
100-109	6	20	57 - 80
110-119	3	23	81 - 92
120-129	2	25	93 - 100

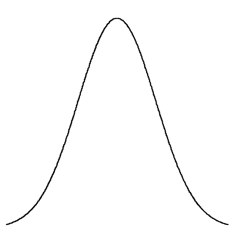
Plug and Chug

$$X_{PR} = X_{ll} + \frac{w(\text{cum } f_{PR} - \text{cum } f_{ll})}{f_i} = 109.5 + \frac{10(21.95 - 20)}{3} = 116$$

- X_{ll} = 109.5
- w = 10
- cum f_{PR} = 21.95
- cum f_{ll} = 20
- f_i = 3
- The score 116 corresponds to the percentile rank of 87.8%

Shapes of Distributions

- A distribution is a graphical means of presenting the frequency of continuous variables
- In psychology many distributions are approximately normal or Gaussian
  - They are bell shaped



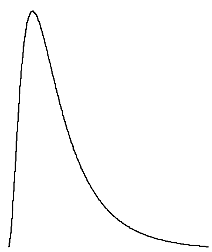
Skewness

- Some distributions are not symmetrical
- They have more observations in one tail of the distribution than in the other
- Such distributions are said to be *skewed*
- Skewness can be either *positive* or *negative*

25

Positively Skewed Distributions

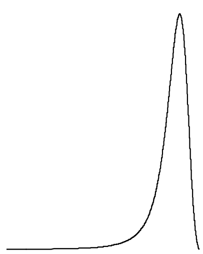
- A *positively skewed distribution* has more large observations than a normal distribution would have



26

Negatively Skewed Distributions

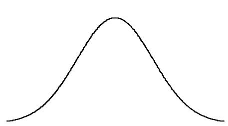
- A *negatively skewed distribution* has more smaller scores than a normal distribution would have



27

Kurtosis

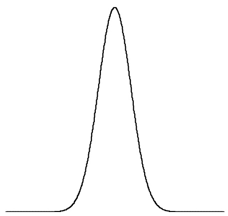
- The *kurtosis* of a distribution is a measure of how dispersed the scores are
- A normal distribution is said to be a *mesokurtic* distribution



28

Leptokurtic

- A *leptokurtic* distribution is less dispersed than a mesokurtic distribution
- That is, the scores tend to cluster more tightly about the center point



Platykurtic

- A *platykurtic* distribution is more dispersed than a mesokurtic distribution
- That is, the scores vary more from the center point than they do in a normal distribution