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.)

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.

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
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.

*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.

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.

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>70-79</td>
<td>1</td>
</tr>
<tr>
<td>80-89</td>
<td>6</td>
</tr>
<tr>
<td>90-99</td>
<td>7</td>
</tr>
<tr>
<td>100-109</td>
<td>6</td>
</tr>
<tr>
<td>110-119</td>
<td>3</td>
</tr>
<tr>
<td>120-129</td>
<td>2</td>
</tr>
</tbody>
</table>
**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.

**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.

<table>
<thead>
<tr>
<th>Class</th>
<th>Freq.</th>
<th>C. Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>70-79</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>80-89</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>90-99</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>100-109</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>110-119</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>120-129</td>
<td>2</td>
<td>25</td>
</tr>
</tbody>
</table>

**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.

**Percentile Rank**

\[
PR = \left( \frac{cum f_i + \left( \frac{|X - X_l|}{w} \right)}{f_i} \right) \times \frac{X_{100}}{N}
\]

- \(cum f_i\) = cumulative frequency of the class below \(X\)
- \(X\) = score to be converted to percentile rank
- \(X_l\) = 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
Cumulative Frequency, $X_i$

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Freq.</th>
<th>C. Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>70-79</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>80-89</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>90-99</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>100-109</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>110-119</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>120-129</td>
<td>2</td>
<td>25</td>
</tr>
</tbody>
</table>

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:

$$110 - \frac{1}{2} = 109.5$$

<table>
<thead>
<tr>
<th>Class</th>
<th>Freq.</th>
<th>C. Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>70-79</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>80-89</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>90-99</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>100-109</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>110-119</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>120-129</td>
<td>2</td>
<td>25</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Class</th>
<th>Freq.</th>
<th>C. Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>70-79</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>80-89</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>90-99</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>100-109</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>110-119</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>120-129</td>
<td>2</td>
<td>25</td>
</tr>
</tbody>
</table>

Calculating the Percentile Rank

\[ PR = \frac{\text{cum } f_i + \frac{X_i - X_l}{w}}{N} \times 100 = \frac{20 + \frac{116 - 109.5}{10}}{25} \times 100 = 87.8 \]

- $\text{cum } f_i = 20$
- $X_i = 116$
- $X_l = 109.5$
- $f_i = 3$
- $w = 10$
- $N = 25$
**Score Corresponding to a Percentile Rank (PR)**

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

\[
X_{\text{PR}} = X_{\ll} + \frac{w(\text{cum f}_{\text{PR}} - \text{cum f}_{\ll})}{f_i}
\]

**What Score Corresponds to a Percentile Rank of 87.8?**

- \( \text{cum f}_{\text{PR}} = \) the percentile rank times the number of scores divided by 100
- \( 87.8 \times 25 / 100 = 21.95 \)

<table>
<thead>
<tr>
<th>Class</th>
<th>Freq</th>
<th>C. Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>70-79</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>80-89</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>90-99</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>100-109</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>110-119</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>120-129</td>
<td>2</td>
<td>25</td>
</tr>
</tbody>
</table>

**Cumulative Frequency\(\ll\)**

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

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

**Cumulative Frequency\(\ll\)**

- The cumulative frequency below the class containing 87.8% of the scores is 20
- \( \text{cum f}_{\ll} = 20 \)
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)

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

Cumulative Frequency

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

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

Plug and Chug

- $X_{ll} = 109.5$
- $w = 10$
- $\text{cum } f_{pr} = 21.95$
- $\text{cum } f_{ll} = 20$
- $f_{1} = 3$
- 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

Positively Skewed Distributions

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

Negatively Skewed Distributions

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

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
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.