Chapter 3: Bivariate data analysis

Contents

3.1 Two-way tables
   ▶ Bivariate data
   ▶ Definition of a two-way table
   ▶ Joint absolute/relative frequency distribution
   ▶ Marginal frequency distribution
   ▶ Conditional frequency distribution

3.2 Correlation
   ▶ Graphical display: scatterplot
   ▶ Types of relationships between numerical variables
   ▶ Measures of linear association
Chapter 2: Bivariate data analysis

Recommended reading

- Peña, D., Romo, J., 'Introducción a la Estadística para las Ciencias Sociales'
  - Chapters 7, 8, 9
- Newbold, P. 'Estadística para los Negocios y la Economía' (2009)
  - Chapter 12

Bivariate data

- **Bivariate data** is obtained when we observe two characteristics (numerical or categorical) of one individual/object.
- **Notation:** A sample of $n$ bivariate observations will be written as $n$ pairs
  $$ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n). $$
In order to summarize the data, we create two-dimensional table ...

- with classes/class intervals of one variable in the rows ...
- and classes/class intervals of the other in the columns

- The corresponding frequencies (absolute or relative) appear in the cells on the cross-section of the given row and the given column

- When one variable is qualitative, the two-way table is called a contingency table

Two-way table: joint absolute frequency distribution

- Two-way table with \( k \) rows and \( m \) columns

\[
\begin{array}{cccc|c}
\text{X} & y_1 & \cdots & y_j & \cdots & y_m & \text{Total} \\
\hline
x_1 & n_{11} & \cdots & n_{1j} & \cdots & n_{1m} & n_{1\bullet} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\
x_i & n_{i1} & \cdots & n_{ij} & \cdots & n_{im} & n_{i\bullet} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\
x_k & n_{k1} & \cdots & n_{kj} & \cdots & n_{km} & n_{k\bullet} \\
\hline
\text{Total} & n_{\bullet1} & \cdots & n_{\bullet j} & \cdots & n_{\bullet m} & n_{\bullet \bullet}
\end{array}
\]

- Notation:
  - \( n_{ij} \) absolute frequency of cell \((i, j)\)
  - Row \( i \) total: \( n_{i\bullet} = n_{i1} + n_{i2} + \cdots + n_{im} \)
  - Column \( j \) total: \( n_{\bullet j} = n_{1j} + n_{2j} + \cdots + n_{kj} \)
  - \( n_{\bullet \bullet} \) sample size
Example Sixty four men were selected and the following two variables were considered

\( Y = \text{Color of man's mother's eyes} \{\text{Bright, Dark}\} \)

\( X = \text{Color of man's eyes} \{\text{Bright, Dark}\} \)

\((x_1, y_1) = (B, B), (x_2, y_2) = (B, D), \ldots, (x_{64}, y_{64}) = (D, D)\)

<table>
<thead>
<tr>
<th>X</th>
<th>Bright</th>
<th>Dark</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bright</td>
<td>23</td>
<td>12</td>
<td>35</td>
</tr>
<tr>
<td>Dark</td>
<td>17</td>
<td>12</td>
<td>29</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>24</td>
<td>64</td>
</tr>
</tbody>
</table>

\( (x_2, y_2) = (B, D) \) means that the 2nd man had Bright eyes and his mother had Dark eyes

\( 23 \) of men had B eyes and their mothers also had B eyes

\( 29 \) of men had Dark eyes

Two-way table: relative joint frequency distribution

\( f_{ij} = n_{ij} / n_{\bullet \bullet} \) relative frequency of cell \((i, j)\)

<table>
<thead>
<tr>
<th>X</th>
<th>( x_1 )</th>
<th>( \cdots )</th>
<th>( x_j )</th>
<th>( \cdots )</th>
<th>( x_k )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_1 )</td>
<td>( f_{11} )</td>
<td>( \cdots )</td>
<td>( f_{1j} )</td>
<td>( \cdots )</td>
<td>( f_{1m} )</td>
<td>( f_{1\bullet} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( x_j )</td>
<td>( f_{i1} )</td>
<td>( \cdots )</td>
<td>( f_{ij} )</td>
<td>( \cdots )</td>
<td>( f_{im} )</td>
<td>( f_{i\bullet} )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( \vdots )</td>
<td>( \ddots )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
</tr>
<tr>
<td>( x_k )</td>
<td>( f_{k1} )</td>
<td>( \cdots )</td>
<td>( f_{kj} )</td>
<td>( \cdots )</td>
<td>( f_{km} )</td>
<td>( f_{k\bullet} )</td>
</tr>
<tr>
<td>Total</td>
<td>( f_{1\bullet} )</td>
<td>( \cdots )</td>
<td>( f_{i\bullet} )</td>
<td>( \cdots )</td>
<td>( f_{m\bullet} )</td>
<td>1</td>
</tr>
</tbody>
</table>

\textbf{Notation:}

\( f_{ij} \) relative frequency of cell \((i, j)\)

Row \( i \) total: \( f_{i\bullet} = f_{i1} + f_{i2} + \cdots + f_{im} \)

Column \( j \) total: \( f_{\bullet j} = f_{1j} + f_{2j} + \cdots + f_{kj} \)

\( n_{\bullet \bullet} \) sample size
### Joint frequency distribution: absolute and relative

**Example**

$Y =$ Weekly number of visits to the cinema  
$X =$ Weekly number of visits to the theater  

(absolute frequencies in black, relative in blue)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>12 0.279</td>
<td>5 0.116</td>
<td>4 0.093</td>
<td>2 0.047</td>
<td>1 0.023</td>
</tr>
<tr>
<td>1</td>
<td>4 0.093</td>
<td>3 0.070</td>
<td>2 0.047</td>
<td>1 0.023</td>
<td>0 0.000</td>
</tr>
<tr>
<td>$X$ 2</td>
<td>3 0.070</td>
<td>3 0.070</td>
<td>2 0.047</td>
<td>0 0.000</td>
<td>0 0.000</td>
</tr>
<tr>
<td>3</td>
<td>1 0.023</td>
<td>0 0.000</td>
<td>0 0.000</td>
<td>0 0.000</td>
<td>0 0.000</td>
</tr>
</tbody>
</table>

### Marginal frequency distribution: absolute and relative

- The marginal frequencies take us back to the univariate case (we ignore the other variable)
- We obtain them by looking at the margins of the joint frequency table (absolute or relative)
- Specifically, the relative marginal frequencies of $X$ are defined as
  
  $$f_{*j} = f_{1j} + \cdots + f_{ij} + \cdots + f_{mj}$$

  and those of $Y$ are
  
  $$f_{ij} = f_{i1} + \cdots + f_{ij} + \cdots + f_{im}$$

- **Absolute marginal frequencies** are obtained analogously
Marginal frequency distribution

Example

\[ Y = \text{Number of workers} \]
\[ X = \text{Number of sales} \]

<table>
<thead>
<tr>
<th></th>
<th>1-24</th>
<th>25-49</th>
<th>50-74</th>
<th>75-99</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-100</td>
<td>0.293</td>
<td>0.122</td>
<td>0.098</td>
<td>0.049</td>
<td>0.561</td>
</tr>
<tr>
<td>101-200</td>
<td>0.098</td>
<td>0.073</td>
<td>0.049</td>
<td>0.024</td>
<td>0.244</td>
</tr>
<tr>
<td>201-300</td>
<td>0.073</td>
<td>0.073</td>
<td>0.049</td>
<td>0.000</td>
<td>0.195</td>
</tr>
<tr>
<td>Total</td>
<td>0.463</td>
<td>0.268</td>
<td>0.195</td>
<td>0.073</td>
<td>1</td>
</tr>
</tbody>
</table>

Marginal (relative) frequencies of \( Y \)

<table>
<thead>
<tr>
<th>Workers</th>
<th>1-24</th>
<th>25-49</th>
<th>50-74</th>
<th>75-99</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_{.j} )</td>
<td>0.463</td>
<td>0.268</td>
<td>0.195</td>
<td>0.073</td>
</tr>
</tbody>
</table>

Marginal (relative) frequencies of \( X \)

<table>
<thead>
<tr>
<th>Sales</th>
<th>1-100</th>
<th>101-200</th>
<th>201-300</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_{i.} )</td>
<td>0.561</td>
<td>0.244</td>
<td>0.195</td>
</tr>
</tbody>
</table>

Conditional frequency distribution

- The conditional distribution of \( Y \), given that \( X = x_i \) (notation \( f(Y|X = x_i) \)) is obtained by calculating
  \[
  \frac{f_{ij}}{f_{.j}}
  \]
  for \( j \)-th class/class interval of \( Y \) (we narrow our data set to the observations satisfying the condition)

- The conditional distribution of \( X \), given that \( Y = y_j \) is obtained by calculating for \( i \)-th class/class interval of \( X \)
  \[
  \frac{f_{ij}}{f_{i.}}
  \]

- Exactly the same conditional distribution is obtained if we replace the relative frequencies by the absolute frequencies in the above definitions

- Conditional frequencies always add up to 1!
### Conditional Frequency Distribution

**Example**

Y = Number of workers  
X = Number of sales

Find the conditional frequency distribution for sales given that the number of workers is between 50 and 74.

<table>
<thead>
<tr>
<th></th>
<th>1-24</th>
<th>25-49</th>
<th>50-74</th>
<th>75-99</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Y</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-100</td>
<td>0.293</td>
<td>0.122</td>
<td>0.098</td>
<td>0.049</td>
<td>0.561</td>
</tr>
<tr>
<td>X 101-200</td>
<td>0.098</td>
<td>0.073</td>
<td>0.049</td>
<td>0.024</td>
<td>0.244</td>
</tr>
<tr>
<td>201-300</td>
<td>0.073</td>
<td>0.073</td>
<td>0.049</td>
<td>0.000</td>
<td>0.195</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>0.463</td>
<td>0.268</td>
<td>0.195</td>
<td>0.073</td>
<td>1</td>
</tr>
</tbody>
</table>

The conditional frequency distribution \( f(X|50 \leq Y \leq 74) \)

<table>
<thead>
<tr>
<th>Sales</th>
<th>1-100</th>
<th>101-200</th>
<th>201-300</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f(x</td>
<td>50 \leq y \leq 74) )</td>
<td>0.500(= ( \frac{0.098}{0.195} ))</td>
<td>0.250(= ( \frac{0.049}{0.195} ))</td>
</tr>
</tbody>
</table>

---

### Conditional Frequency Distribution

**Example**

Y = Number of workers  
X = Number of sales

Find the conditional frequency distribution for workers when the number of sales is between 101 and 200.

<table>
<thead>
<tr>
<th></th>
<th>1-24</th>
<th>25-49</th>
<th>50-74</th>
<th>75-99</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Y</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-100</td>
<td>0.293</td>
<td>0.122</td>
<td>0.098</td>
<td>0.049</td>
<td>0.561</td>
</tr>
<tr>
<td>X 101-200</td>
<td>0.098</td>
<td>0.073</td>
<td>0.049</td>
<td>0.024</td>
<td>0.244</td>
</tr>
<tr>
<td>201-300</td>
<td>0.073</td>
<td>0.073</td>
<td>0.049</td>
<td>0.000</td>
<td>0.195</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>0.463</td>
<td>0.268</td>
<td>0.195</td>
<td>0.073</td>
<td>1</td>
</tr>
</tbody>
</table>

The conditional frequency distribution \( f(Y|101 \leq X \leq 200) \)

<table>
<thead>
<tr>
<th>Workers</th>
<th>1-24</th>
<th>25-49</th>
<th>50-74</th>
<th>75-99</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f(y</td>
<td>101 \leq x \leq 200) )</td>
<td>0.402</td>
<td>0.299</td>
<td>0.201</td>
</tr>
</tbody>
</table>
Types of relationships

There are different ways in which two variables may be related:
(i) absence of relationship (ii) negative linear relationship
(iii) positive linear relationship and (iv) nonlinear relationship
Measures of linear association: covariance

- We look for a descriptive measure that would indicate whether there is a linear relationship between $x$ and $y$

- Sample covariance is defined as

$$s_{xy} = \frac{1}{n-1} \left( \sum_{i=1}^{n} x_i y_i - \bar{x} \bar{y} \right)$$

- Covariance and linear transformation: if we transform $Y$ variable according to $Z = a + bY$, then the covariance between $X$ and the new variable $Z$ is

$$s_{xz} = bs_{xy}$$

Properties of the covariance

- If the covariance is 'much larger than 0', it's because there exists a positive linear relationship between the variables

- If the covariance is 'much smaller than 0', it's because there exists a negative linear relationship

- If the covariance is 'small', it's because:
  (i) the linear relationship does not exists; or
  (ii) the relationship is nonlinear.

- Drawbacks:
  - What does it mean large/small?
  - What are the units of covariance?
Measures of linear association: correlation

- Correlation overcomes the issues with covariance
- Sample correlation is defined as
  \[ r_{x,y} = \frac{s_{xy}}{s_x s_y} \]

- Properties:
  - The correlation is bounded: \(-1 \leq r_{x,y} \leq 1\), so the terms large and small now make sense
  - It is unitless

Example Three variables are measured over 91 countries: female life expectancy, male life expectancy and gross national product.

- The covariances between the three pairs of two variables are 105.15, 50066.04 and 57917.93, respectively.

- On the other hand, the correlations between the three pairs of two variables are 0.98, 0.64 and 0.65, respectively.

- Therefore, even if the covariances between male and female life expectancy and gross national product are larger than the covariance between male and female life expectancies, the correlation is larger for these last two variables.