MODUL STATISTIKA UNTUK BISNIS DAN MANAJEMEN

Disampaikan pada matrikulasi Program Studi Magister Manajemen

Fakultas Ekonomi dan Bisnis

Universitas Brawijaya Malang

Oleh
Dr. Ananda Sabil Hussein

Malang
2014
Describing Data: Frequency Tables, Distribution & Graphic Presentation

A. Constructing a Frequency Table

Frequency table is a grouping of qualitative data into mutually exclusive classes showing the number of observations in each class.

1. Relative Class Frequency

A relative frequency table shows the fraction of the number of frequencies in each class. **Example** 0.625, found by 50 divided by 80, is the fraction of domestic vehicles sold last month. The relative frequency distribution is shown in Table 2-2.

![Table 2-2 Relative Frequency Table of Vehicles Sold By Type At Whitner Autoplex Last Month](image)

<table>
<thead>
<tr>
<th>Vehicle Type</th>
<th>Number Sold</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domestic</td>
<td>50</td>
<td>0.625</td>
</tr>
<tr>
<td>Foreign</td>
<td>30</td>
<td>0.375</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>1.000</td>
</tr>
</tbody>
</table>

2. Graphic Presentation of Qualitative Data

A **bar chart** is a graphic representation of a frequency table and **pie chart** is a pie chart that shows the proportion or percent that each class represent of the total number of frequencies.

B. Constructing Frequency Distributions: Quantitative Data

Frequency distribution is a grouping of data into mutually exclusive classes showing the number of observations in each class.

**Example and Solution**

Ms. Kathryn Ball of AutoUSA wants tables, charts, and graphs to show the typical selling price on various dealer lots. Table 2-4 reports only the price of the 80 vehicles sold last month at Whitner Autoplex. What is the typical selling price? What is the highest selling price? What is the lowest selling price? Around what value do the selling prices tend to cluster? **TABLE 2-4 Prices of Vehicles Sold Last Month at Whitner Autople**
Step 1: Decide on the number of classes. This guide suggests you select the smallest number for the number of classes such that \(2^k\) (in words, 2 raised to the power of is greater than the number of observations \(n\)). In the Whitner Autoplex example, there were 80 vehicles sold. So \(n = 80\). If we try \(k = 6\), which means we would use 6 classes, then \(2^6 = 64\), somewhat less than 80. Hence, 6 is not enough classes if we let \(k = 7\), then \(2^7 = 128\), which is greater than 80. So the recommended number of classes is 7. 

**TABLE 2-5 An Example of Too Few Classes**

<table>
<thead>
<tr>
<th>Vehicle Selling Price ($)</th>
<th>Number of Vehicles</th>
</tr>
</thead>
<tbody>
<tr>
<td>15,000 up to 24,000</td>
<td>48</td>
</tr>
<tr>
<td>24,000 up to 33,000</td>
<td>30</td>
</tr>
<tr>
<td>33,000 up to 42,000</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
</tr>
</tbody>
</table>

Step 2: Determine the class Interval or width. Generally the class interval should be the same for all classes. The classes all taken together must cover at least the distance from the lowest value in the data up to the highest value. Expressing these words in a formula:

\[
i \geq \frac{H - L}{k}
\]
\[ i = \text{Class interval} \]
\[ H = \text{Highest observed value} \]
\[ L = \text{Lowest observed value} \]
\[ K = \text{Number of classes}. \]

In the Whitner Autoplex case, the lowest value is $15,546 and the highest value is $35,925. If we need 7 classes, the interval should be at least \((35,925 - 15,546)/7 = 2,911\). In practice this interval size is usually rounded up to some convenient number, such as a multiple of 10 or 100. The value of $3,000 might readily be used in this case.

**Step 3: Set the individual class limits.** For example in this text we will generally use the format $1,300 up to $1,400 and $1,400 up to $1,500 and so on. With this format it is clear that $1,399 goes into the first class and $1,400 in the second.

**Step 4: Tally the vehicle selling prices into the classes.**

<table>
<thead>
<tr>
<th>Class</th>
<th>Tallies</th>
</tr>
</thead>
<tbody>
<tr>
<td>$15,000 up to $18,000</td>
<td>III</td>
</tr>
<tr>
<td>$18,000 up to $21,000</td>
<td>IIIIII</td>
</tr>
<tr>
<td>$21,000 up to $24,000</td>
<td>IIIIII</td>
</tr>
<tr>
<td>$24,000 up to $27,000</td>
<td>IIIIII</td>
</tr>
<tr>
<td>$27,000 up to $30,000</td>
<td>III</td>
</tr>
<tr>
<td>$30,000 up to $33,000</td>
<td>I</td>
</tr>
<tr>
<td>$33,000 up to $36,000</td>
<td>I</td>
</tr>
</tbody>
</table>

**Step 5: Count the number of items in each class.**

<table>
<thead>
<tr>
<th>Selling Prices ($ thousands)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 up to 18</td>
<td>8</td>
</tr>
<tr>
<td>18 up to 21</td>
<td>23</td>
</tr>
<tr>
<td>21 up to 24</td>
<td>17</td>
</tr>
<tr>
<td>24 up to 27</td>
<td>18</td>
</tr>
<tr>
<td>27 up to 30</td>
<td>8</td>
</tr>
<tr>
<td>30 up to 33</td>
<td>4</td>
</tr>
<tr>
<td>33 up to 36</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>80</strong></td>
</tr>
</tbody>
</table>
1. Class Intervals and Class Midpoints
   ✓ Class intervals is the difference between the limits of two consecutive classes.
   ✓ Class midpoint is halfway between the limits of consecutive classes.
   ✓ Class frequency show the number of observations in each class.

2. Relative Frequency Distribution
   ✓ A relative frequency distribution shows the percent of observations in each class.

C. Graphic Presentation of a Frequency Distribution
   There are three methods for graphically portraying a frequency distribution.
   1. Histogram portrays the number of frequencies in each class in the form of a rectangle.
   2. A frequency polygon consists of line segments connecting the points formed by the intersection of the midpoint and the class frequency.
   3. A cumulative frequency distribution shows the number or percent of observations below given values

CHAPTER 3 Describing Data: Numerical Measures

A. The Population Mean
   To find the population mean, we use the following formula.

\[
\mu = \frac{\sum X}{N}
\]

where:
\( \mu \) represents the population mean. It is the Greek lowercase letter “mu.”
\( N \) is the number of values in the population.
\( X \) represents any particular value.
\( \sum \) is the Greek capital letter “sigma” and indicates the operation of adding.
\( \sum X \) is the sum of the \( X \) values in the population.
EXAMPLE

There are 12 automobile manufacturing companies in the United States.

<table>
<thead>
<tr>
<th>Company</th>
<th>Number of Patents Granted</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Motors</td>
<td>511</td>
</tr>
<tr>
<td>Nissan</td>
<td>385</td>
</tr>
<tr>
<td>DaimlerChrysler</td>
<td>275</td>
</tr>
<tr>
<td>Toyota</td>
<td>257</td>
</tr>
<tr>
<td>Honda</td>
<td>249</td>
</tr>
<tr>
<td>Ford</td>
<td>234</td>
</tr>
<tr>
<td>Mazda</td>
<td>210</td>
</tr>
<tr>
<td>Chrysler</td>
<td>97</td>
</tr>
<tr>
<td>Porsche</td>
<td>50</td>
</tr>
<tr>
<td>Mitsubishi</td>
<td>36</td>
</tr>
<tr>
<td>Volvo</td>
<td>23</td>
</tr>
<tr>
<td>BMW</td>
<td>13</td>
</tr>
</tbody>
</table>

Is this information a sample or a population? What is the arithmetic mean number of patents granted? This is a population because we are considering all the automobile manufacturing companies obtaining patents. From formula (3-1) we can find the result:

\[
\mu = \frac{511 + 385 + \cdots + 13}{12} = \frac{2340}{12} = 195
\]

How do we Interpret the value of 195? The typical number of patents received by an automobile manufacturing company is 195. Because we considered all the companies receiving patents, this value is a population parameter.

B. The Sample Mean

To find the mean for a sample:

\[
\bar{X} = \frac{\sum X}{n}
\]

where:
- \(\bar{X}\) is the sample mean. It is read “X bar.”
- \(n\) is the number of values in the sample.
C. **Properties of the Arithmetic Mean**

The arithmetic mean is a widely used measure of location. It has several important properties:

1. **Every set of interval or ratio level data has a mean.**
2. **All the values are included in computing the mean.**
3. **The mean is unique.** That is, there is only one mean in a set of data.
4. **The sum of the deviations of each value from the mean is zero.**

D. **The Weighted Mean**

The formula for determining the weighted mean is:

\[
\bar{X}_w = \frac{w_1 X_1 + w_2 X_2 + w_3 X_3 + \cdots + w_n X_n}{w_1 + w_2 + w_3 + \cdots + w_n}
\]

This may be shortened to:

\[
\bar{X}_w = \frac{\sum (wX)}{\sum w}
\]

Note that the denominator of a weighted mean is always the sum of the weights.

E. **The Median**

The median is the value in the middle of a set of ordered data and to find the median, sort the observations from smallest to largest and identify the middle value.

F. **The Mode**

The mode is the value that occurs most often in a set of data. First, the mode can be found for nominal-level data. Second, a set of data can have more than one mode.

G. **The geometric mean** is the \( n \)th root of the product of \( n \) positive values.

Formula for the geometric mean:
**H. Measures of Dispersion**

- The dispersion is the variation or spread in a set of data A. The range is the difference between the largest and the smallest value in a set of data.
- The formula for the range is: \( \text{Range} = \text{Largest value} - \text{Smallest value} \)
- The major characteristics of the range are: only two values are used in its calculation, it is influenced by extreme values, and easy to compute and to understand.
- The mean absolute deviation is the sum of the absolute values of the deviations from the mean divided by the number of observations.
- The formula for computing the mean absolute deviation is

\[
MD = \frac{\sum |X - \bar{X}|}{n}
\]

**I. Variance and Standard Deviation**

- Variance and Standard Deviation are also based on the deviations from the mean.
- The formula for the population variance is

\[
\sigma^2 = \frac{\sum (X - \mu)^2}{N}
\]
The formula for the sample variance is:

\[
\text{Sample Variance} \quad s^2 = \frac{\sum(X - \bar{X})^2}{n - 1} \quad [3-10]
\]

J. The standard deviation is the square root of the variance.

The major characteristics of the standard deviation are: it is in the same units as the original data, it is the square root of the average squared distance from the mean, it cannot be negative, and the most widely reported measure of dispersion.

The formula for the sample standard deviation is

\[
\text{Sample Standard Deviation} \quad s = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}} \quad [3-11]
\]

The formula for the standard deviation of grouped data is

\[
\text{Standard Deviation, Grouped Data} \quad s = \sqrt{\frac{\sum(M - \bar{X})^2}{n - 1}} \quad [3-13]
\]

CHAPTER 4 Describing Data: Displaying and Exploring Data

A. Dot Plots

A dot plot shows the range each of the values. Dot plots report the details of each observation. They are useful for comparing two or more data sets.

Dot plots are most useful for smaller data sets, whereas histograms tend to be most useful for large data sets.

B. Steam and Leaf Display

One technique to present a set of data, each numerical value is divided into two parts.

The leading digits becomes the stem and the trailing digit the leaf.

The stems are located along the vertical axis, and the leaf values are stacked against each other along the horizontal axis.
C. Other Measures of Dispersion

1. Quartiles, Deciles, and Precentiles
   - Quartiles divide a set of observations into four equal parts.
   - 25% of the observations are less than the 1^{st} quartile, 50% are less than the 2^{nd} quartile, and 75% are less than the 3^{rd} quartile.
   - The interquartile range is the difference between the third and the first quartile.
   - The formula for the Location of Percentile:

   \[
   L_p = (n + 1) \frac{P}{100}
   \]  

2. Box Plots
   - A box plot is a graphic display of a set of data and drawn enclosing the regions between the first and third quartiles.
   - A line is drawn inside the box at the median value.
   - Dotted line segments are drawn from the third quartile to the largest value to show the highest 25 percent of the values and from the first quartile to the smallest value to show the lowest 25 percent of the values.
   - A box plot is based on five statistics the maximum and minimum values, the first and third quartiles, and the median

D. SKEWNESS

- The coefficient of skewness is a measure of the symmetry of a distribution.
- The formula developed by Pearson is:
The coefficient of skewness computed by statistical software is:

\[
sk = \frac{n}{(n-1)(n-2)} \left[ \sum \left( \frac{X - \overline{X}}{s} \right)^3 \right]
\]

E. Describing the Relationship between Two Variables

✓ A scatter diagram is a graphic tool to portray the relationship between two variables.
✓ Both variables are measured with Interval or ratio scales.
✓ If the scatter of points moves from the lower left to the upper right, the variables under consideration are directly or positively related.
✓ If the scatter of points moves from the upper left to the lower right, the variables are inversely or negatively related.
✓ A contingency table is used to classify nominal scale observations according to two characteristics