Graphical Displays of Qualitative Data

Previously: Frequency and relative frequency tables

Next:

1. Pie chart

But,
Ex. Recall example 2.8 pg 50. Political affiliation of 40 students.

Responses: 13 Democratic
18 Republican
 9 Other

<table>
<thead>
<tr>
<th>Party</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrats</td>
<td>0.325</td>
</tr>
<tr>
<td>Republican</td>
<td>0.450</td>
</tr>
<tr>
<td>Other</td>
<td>0.225</td>
</tr>
</tbody>
</table>

Political Affiliation

- Democratic: 32.5%
- Republican: 45.0%
- Other: 22.5%
2. Bar Chart

Political Affiliation

Relative Frequency

Democratic  Republican  Other

Party
2.4 Stem and Leaf Diagrams

Steps:

1.

2.

3.

4.

Ex. Consider the following set of student examination marks

62  55  48  73  73  75  39
82  94  81  85  68  65  29
61  51  69  76  89  78  77
70  67  73  56
Ex. Examination marks

Stem-and-Leaf Display: Marks

Stem-and-leaf of Marks     N  = 25
Leaf Unit = 1.0

2 9
3 9
4 8
5 156
6 125789
7 0335678
8 1259
9 4

Stem-and-Leaf Display: Marks

Stem-and-leaf of Marks     N  = 25
Leaf Unit = 1.0

2 9
3
3 9
4
4 8
5 1
5 56
6 12
6 5789
7 0333
7 5678
8 12
8 59
9 4
2.5 Distribution shapes: Symmetry and Skewness

Shape

- important aspect of the dist’n of quantitative data

- plays a role in determining the correct method of statistical analysis

Distribution Shapes?

1. **Symmetric Dist’n**
   A dist’n that can be divided into two pieces that are
2. **Skewed Dist’n**
   A dist’n that is not symmetric is either
• Population and Sample Distributions

The dist’n of a population is called the population dist’n.

The dist’n of a sample is called the sample dist’n.

IMPORTANT!!

See figure 2.11, pg 75, in textbook.
3.1 Measures of Center

Measures of central location:
- 
- 
- 

1. **Mean**

Defined as the sum of the observations divided by the number of observations.

The mean is:
Ex. Number of days six patients of heart transplant survived:

3  15  46  64  64  623

The **sample mean** is

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

- Median

observations when arranged from smallest to largest.
Ex. Transplant data

3  15  46  64  64  623

Median =

Notice:

- If # of observations is odd,

- If # of observations is even,
- **Mode**

  Defined as the value that

  Ex. Transplant data

  $3 \hspace{1em} 15 \hspace{1em} 46 \hspace{1em} 64 \hspace{1em} 64 \hspace{1em} 623$

  Hence, Mode =

- **Note:**

  - If the greatest frequency is 1, then the

  - If the greatest frequency is 2 or greater, then any value that occurs with that
Ex. Transplant data.

\begin{align*}
3 & \quad 15 & \quad 46 & \quad 64 & \quad 64 & \quad 623 \\
\text{Mean} & = 135.8 \\
\text{Median} & = 55 \\
\text{Mode} & = 64
\end{align*}

Which of the three is a better measure of central location?

(See example 3.4 in textbook)
Relationship between mean and median

1. If the distribution is symmetrical:

2. If the distribution is skewed to the right (positively skewed):

3. If the distribution is skewed to the left (negatively skewed):
• **Population and Sample Mean**

  The mean of a population is called the **population mean**.

  The mean of a sample is called the **sample mean**.

**IMPORTANT!!**

For a particular population:
3.2 The sample mean

First, need a symbol to represent summation.

\[ \sum_{i=1}^{n} x_i = \]

Ex. Let \( x_1 = 2 \), \( x_2 = 3 \) and \( x_3 = 4 \) then

The **sample mean** of \( n \) measurements \( x_1, x_2, ..., x_n \)
is defined as:

\[ \bar{x} = \]
For example above:

\[ \bar{x} = \] 

Other important Sums:

Ex. \( x_1 = 2, \ x_2 = 3 \) and \( x_3 = 4 \) then

1. \[ \sum_{i=1}^{3} x_i^2 = \]

2. \[ \sum_{i=1}^{3} (x_i - \bar{x}) = \]

3. \[ \sum_{i=1}^{3} (x_i - \bar{x})^3 = \]
3.3 Measures of Variation; The Sample Standard Deviation

Ex. We have seven comparable fast food bars sample from cities A and B. Net profit (cents) per dollar sales are:

A: 5 4 6 5 5 3 7
B: 5 9 -4 23 5 -10 7

Both samples have

Dotplot: City A, City B

-14.0 -7.0 0.0 7.0 14.0 21.0
Measures of variation:

- 
- 
- 
- 

• **Range (pg104)**

  Range =

  Ex. City A: Largest value =
  Smallest value =

  Range =

  City B:

  Range =

  The range is:

  -
  -
  -
  -
• **Sample Standard Deviation** (pg105)

Measures the extent to which values differ from the mean:

Summing over all data values:

\[
\sum_{i=1}^{n} (x_i - \bar{x}) = \text{always!}
\]

Instead, look at squared differences
Find the sum of squared deviations:

and finally divide by $n-1$ (average of squared deviations).

**Variance of a sample** of $n$ measurements having mean $\overline{x}$ is:

$$s^2 =$$
Ex. City A

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$(x_i - \bar{x})$</th>
<th>$(x_i - \bar{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$$\bar{x} = 0$$

Hence, $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = $
Units of variance:
- Not very useful!

- Standard Deviation

Sample standard deviation

\[ s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}} \]

Ex. City A: \( s_A = \)

City B: \( s_B = \)

(Units are the same as the original data.)
• **Shortcut/Computational formula for sample variance**

\[ s^2 = \]

Ex. For City A:

\[ \sum_{i=1}^{7} x_i^2 = \]

\[ \sum_{i=1}^{7} x_i = \]

\[ s^2 = \quad = 1.66 \]
• Computing formula for sample SD

\[
s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2}{n - 1}}
\]

Note:

Ex.  \( s_B = 10.41 \)

Hence, data from city B displays
Empirical Rule (pg 112)

If a data set has a mound-shaped dist’n, the interval

\[ \bar{x} - s \] contains approx

\[ \bar{x} - 2s \] contains approx

\[ \bar{x} - 3s \] contains approx
Ex. Assume a sample has $x=40$ and $s=5$.

What can you say about the dist’n of the values in the sample?

If original dist’n **mound-shaped** then:

- Approx 68% of data lie
- Approx 95% of data lie
- Approx 99.7% of data lie