Chapter 4
Numerical Summaries – Mean and Standard Deviation

Data representations

- The histogram of observed data summarizes a large amount of information describing the process we have observed. Often more concise representations are needed.
  - Measures of central tendency – average, median, mode
  - Measures of variability – Standard deviation (standard error, root-mean-square), range and quartile and inter-quartile range
  - Inter-quartile range
  - Energy of the data (sum-squared)
  - Etc.

The average

- If we have to summarize a histogram, or any bar-plot for that matter, in only a few words what would these be?

Cross-sectional vs. Longitudinal Studies

- The avg. height of men appears to decrease with age. Should we conclude the avg. person’s getting shorter with time?
  - No, because this is a cross-sectional study – different subjects are compared to each other at one point in time.
  - In longitudinal studies – subjects/units are followed over time and compared with themselves.
  - Note that the people on the 20-30 yrs range are completely different from the folks in the 60-70 yrs of age. There’s evidence that with time men may be getting taller – an effect which is heavily confound with the effects of aging.
Average vs. Median

- Avg. weight for women 146 lb. Should we expect 50% below and 50% above the average?
  - No, in fact 41% are above and 59% are below the avg.
  - The histogram balances when supported on the average.
- The median of a histogram is the value in the middle with 50% of the observations above and 50% below the median.

Root Mean Square (R.M.S.)

- Consider {0, 5, -8, 7, -3}, the mean is: 0.2. But it’s also the mean of {0.1, 0.3, 0, 0.4, 0.2}. Of course, the 2 sequences of 5 numbers are very very different (e.g., size, sign, integer vs. double, etc.) So, the mean does not really represent all the info about the data!
- R.M.S. \((\{a_1, a_2, a_3, \ldots, a_n\})\) is: \(R.M.S. = \sqrt{\frac{1}{N} \sum_{k=1}^{N} a_k^2}\)

Standard Deviation (SD)

- The standard deviation is a measure of the spread of the data around its average. Most numbers in the data will be within 1 SD away from the average, and very few will be 2 SD’s, or more, away from the average.
- With the women’s height example we saw, 6,566 women ages 18-74 were surveyed, avg. height was 63.5 in and the SD was 2.5 in.
- Rule of thumb for data spreading:
  - Roughly 68% of all numbers from a list are within 1 SD of the average, and the other ~32% will be farther away. About 95% of the values will be within 2 SD’s away from the average.

Calculating the Standard Deviation

- SD = (almost) R.M.S. deviation from the average.
  - Let \((a_1, a_2, \ldots, a_N)\) are the observed values, then:
    \[ SD(\{a_1, a_2, \ldots, a_N\}) = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (a_k - \mu)^2} \]
  - Where the average (mean) \(\mu = \frac{1}{N} \sum a_k\)
  - Example, \(\{20, 10, 15, 15, 10\}\), \(\mu = \frac{1}{5} (20 + 10 + 15 + 15 + 10) = 15\)
    \[ SD = \sqrt{\frac{1}{N} \sum (a_k - \mu)^2} = \sqrt{\frac{1}{5} (5^2 + 5^2 + 0^2 + 0^2 + 5^2)} = \sqrt{\frac{50}{5}} = 4.1 \]

Calculating the Standard Deviation

- SD = (almost) R.M.S. deviation from the average.
  - Let \((a_1, a_2, \ldots, a_N)\) are the observed values, then:
    \[ SD(\{a_1, a_2, \ldots, a_N\}) = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (a_k - \mu)^2} \]
  - Note the difference between our and the textbook definition of SD, see Ch. 26.
Inter-quartile Range (IQR)

We talked about this earlier
At the end of Ch. 01
→ Chapter 5

Quartiles

The first quartile ($Q_1$) is the median of all the observations whose position is strictly below the position of the median, and the third quartile ($Q_3$) is the median of those above.

Five number summary

The five-number summary = (Min, Q₁, Med, Q₃, Max)

Inter-quartile Range

$IQR = Q_3 - Q_1$

Box plot compared to dot plot

Box plot

Dot plot

Construction of a box plot

Figure 2.4.4
Construction of a box plot.

Figure 2.4.3
Box plot for SYSVOL.

SYSVOL

50 100 150 200

Data

Q₁ Med Q₃

1.5 IQR

(pull back until hit observation)

(pull back until hit observation)

Scale

**Comparing 3 plots of the same data**

Stem-and-leaf of strength

N = 33

Leaf Unit = 10

<table>
<thead>
<tr>
<th>Strength</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 19 8</td>
</tr>
<tr>
<td>5</td>
<td>20 0334</td>
</tr>
<tr>
<td>10</td>
<td>21 20 00233</td>
</tr>
<tr>
<td>15</td>
<td>22 5568899</td>
</tr>
<tr>
<td>20</td>
<td>22 00011112</td>
</tr>
<tr>
<td>22.5</td>
<td>5 23 014</td>
</tr>
<tr>
<td>23</td>
<td>2 23</td>
</tr>
<tr>
<td>24</td>
<td>2 2</td>
</tr>
<tr>
<td>25.2</td>
<td>2 1</td>
</tr>
<tr>
<td>25.9</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 2.4.5 Three graphs of the breaking-strength data for gear teeth in positions 4 & 10 (Minitab output).

**Mean from a frequency table**

\[
\bar{x} = \frac{1}{n} \text{Sum of (value} \times \text{frequency of occurrence)} = \frac{1}{n} \text{(Sum of all observations)}
\]

**Frequency Table**

TABLE 2.5.1 Word Lengths for the First 100 Words on a Randomly Chosen Page

<table>
<thead>
<tr>
<th>Value u</th>
<th>Frequency f</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>

**Frequency Table for the Occurrence of Fish Species in Ocean Strata**

<table>
<thead>
<tr>
<th>No. of strata in which species occur</th>
<th>Frequency (No. of species)</th>
<th>Percentage of species ( \frac{f_j}{n} \times 100 )</th>
<th>Cumulative Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>117</td>
<td>35.5</td>
<td>35.5</td>
</tr>
<tr>
<td>2</td>
<td>61</td>
<td>18.5</td>
<td>53.9</td>
</tr>
<tr>
<td>3</td>
<td>37</td>
<td>11.2</td>
<td>65.2</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>7.3</td>
<td>72.4</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>7.0</td>
<td>79.4</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>3.6</td>
<td>83.0</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>4.2</td>
<td>87.3</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>3.0</td>
<td>90.3</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>2.7</td>
<td>93.0</td>
</tr>
<tr>
<td>10+</td>
<td>23</td>
<td>7.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>

\( n = 330 \) \( \times 100 \)

Source: Haedrich and Merrett [1988]

**Figure 2.5.1** Bar graph for species data.