Chapter 14. Descriptive Statistics

Sections 1 and 2. Graphical descriptions of data

Data is any information we gather from experiments or surveys. We’ll usually assume our data consists of a bunch of numbers. Suppose we give a quiz to a class, and the scores are:

\[\{5, 3, 9, 7, 5, 7, 5, 8, 5\}\]

This is called the data set. Especially when the data set is large, it is hard to understand the data set as a whole; we need to find ways to summarize it.

One method is to make a frequency table:

<table>
<thead>
<tr>
<th>Quiz score</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

There are also various graphical representations of the data: bar graphs, pictograms, pie charts, histograms; look at the book’s figures 1-11 of this chapter.

However we won’t dwell on this aspect of statistics.
Section 3. Numerical summaries of data

Often we try to give some numerical summaries of the data set.

The size $N$ of a data set is the number of numbers in it. Take our example:

\{5, 3, 9, 7, 5, 7, 5, 8, 5\} has $N = 9$.

We can also get this answer from the frequency table by adding the frequencies: $1+4+2+1+1=9$.

<table>
<thead>
<tr>
<th>Quiz score</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The average (or mean) of a data set is the sum of the numbers divided by $N$:

$$\frac{5+ 3 + 9 + 7 + 5 + 7 + 5 + 8 + 5}{9} = \frac{54}{9} = 6.$$  

With a large data set it’s easier to use the frequency table:

$$\frac{3*1 + 5*4 + 7*2 + 8*1 + 9*1}{9} = \frac{3 + 20 +14 + 8 + 9}{9} = 6.$$  

Multiply each number by its frequency, add these products up, and then divide by $N$. 

The **median** of a data set is another way of finding a type of middle point of the data set: half the numbers of the data set are smaller than the median, and half are larger.

The median of \( \{1, 3, 5, 9, 11\} \) is 5. First make sure that the numbers are arranged in increasing order. Then locate the point that separates the numbers into two equal parts.

\( \{1, 9, 6, 6, 2, 5, 4\} \) in increasing order is \( \{1, 2, 4, 5, 6, 6, 9\} \), and we see the median is 5, since three numbers are smaller than 5 and three are larger.

\( \{1, 2, 4, 8, 9, 12\} \) has \( N=6 \), which is even. In this case we can divide the numbers into the smaller three: 1, 2, 4 and the larger three: 8, 9, 12. Any number between 4 and 8 would separate the data set into two equal parts; in this case we let the median be the average of these two middle numbers:

\[
\text{median} = \frac{4 + 8}{2} = 6. 
\]

In the example we looked at earlier, there were nine numbers: 3, 5, 5, 5, 5, 7, 7, 8, 9; the median is the fifth number, which is 5.

Recall that the average was 6, so the average and median are different (this is usually the case). The average and the median are different ways of describing a “middle” or “typical” number from the data set. Each has its advantages.
Suppose a company has a boss and nine employees. Suppose the boss makes $200,000 and the employees $20,000 each. The median income is $20,000, which is the “typical income”; the average of these incomes is

\[
\frac{20,000 \times 9 + 200,000}{10} = \frac{380,000}{10} = 38,000.
\]

This is not close to anyone’s actual income in this company, but it does have the good property that if you multiply the average income by the number of people you get the total income.

The median of a data set splits the numbers into two halves: the smaller numbers and the larger numbers. Sometimes it is useful to split the numbers into four quarters. Here is an example: consider the data set \{1, 3, 13, 3, 5, 9, 9, 11, 13, 13, 8, 6, 5, 4\}. \(N = 15\).

1. **Sort the data:** 1, 3, 3, 4, 5, 5, 5, 6, 8, 9, 9, 11, 13, 13, 13

2. **Find the median:** \(\frac{1}{2} \times 15 = 7.5\), round up to 8; so the median is the 8\(^{th}\) number; thus, the median is 6.

3. **Find the first quartile, Q1:** \(\frac{1}{4} \times 15 = 3.75\), round up to 4; so Q1 is the 4\(^{th}\) number, so Q1 = 4.

4. **Find the third quartile, Q3:** \(\frac{3}{4} \times 15 = 11.25\), round up to 12; so Q3 is the 12\(^{th}\) number, so Q3 = 11.

5. **The median, M** is also called the second quartile, Q2.

1, 3, 3, \(4 = Q1\), 5, 5, 5, \(6 = M = Q2\), 8, 9, 9, \(11 = Q3\), 13, 13, 13
The smallest number in the data set is 1 (call this \textbf{min}, for minimum), and the largest is 13 (call this \textbf{max}, for maximum).

The \textbf{five-number summary} of a data set is: \textbf{min, Q1, M, Q3, max}.

For the above example this is: 1, 4, 6, 11, 13

Consider the already sorted data: \{2, 3, 5, 5, 6, 8, 8, 9\}; \text{N}=8;
\frac{1}{4} \times 8 = 2, \text{ so Q1 is the average of the 2\textsuperscript{nd} and 3\textsuperscript{rd} numbers } = (3+5)/2
\frac{1}{2} \times 8 = 4, \text{ so M is the average of the 4\textsuperscript{th} and 5\textsuperscript{th} numbers } = (5+6)/2
\frac{3}{4} \times 8 = 6, \text{ so Q3 is the average of the 6\textsuperscript{th} and 7\textsuperscript{th} numbers } = (8+8)/2

so we get: 2, 3, 4=Q1, 5, 5, 5.5=M=Q2, 6, 8 =Q3= 8, 9

The five-number summary is often represented graphically with a \textbf{box plot} (this won't be on the test):