Outline

Graphical and Numerical Representation of Data

Graphical representation of data
  Pie Charts

Histograms
  Drawing a Histogram
  Vertical Scale
Graphical and Numerical Representation of Data

In this section we will consider graphical representations of the distribution of a set of data. The goal is to identify the range of values and the most likely values as well as properties like symmetry, uni- or multi-modality, tail behavior, etc. We will also consider the common classification of data into different types.

To quantify the central value of the distribution of the sample we define the average and the median. To quantify the dispersion of the sample with respect to its central value we define the standard deviation.
Graphical representation of data

Pie Charts

The following pie charts correspond to the proportion of ice-cream flavors sold annually by a given brand
Descriptive Statistics
Pie Charts are a bad idea!

From the R manual page for the `pie` function:

About piecharts... Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data.
Cleveland (1985), page 264: "Data that can be shown by pie charts always can be shown by a dot chart. This means that judgements of position along a common scale can be made instead of the less accurate angle judgements." This statement is based on the empirical investigations of Cleveland and McGill as well as investigations by perceptual psychologists.

The same data are represented now with bar charts. Notice that this representation allows a very clear quantification of the differences between the ice-cream types.
Descriptive Statistics

![Graph showing Descriptive Statistics]

- Blueberry
- Cherry
- Apple
- Boston Cream
- Other
- Vanilla Cream

- Vanilla Cream
- Other
- Boston Cream
- Apple
- Cherry
- Blueberry

April 8th, 2015 | UCSC — AMS5
Histograms

Information is available from 131 hospitals.

We show a histogram of the average length of stay measured in days for each hospital.

The area of each block is proportional to the number of hospitals in the corresponding class interval.
Histogram of the average length of stay in hospital

length of stay (days)
Important Note

In this example all the intervals have the same length, so the heights of the blocks give all the information about the number of hospitals in each class.
Descriptive Statistics

There are 7 class intervals corresponding to

- 6 to 8 days
- 8 to 10 days
- 10 to 12 days
- 12 to 14 days
- 14 to 16 days
- 16 to 18 days
- 18 to 20 days
Note that the class that corresponds to 14 to 16 days is empty and that the class with the highest count of hospitals is the one of 8 to 10 days.
Descriptive Statistics

Drawing a Histogram

The starting point of a histogram is a distribution table.

Consider the distribution of families by income in the US in 1973.

In this table class intervals include the left point, but not the right point. It is important to specify which of the endpoints are included in each class.
## Descriptive Statistics

<table>
<thead>
<tr>
<th>Income level in $</th>
<th>percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 1,000</td>
<td>1</td>
</tr>
<tr>
<td>1,000 – 2,000</td>
<td>2</td>
</tr>
<tr>
<td>2,000 – 3,000</td>
<td>3</td>
</tr>
<tr>
<td>3,000 – 4,000</td>
<td>4</td>
</tr>
<tr>
<td>4,000 – 5,000</td>
<td>5</td>
</tr>
<tr>
<td>5,000 – 6,000</td>
<td>5</td>
</tr>
<tr>
<td>6,000 – 7,000</td>
<td>5</td>
</tr>
<tr>
<td>7,000 – 10,000</td>
<td>15</td>
</tr>
<tr>
<td>10,000 – 15,000</td>
<td>26</td>
</tr>
<tr>
<td>15,000 – 25,000</td>
<td>26</td>
</tr>
<tr>
<td>25,000 – 50,000</td>
<td>8</td>
</tr>
<tr>
<td>50,000 and over</td>
<td>1</td>
</tr>
</tbody>
</table>
Once the distribution table is available the next step is to draw a horizontal axis specifying the class intervals. Then we draw the blocks remembering that:

In a histogram, the areas of the blocks represent percentages

So, it is a mistake to set the heights of the blocks equal to the percentages in the table.

To figure out the height of a block divide the percentage by the length of the interval.
The table needed to calculate the heights of the blocks looks like:
## Descriptive Statistics

<table>
<thead>
<tr>
<th>Income level in $</th>
<th>percent</th>
<th>length (× $1,000)</th>
<th>height</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – 1,000</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1,000 – 2,000</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2,000 – 3,000</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3,000 – 4,000</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>4,000 – 5,000</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>5,000 – 6,000</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>6,000 – 7,000</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>7,000 – 10,000</td>
<td>15</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>10,000 – 15,000</td>
<td>26</td>
<td>5</td>
<td>5.2</td>
</tr>
<tr>
<td>15,000 – 25,000</td>
<td>26</td>
<td>10</td>
<td>2.6</td>
</tr>
<tr>
<td>25,000 – 50,000</td>
<td>8</td>
<td>25</td>
<td>.32</td>
</tr>
<tr>
<td>50,000 and over</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This is the resulting histogram. Notice that the class interval of incomes above $50,000 has been ignored.
Descriptive Statistics

Distribution of family income in the US in 1973
**Vertical Scale**

What is the meaning of the vertical scale in a histogram?

Remember that the area of the blocks is proportional to the percents. A high height implies that large chunks of area accumulate in small portions of the horizontal scale.

This implies that the density of the data is high in the intervals where the height is large. In other words, the data are more crowded in those intervals.