Chapter 1.1 – Displaying Distributions with Graphs

The distribution of a variable describes WHAT values the variable takes and HOW often it takes these values.

Depending on the type of the data (categorical or quantitative) we need to use different graphical and numerical tools to analyze and summarize the data at hand.

We will start by describing data graphically:
- **bar graphs**, **pie charts** and **pareto charts** can be used to graphically summarize categorical data.
- a common graphical display for quantitative data is a **histogram**.

**Strategies for data analysis:**
- examine each variable in data set individually and look for possible relationships among variables
- start with graphs and add numerical summaries as needed
Chapter 1.1 – Displaying Distributions with Graphs

Categorical variables: Bar graphs and Pie charts

Example: data on loan amounts owed by Asia to several foreign loaners:

<table>
<thead>
<tr>
<th>country</th>
<th>loan amount*</th>
<th>percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japan</td>
<td>97.2</td>
<td>35.8</td>
</tr>
<tr>
<td>Germany</td>
<td>31.7</td>
<td>11.7</td>
</tr>
<tr>
<td>France</td>
<td>24.6</td>
<td>9.1</td>
</tr>
<tr>
<td>US</td>
<td>23.8</td>
<td>8.8</td>
</tr>
<tr>
<td>Great Britain</td>
<td>16.3</td>
<td>5</td>
</tr>
<tr>
<td>others</td>
<td>80.9</td>
<td>29.8</td>
</tr>
<tr>
<td>Total</td>
<td>274.5</td>
<td>≈ 100**</td>
</tr>
</tbody>
</table>

* loan amount in Billions of Dollars.
** due to rounding percentages will add up to 100.2

Chapter 1.1 – Displaying Distributions with Graphs

The text in the lower left corner reads:

“The Areas of the blue, red, & black wedges are each measured from the center as the common vertex.

The blue wedges measured from the center of the circle represent area for the deaths from Preventable or Mitigable Zymotic diseases, the red wedges measured from the center the deaths from wounds, & the black wedges measured from the center the deaths from all other causes.

The black line across the red triangle in Nov. 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red, in January & February 1855, the blue coincides with the black.

The entire areas may be compared by following the blue, the red, & the black lines enclosing them.”

Source: Nightingale, Florence. Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army, 1858.
Chapter 1.1 – Displaying Distributions with Graphs

Pie Chart
A pie chart shows the amount of data that belong to each category as a proportional part of the circle

- useful when only one variable is of interest
- easy to compare: relative size of the parts to each other as well as the size of each part compared to the total
- percentage have to add up to 100, i.e. we need to account for all possible categories*

* Assume, we did not have complete information on loan amounts for all other countries: a pie chart can no longer be constructed because loan amounts for Japan, Germany, France, US and Great Britain do not account for 100% of loan amount given to Asia. (see also AYK problem 1.3, page 9)

Bar Graph
A bar graph shows the amount of data that belong to each category as proportionally sized rectangular areas. They are more flexible than pie charts because we don’t need to account for all possible categories of the variable.

- Bar graphs are valuable presentation tools as they are effective at reinforcing differences in magnitude (note that bars have to be of equal widths and are equally spaced)
- Bar graphs are (obviously) useful when the observed outcomes of the variable, our data, can be placed into different categories
- Bar graphs can be either horizontal or vertical
Chapter 1.1 – Displaying Distributions with Graphs

Comments:
- Often classes have a “natural order”, so it makes sense to put the bars in that order. For example consider how many Fr, So, J, S are enrolled in this 226 section.
- Sometimes, however, it is more useful to arrange the bars with respect to their magnitude, i.e. order them from tallest to shortest in order to highlight the categories with the highest frequencies (most important classes). This type of bar graph is called a Pareto chart.

Quantitative Variables: Frequency Tables and Histograms

Numerical data (observations are numbers rather than categories) are very common in business. One of the most common ways to summarize observations from a quantitative variable is a histogram.

Idea:
- group values into “classes” (intervals) of equal width
- count how many observations fall into each class
- draw a bar graph for the counts (keeping the intervals in numerical order, adjacent but non-overlapping)

Example: Ages and annual salaries for the CEOs of the 60 top ranked small companies in America in 1993. Source: Forbes, Nov. 8, 1993, “America’s Best Small Companies.”
Chapter 1.1 – Histograms

**General Guidelines for Drawing a Histogram**

- all classes have same width
- classes must not overlap, i.e. each observation can only belong to one class
- reasonable number of classes ($\sqrt{n}$ for samples of size $n \leq 150$ observations)
- there is no “best choice” of class width and number of classes ⇒ use good judgement to decide (see also next example for effect of class width)
- frequency $f$ for each class corresponds to number of observations that belong to that class
- sum of all class frequencies must add to $n$, the size of the sample

**Note:** Classifying data into classes leads to a **loss of information**, always be aware of that!!

---

**Different shapes of distributions/histograms**

We distinguish the following main shapes of distributions

- symmetric
- skewed to the right
- skewed to the left
- uniform/rectangular
- J-shaped

in addition we distinguish between **bimodal** and **multi-modal** distributions as opposed to uni-modal distributions.
Chapter 1.1 – Histograms

**Example:** Boston Housing Data - housing values for Boston suburbs, 506 observations on 14 different variables

<table>
<thead>
<tr>
<th>variable</th>
<th>description</th>
<th>type</th>
<th>units</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRIM</td>
<td>crime rate</td>
<td>numeric</td>
<td>per thousand</td>
</tr>
<tr>
<td>ZN</td>
<td>land value</td>
<td>numeric</td>
<td>$1000</td>
</tr>
<tr>
<td>INDUS</td>
<td>index of accessibility to employment centers</td>
<td>numeric</td>
<td></td>
</tr>
<tr>
<td>CHAS</td>
<td>Charles River dummy variable</td>
<td>numeric</td>
<td></td>
</tr>
<tr>
<td>NOX</td>
<td>nitric oxide concentration</td>
<td>numeric</td>
<td>ppm</td>
</tr>
<tr>
<td>RM</td>
<td>average number of rooms</td>
<td>numeric</td>
<td></td>
</tr>
<tr>
<td>AGE</td>
<td>proportion of owner-occupied units built prior to 1940</td>
<td>numeric</td>
<td></td>
</tr>
<tr>
<td>DIS</td>
<td>weighted distance to five Boston employment centers</td>
<td>numeric</td>
<td>miles</td>
</tr>
<tr>
<td>TAX</td>
<td>full-value property-tax rate</td>
<td>numeric</td>
<td>$1000</td>
</tr>
<tr>
<td>PTRATIO</td>
<td>pupil-teacher ratio</td>
<td>numeric</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>proportion of blacks by town</td>
<td>numeric</td>
<td></td>
</tr>
<tr>
<td>LSTAT</td>
<td>proportion of lower status of the population</td>
<td>numeric</td>
<td></td>
</tr>
</tbody>
</table>

Example: Boston Housing Data cont’d — different shapes of distributions

- Most outliers
- Encoded values
- Value of home
- Bar chart for the only categorical variable (variable 6)

**Describing distributions/histograms**

- overall shape of histogram
- center
- spread
- possible outliers
Chapter 1.1 – Stem-and-Leaf Plot (Stemplot)

- shows the actual digits
- Each numerical value is divided into two components
  - leading digits – stem
  - trailing digits – leaf

How to make a stem-and-leaf plot:

1. Separate each observation into a stem consisting of all but the last (most right) digit and a leaf (the last digit). Stems may have as many digits as needed, but each leaf contains only a single digit.
2. Write stems in a vertical column with smallest at the top, then draw a vertical line at the right if this column
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.
4. Add legend on how to read stem plot.

Example: random sample of package design ratings ranging from 0 - 45:

22 21 30 25 21 26 43 30 27 30 27 36 28 33 38 35 19 30 34 41
Chapter 1.1 – Time Plots

A trader at the stock exchange in Frankfurt.
Photo from NY Times, January 21, 2008

Stocks Plunge in Europe and Asia on U.S. Recession Fear

Time Plots

- A time-plot displays data that are observed over a given period of time.
- Helps us understand the behavior of the data over time.
- Always put time on the horizontal axis of your plot and the variable of interest in the vertical axis.

Time Series average number of occupied rooms By month

important features:
- trend
- seasonal variation
Chapter 1.1 – Time Plots

A trend in a time series refers to a long-term, persistent rise or fall, i.e., a positive (upward) or negative (downward) trend.

The southern oscillation is defined as the barometric pressure difference between Tahiti and the Darwin Islands at sea level.

Specifically, repeated southern oscillation values less than -1 typically define an El Niño.

Chapter 1.1 – Time Plots

A pattern in a time series that repeats itself at known regular intervals is called seasonal variation.

Time Series Sunspots By Year

The southern oscillation is a predictor of El Niño which in turn is thought to be a driver of world-wide weather.