Introductory Statistics Lectures

Visualizing Data

Descriptive Statistics I

Anthony Tanbakuchi
Department of Mathematics
Pima Community College

(Compile date: Tue May 19 14:48:11 2009)

Contents

1 Visualizing Data 1
1.1 Frequency distributions: univariate data . 2
Standard frequency distributions 3
Relative frequency distributions 4
Cumulative frequency distributions . . 5
1.2 Visualizing univariate quantitative data 5
Histograms 5
Dot plots 7
Stem and leaf plots . . . 7
1.3 Visualizing univariate qualitative data 8
Bar plot 8
Pareto chart 9
1.4 Visualizing bivariate quantitative data 11
Scatter plots 11
Time series plots 12
1.5 Summary 13

1 Visualizing Data

The Six Steps in Statistics

1. Information is needed about a parameter, a relationship, or a hypothesis needs to be tested.
2. A study or experiment is designed.
3. Data is collected.
4. **The data is described and explored.**
5. The parameter is estimated, the relationship is modeled, the hypothesis is tested.
6. Conclusions are made.

Descriptive statistics

Definition 1.1

Descriptive statistics.

summarize and describe collected data (sample).

Contrast with inferential statistics which makes inferences and generalizations about populations.

Useful characteristics for summarizing data

Example 1 (Heights of students in class). If we are interested in heights of students and we measure the heights of students in this class, how can we summarize the data in useful ways?

What to look for in data

shape are most of the data all around the same value, or are they spread out evenly between a range of values? How is the data **distributed**?

center what is the average value of the data?

variation how much do the values vary from the center?

outliers are any values significantly different from the rest?

Class height data

Load the class data, then look at what variables are defined.

```
R: load("ClassData.RData")
R: ls()
[1] "class.data"
R: names(class.data)
[1] "gender"             "height"
[3] "forearm"            "height_mother"
[5] "corrective_lenses"  "hair_color"
[7] "transportation"     "work_hours"
[9] "credits"
```

```
R: class.data$height
 [1] 65 68 71 66 68 65 62 68 77 62 69 70 65 63 67 73 68 70
```

Task: describe student heights

1. Determine general distribution (shape) of data.
2. Find a measure(s) of the center.
3. Find a measure of the variation.
4. Look for outliers.

1.1 Frequency distributions: univariate data

Definition 1.2

univariate data.

measurements made on only one variable per observation.

Definition 1.3

bivariate data.

measurements made on two variables per observation.

Definition 1.4

multivariate data.

measurements made on many variables per observation.

STANDARD FREQUENCY DISTRIBUTIONS

FREQUENCY DISTRIBUTION. DEFINITION 1.5

A table listing the **frequency** (number of times) data values occur in each interval. Good first summary of data!

Steps:

1. Choose number of classes (typically 5-20)
2. Calculate the class width:

$$\text{class width} \approx \frac{\text{max value} - \text{min value}}{\text{number of classes}} \tag{1}$$

3. Choose starting point, typically min data value or 0.
4. List in table lower class limits and then upper limits
5. Tally the data in each class, can use tick marks.

Given:

```
R: class.data$height
  [1] 65 68 71 66 68 65 62 68 77 62 69 70 65 63 67 73 68 70
```

(min=62, max=77, n=18)

Question 1. Construct a frequency table for the class heights:

Information in a frequency distribution

1. shape
2. center
3. variation
4. outliers

RELATIVE FREQUENCY DISTRIBUTIONS

DEFINITION 1.6 RELATIVE FREQUENCY DISTRIBUTION.

A table of **relative** frequency counts.

$$\text{relative frequency } \% = \frac{\text{class freq}}{\text{total number}} 100\% \tag{2}$$

Useful for comparing data.

Example 2. Add a relative frequency column onto the student height table.

CUMULATIVE FREQUENCY DISTRIBUTIONS

Cumulative frequency distribution. Definition 1.7

Table listing the **total counts less than** the upper class limits for each interval.

Steps to construct:

1. Make a frequency distribution
2. Make a second table and change the intervals to **less than** [**upper limit**].
3. Each frequency is the sum of all the previous frequencies.
   We **accumulate** the frequencies.

Example 3. Make a cumulative frequency distribution for student heights.

1.2 Visualizing univariate quantitative data

HISTOGRAMS

Histogram. Definition 1.8

A bar plot of a frequency histogram table.

x-axis class boundaries (or midpoints)

y-axis frequency (counts)

(No spaces between bars.)

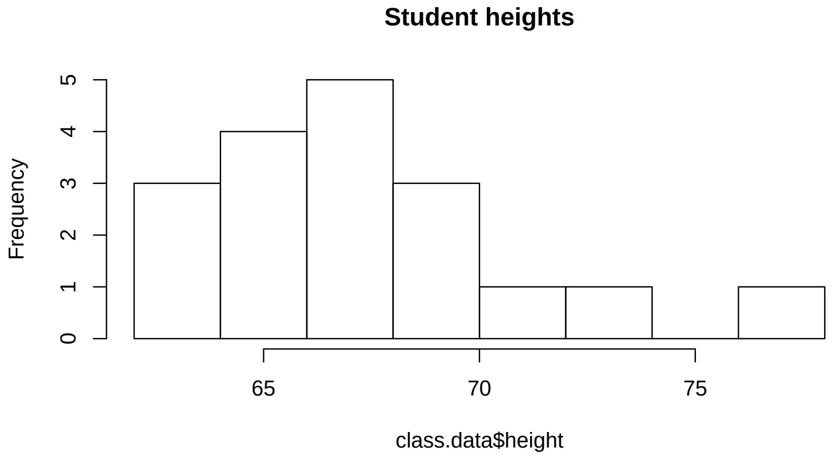
Histogram: R Command

`hist(x)`

Where `x` is a vector of data.

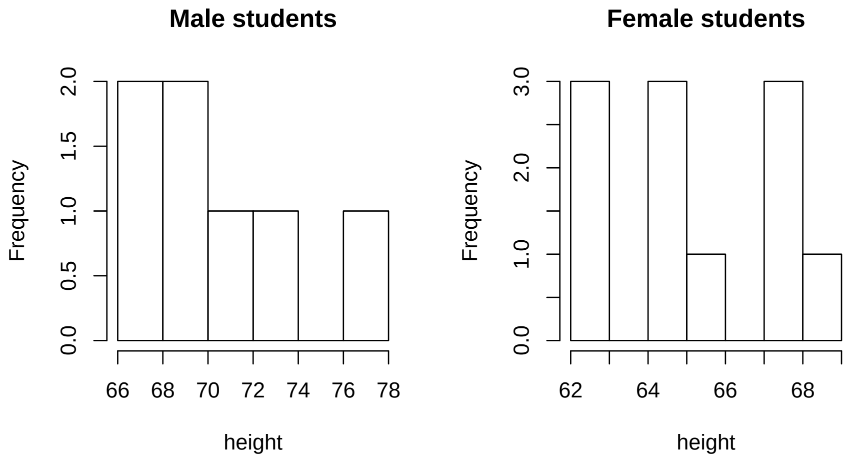
Example 4. Histogram of student heights

```
R: hist(class.data$height, main = "Student heights")
```



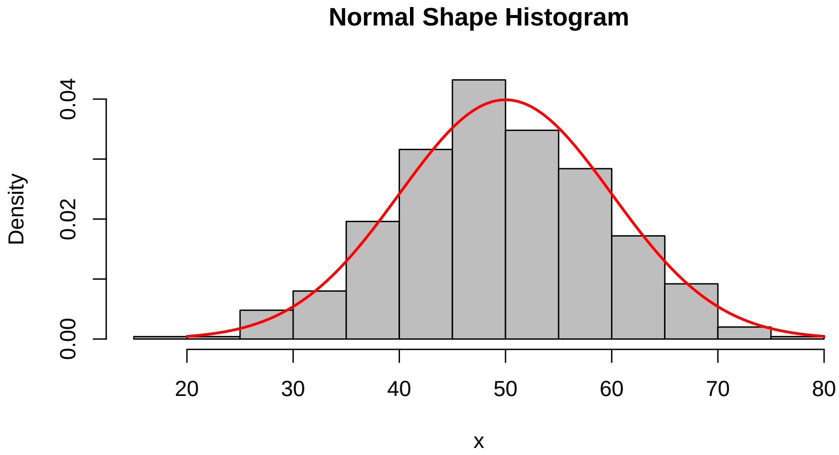
Example 5. Comparison of student heights versus gender:

```
R:  par(mfrow = c(1, 2))
R:  hist(class.data$height[class.data$gender == "MALE"],
+       main = "Male students", xlab = "height")
R:  hist(class.data$height[class.data$gender == "FEMALE"],
+       main = "Female students", xlab = "height")
```



Normal Shaped Distribution

Example of a normal shaped histogram (sometimes called a "bell shape").



Red Line: Ideal Normal Distribution

DOT PLOTS

STEM AND LEAF PLOTS

Stem and leaf plot. Definition 1.9

A character based plot similar to a histogram that breaks up data into the stem and leaf "plot". Quick and easy histogram.

stem like the classes of a histogram, generally the most significant digit of the data.

leaf digit to the right of the stem.

For each data point it's leaf value (digit) is recorded to the right of the stem. More leaves indicate more values in the stem's class.

Stem and leaf plot: R Command

`stem(x)`

Where `x` is a vector of data.

Example 6. Stem and leaf plot of student heights

```
R: stem(class.data$height)
  The decimal point is 1 digit(s) to the right of the |

  6 | 223
  6 | 5556788889
  7 | 0013
  7 | 7
```

Below is a stem and leaf plot for a small set of data stored in `x`.

```
R: stem(x)
  The decimal point is at the |

  5 | 558
  6 | 04
  7 |
  8 | 3
```

Question 2. Reconstruct the data set from the above plot.

Other ways to visualize frequency data

relative frequency histogram bar plot with vertical axis in **percent** rather than counts.

frequency polygon just like a histogram but uses line segments instead of bars.

ogive "oh-jive" line graph of cumulative frequencies.

1.3 Visualizing univariate qualitative data

Univariate qualitative data is categorical data concerning one variable.

Example 7 (Univariate qualitative data). student hair color is a single variable that is qualitative.

```
R: class.data$hair_color
 [1] BROWN BROWN BLACK BLOND BROWN BLACK BROWN BROWN BROWN
[10] BROWN BLOND BROWN BROWN BROWN BROWN BROWN BLACK BROWN
Levels: BLACK BLOND BROWN
```

Summarizing qualitative data

R Command

Frequency table:
`table(x)`
Where `x` is a vector of data.

Example 8. Summarizing student hair color.

```
R: table(class.data$hair_color)
BLACK BLOND BROWN
    3     2    13
```

We will now look at methods for visualizing these types of tables.

BAR PLOT

Definition 1.10

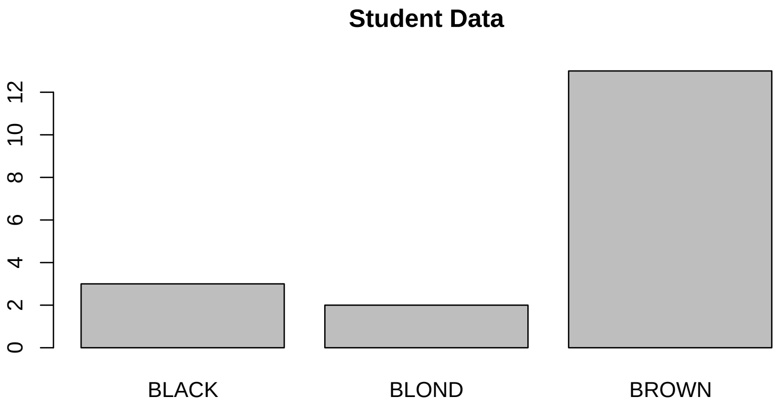
Bar plot.
Like a histogram but for qualitative data frequencies. Each bar represents the count of a category.
x-axis categories
y-axis counts or proportions.
Misleading if y-axis does not start at 0.
Makes sense to put spaces between bars since data is not continuous.

R Command

Bar plot:
`t=table(x); barplot(t)`
Where `x` is a vector of categorical data and we plot the frequency table `t` .

Example 9. Bar plot of student hair color

```
R: t = table(class.data$hair_color)
R: barplot(t, main = "Student Data")
```



PARETO CHART

Pareto chart. Definition 1.11

A bar plot of categorical data frequencies **sorted in decreasing frequency**.

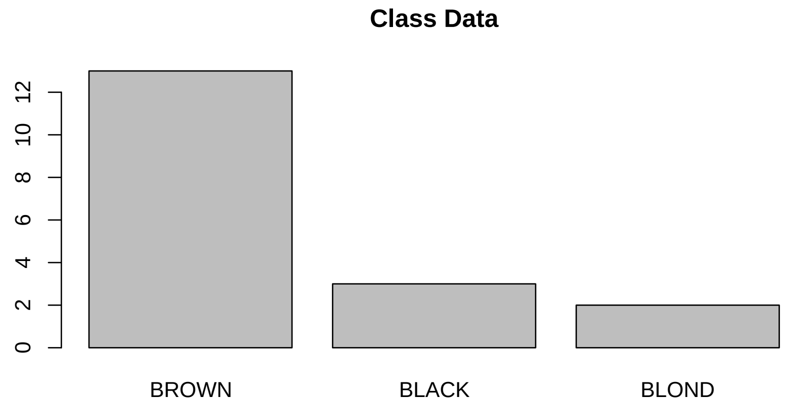
Pareto chart: R Command

```
t=table(x); t=sort(t, decreasing=TRUE); barplot(t)
```

Where x is a vector of categorical data and we plot the **sorted** frequency table t .

Example 10. Pareto chart of student hair color.

```
R: t = table(class.data$hair_color)
R: t = sort(t, decreasing = TRUE)
R: t
BROWN BLACK BLOND
   13     3     2
R: barplot(t, main = "Class Data")
```



Pie charts

DEFINITION 1.12

PIE CHART.

Used to display relative frequencies (or proportions) for categorical data. The **area** represents the relative frequency.

Although widely found in the media, pie charts are no longer commonly used by statisticians since **it's hard to gauge the areas visually**.

R COMMAND

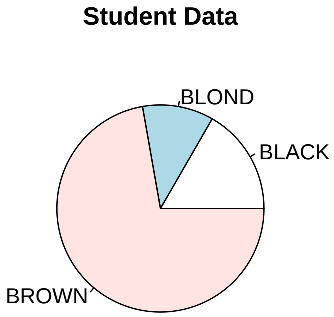
PIE CHART:

`t=table(x); pie(t)`

Where `x` is a vector of categorical data and we plot the frequency table `t`.

Example 11. Pie chart of student hair color.

```
R: t = table(class.data$hair_color)
R: pie(t, main = "Student Data")
```



1.4 Visualizing bivariate quantitative data

Bivariate quantitative data. Definition 1.13

is **paired** data dealing with two variables. For each measurement of the first variable there is a corresponding measurement in the second variable.

Example 12 (Bivariate data). We may be interested to study the relationship between **height** and **weight** of students. For each student in our study we measure both variables, the height and weight.

We will not look at ways to visualize bivariate quantitative data.

SCATTER PLOTS

Scatter plot. Definition 1.14

A plot of paired (x, y) data. The two variables are represented by x and y.
horizontal axis x values.
vertical axis y values.

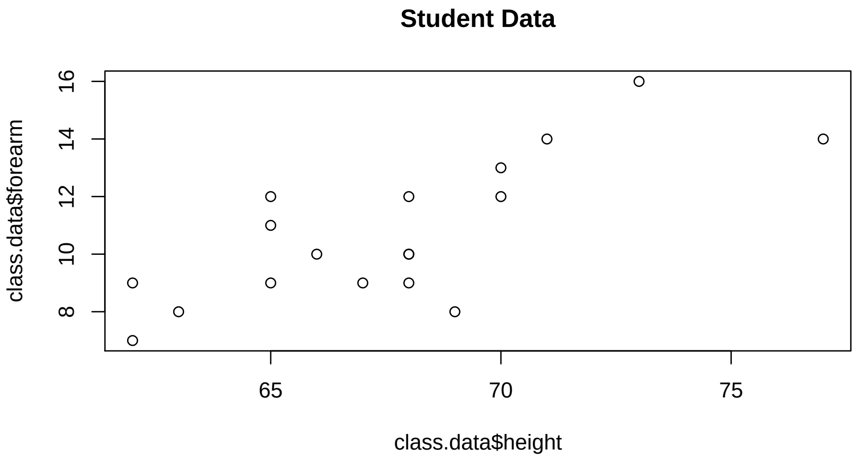
Scatter plot: R Command

```
plot(x,y)
```

Where both `x` and `y` are **ordered** vectors of data. Ordered meaning the first element of `x` corresponds to the first element of `y`.

Example 13. Scatter plot of student height vs forearm length.

```
R: plot(class.data$height, class.data$forearm, main = "Student ↩
   Data")
```



TIME SERIES PLOTS

Definition 1.15

Time series plot.

Plots data points as a function of time.

x-axis time

y-axis values of measured variable.

Example 14. Measure the height of a bean sprout over time. Plot the pairs of data (time, height) using a scatter diagram with x variable being time.

R Command

Time series plot plot:

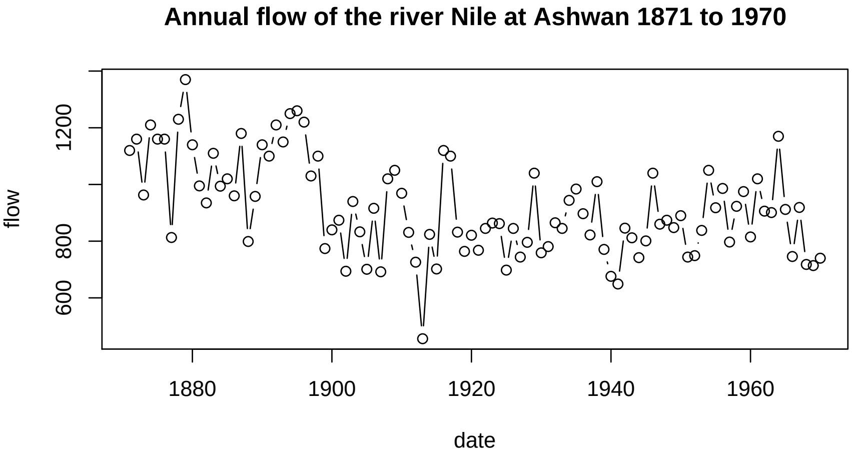
`plot(t,y, type="b")`

Where both `t` and `y` are **ordered** vectors of data. Ordered meaning the first element of `t` corresponds to the first element of `y` . The optional argument `type="b"` makes a plot with both points and lines.

Example 15. Time series plot of Nile River flow data.

```
R: date = 1871:1970
R: flow = c(1120, 1160, 963, 1210, 1160, 1160, 813,
+     1230, 1370, 1140, 995, 935, 1110, 994, 1020,
+     960, 1180, 799, 958, 1140, 1100, 1210, 1150,
+     1250, 1260, 1220, 1030, 1100, 774, 840, 874,
+     694, 940, 833, 701, 916, 692, 1020, 1050,
+     969, 831, 726, 456, 824, 702, 1120, 1100,
+     832, 764, 821, 768, 845, 864, 862, 698, 845,
+     744, 796, 1040, 759, 781, 865, 845, 944, 984,
+     897, 822, 1010, 771, 676, 649, 846, 812, 742,
+     801, 1040, 860, 874, 848, 890, 744, 749, 838,
+     1050, 918, 986, 797, 923, 975, 815, 1020,
+     906, 901, 1170, 912, 746, 919, 718, 714, 740)
```

```
R: plot(date, flow, type = "b", main = "Annual flow of the river ↩
    Nile at Ashwan 1871 to 1970")
```



If you want to make a time series plot, you will typical define `t=1871:1970` , `y=c(`y_1`,`y_2`,...)` , then `plot(t, y, type="b")` .

1.5 Summary

Summary

- Univariate quantitative data (look for shape, center, variation, outliers)
  1. Frequency table: standard, cumulative, relative
  2. Histogram: visual for frequency table `hist(x)`, `stem(x)`
- Univariate qualitative data:
  1. Frequency table: `t=table(x)`
  2. Pareto chart: `barplot(sort(t, decreasing = TRUE))`
- Bivariate data:
  1. Scatter plot: `plot(x, y)`
  2. Time series plot: `plot(t, y, type="b")`

Additional Examples to try

Example 16. Take a look at the built in `morley` table of data in R. This is the classical data from the Michaelson and Morley experiment that measured the speed of light (values are in km/s minus 299,000). The data consists of five experiments, each consisting of 20 consecutive runs.

- Make a histogram of the speed of light. What does the histogram tell you about the data?
- Make a stem and leaf plot of the data. How does it compare with the histogram?
- Are there any outliers?

- What is the range of the data?
- What value are most of the measurements around?

Example 17. Take a look at the built in `women` table of data in R. This data set gives the average heights and weights for American women aged 30 to 39. Make a scatter plot of height vs. weight. Are any trends visible?