Exploratory data analysis (Chapter 2)

Fall 2011

- Data Examples

Example 1: Survey Data

1. Data collected from a Stat 371 class in Fall 2005
2. They answered questions about their: gender, major, year in school, miles from home, height, blood type, number of brothers, number of sisters.

Example 2: Milk Production

1. Milk data: milk yields (lbs/day) were collected from a herd of 14 cows on a single day.
2. Data: 44, 55, 37, 32, 37, 26, 23, 41, 34, 19, 30, 39, 46, 44.
3. In R

```
milk = c(44, 55, 37, 32, 37, 26, 23, 41, 34, 19,
```

Data Summaries

- What is the ‘best’ way to summarize these data sets?
- First step is to summarize each variable in the data set.
- Then, the best way to summarize a variable depends on its characteristics.
- Consider numerical summaries and graphical summaries.

- Introduction and Definitions (Section 2.1)

Exploratory Data Analysis

- Exploratory Data Analysis involves both graphical displays of data and numerical summaries of data.
- A data set is often represented as a matrix.
- There is a row for each unit.
- There is a column for each variable.
- A unit is an object that can be measured, such as a person, or a thing.
- A variable is a characteristic of a unit that can be assigned a number or a category.
- For the survey data, each respondent is a unit.
- Variables include sex, major, year in school, miles from home, height, and blood type.

Data

Earlier Cow Study

treatment	level	lactation	age	initial.weight	dry	milk	fat	solids	final.weight	protein
control	0	3	49	1360	15.429	45.552	3.88	8.96	1442	3.67
control	0	3	47	1498	18.799	66.221	3.40	8.44	1565	3.03
control	0	2	36	1265	17.948	63.032	3.44	8.70	1315	3.40
control	0	2	33	1190	18.267	68.421	3.42	8.30	1285	3.37
control	0	2	31	1145	17.253	59.671	3.01	9.04	1182	3.61
control	0	1	22	1035	13.046	44.045	2.97	8.60	1043	3.03
low	0.1	6	89	1369	14.754	57.053	4.60	8.60	1268	3.62
low	0.1	4	74	1656	17.359	69.699	2.91	8.94	1593	3.12
low	0.1	3	45	1466	16.422	71.337	3.55	8.93	1390	3.30
low	0.1	2	34	1316	17.149	68.276	3.08	8.84	1315	3.40
low	0.1	2	36	1164	16.217	74.573	3.45	8.66	1168	3.31
low	0.1	2	41	1272	17.986	66.672	3.43	9.19	1188	3.59
medium	0.2	3	45	1362	19.998	76.604	4.29	8.44	1273	3.41
medium	0.2	3	49	1305	19.713	64.536	3.94	8.82	1305	3.21
medium	0.2	3	48	1268	16.813	71.771	2.89	8.41	1248	3.06
medium	0.2	3	44	1315	15.127	59.323	3.13	8.72	1270	3.26
medium	0.2	2	40	1180	19.549	62.484	3.36	8.51	1285	3.21
medium	0.2	2	35	1190	19.142	70.178	3.92	8.94	1168	3.28
high	0.3	5	81	1458	20.458	71.558	3.69	8.48	1432	3.17
high	0.3	3	49	1515	19.861	56.226	4.96	9.17	1413	3.72
high	0.3	3	48	1310	18.379	49.543	3.78	8.41	1390	3.67
high	0.3	3	46	1215	18.000	55.351	4.22	8.94	1212	3.80
high	0.3	3	49	1346	19.636	64.509	4.16	8.74	1318	3.31
high	0.3	3	46	1428	19.586	74.430	3.92	8.75	1333	3.37

Variables

- Variables are either quantitative or categorical.
- In a categorical variable, measurements are categories.
- Examples include blood type, sex.
- The variable year in school is an example of an ordinal categorical variable, because the levels are ordered.
- Quantitative variables record a number for each unit.
- Examples include height, which is continuous and number of sisters, which is discrete.
- Often, continuous variables are rounded to a discrete set of values (such as heights to the nearest inch or half inch).
- We can also make a categorical variable from a continuous variable by dividing the range of the variable into classes (So, for example, height could be categorized as short, average, or tall).
- Identifying the types of variables can be important because some methods of statistical analysis are appropriate only for a specific type of variable.

Samples

- A sample is a collection of units on which we have measured one or more variables.
- The number of observations in a sample is called the sample size.
- Common notation for the sample size is n.
- The textbook adopts the convention of using uppercase letters for variables and lower case letters for observed values.

Types of variables (Let’s Summarize)

Examples: data from the survey.

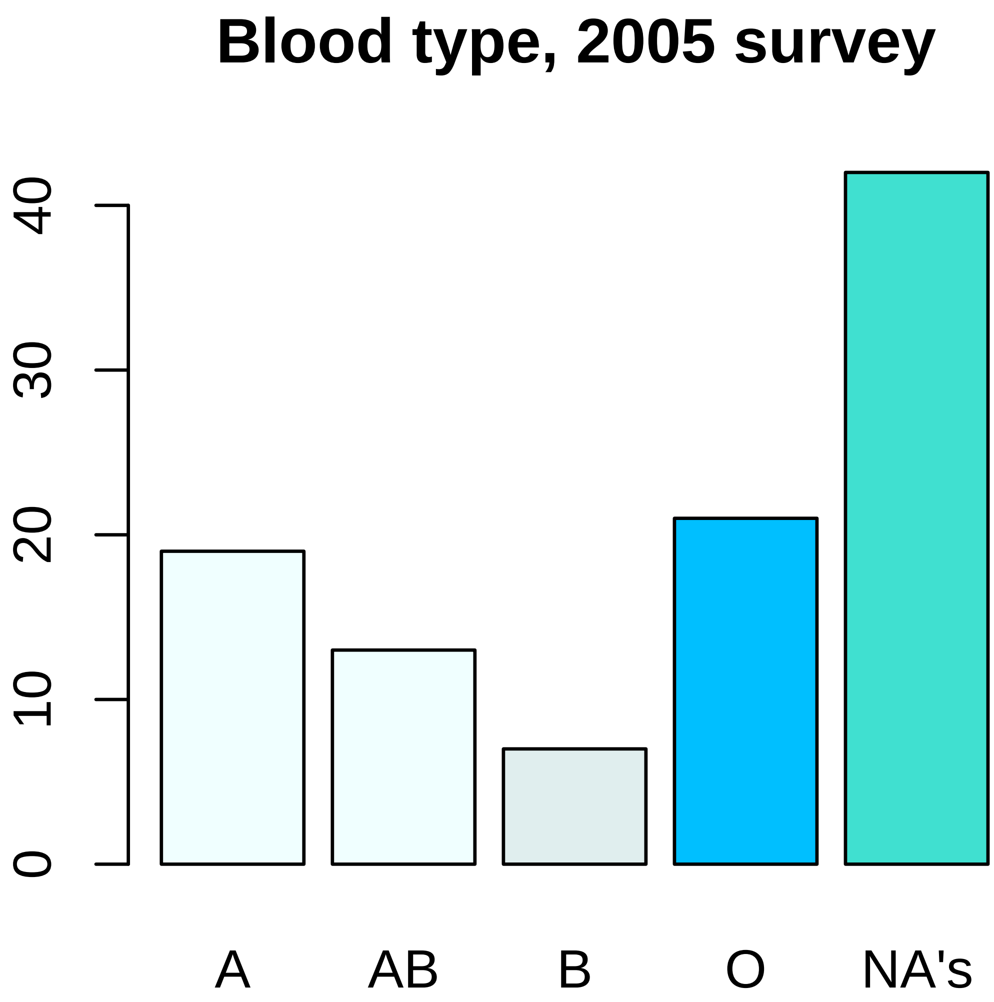
- **Categorical** (qualitative)
  - nominal: Sex, blood type
  - ordinal: Year in school
- **Numerical** (quantitative)
  - continuous: Height, Miles from home
  - discrete: # brothers, # sisters

- Visual Summaries (Graphical Displays, Section 2.2-2.3)

Summaries of Categorical Variables

- A frequency distribution is a list of the observed categories and a count of the number of observations in each.
- A frequency distribution may be displayed with a table or with a bar chart.
- For ordinal categorical random variables, it is conventional to order the categories in the display (table or bar chart) in the meaningful order.
- For non-ordinal variables, two conventional choices are alphabetical and by size of the counts.
- The vertical axis of a bar chart may show frequency or relative frequency.
- It is conventional to leave space between bars of a bar chart of a categorical variable.

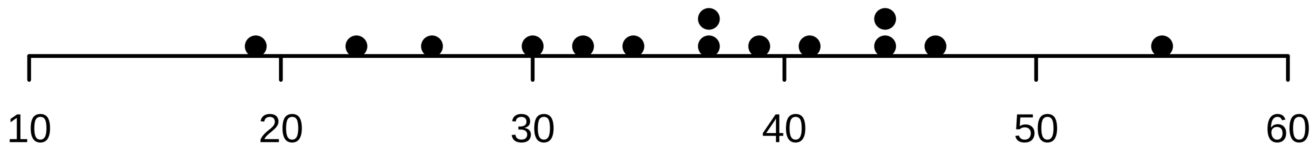
Example: Bar Chart for Blood Type



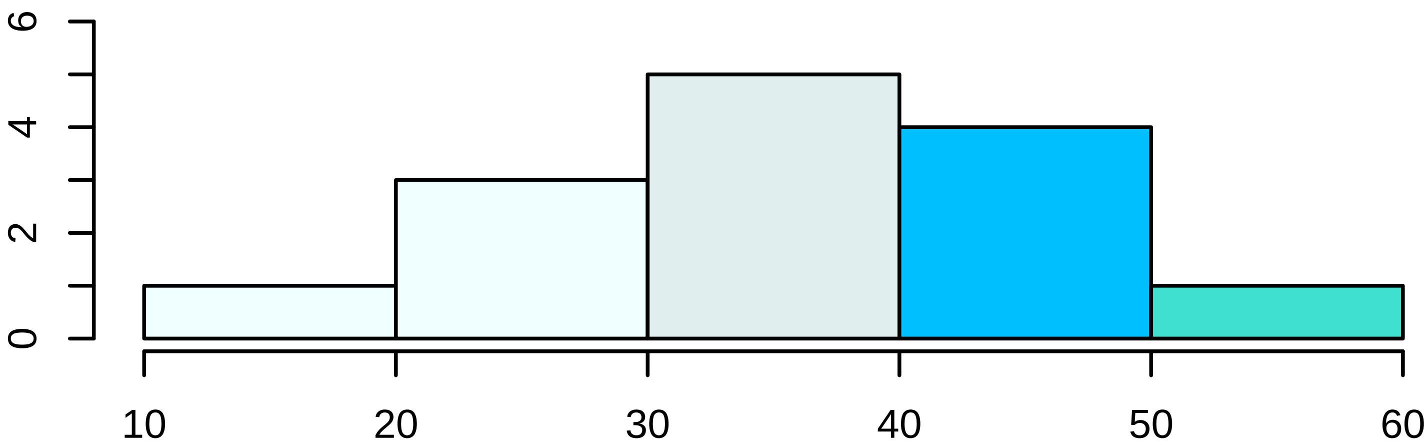
Summaries of Quantitative Variables

- Quantitative variables from very small samples can be displayed with a dotplot.
- Histograms are a more general tool for displaying the distribution of quantitative variables.
- A histogram is a bar graph of counts of observations in each class, but no space is drawn between classes.
- If classes are of different widths, the bars should be drawn so that areas are proportional to frequencies.
- Selection of classes is arbitrary. Different choices can lead to different pictures.
- Too few classes is an over-summary of the data.
- Too many classes can cloud important features of the data with noise.

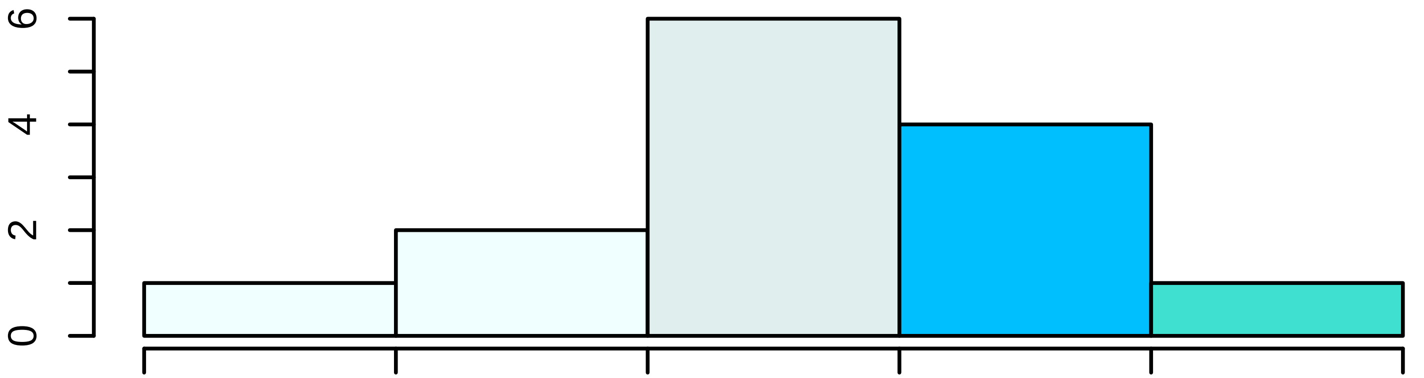
Dot−plot of milk



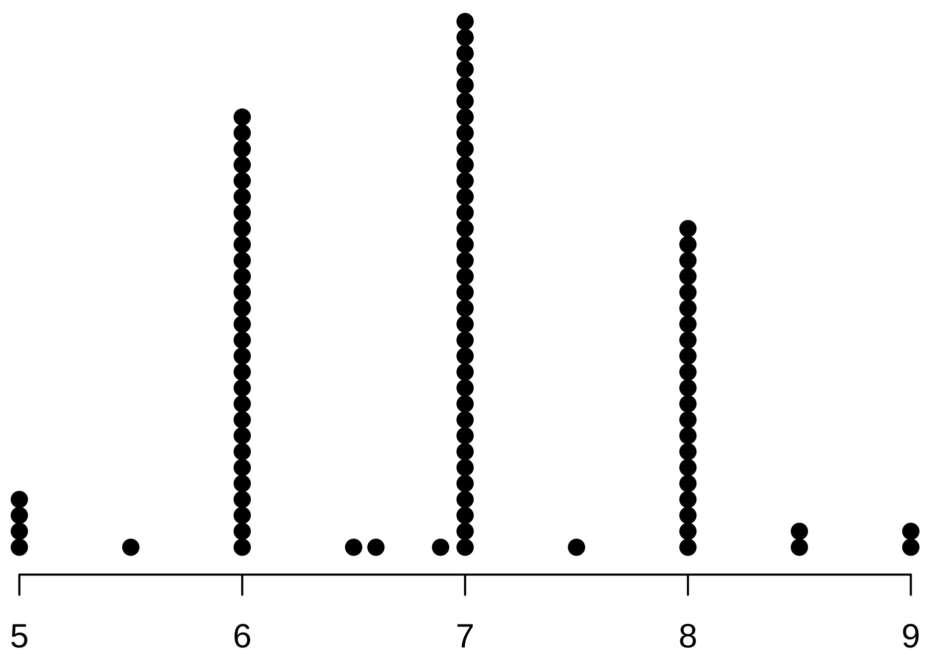
Histogram of milk



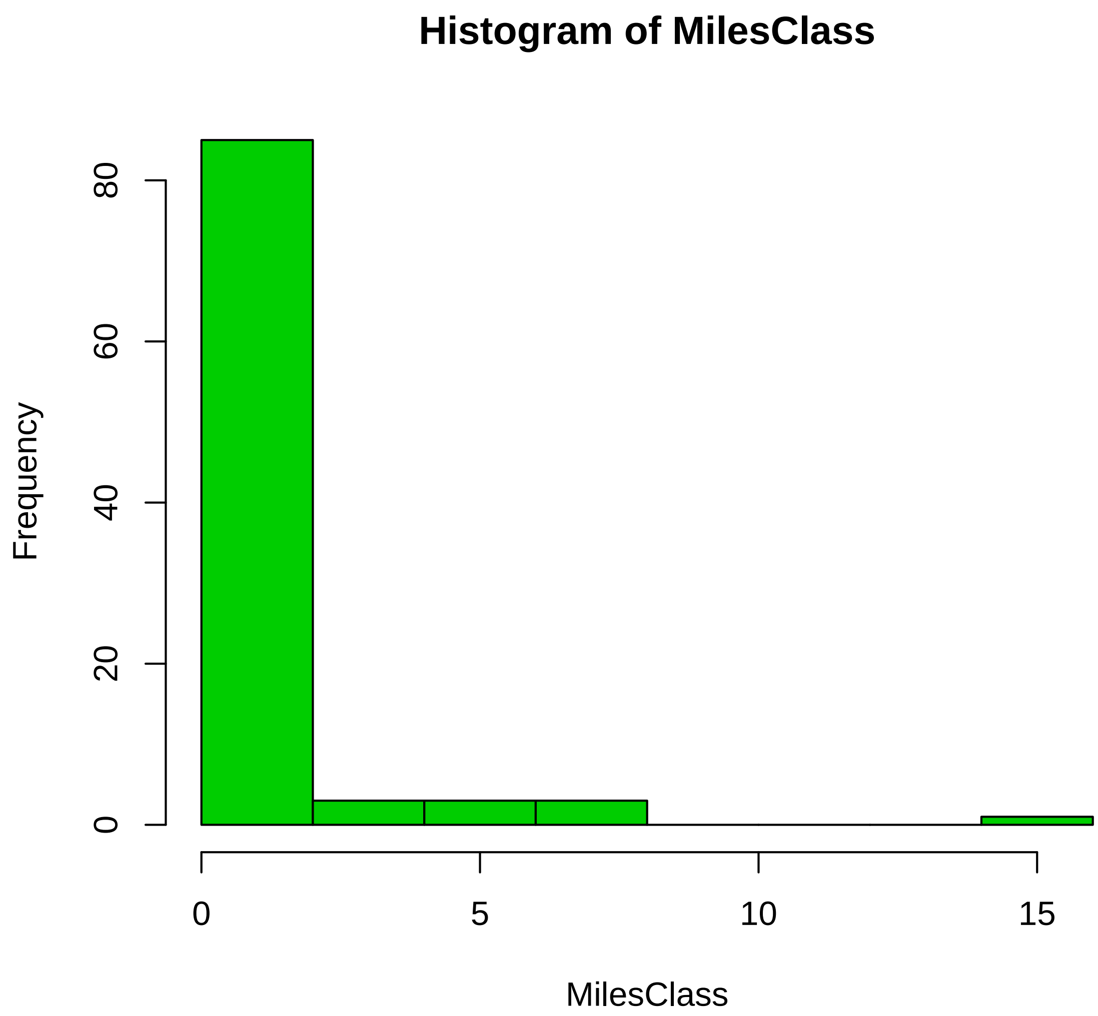
Histogram of milk



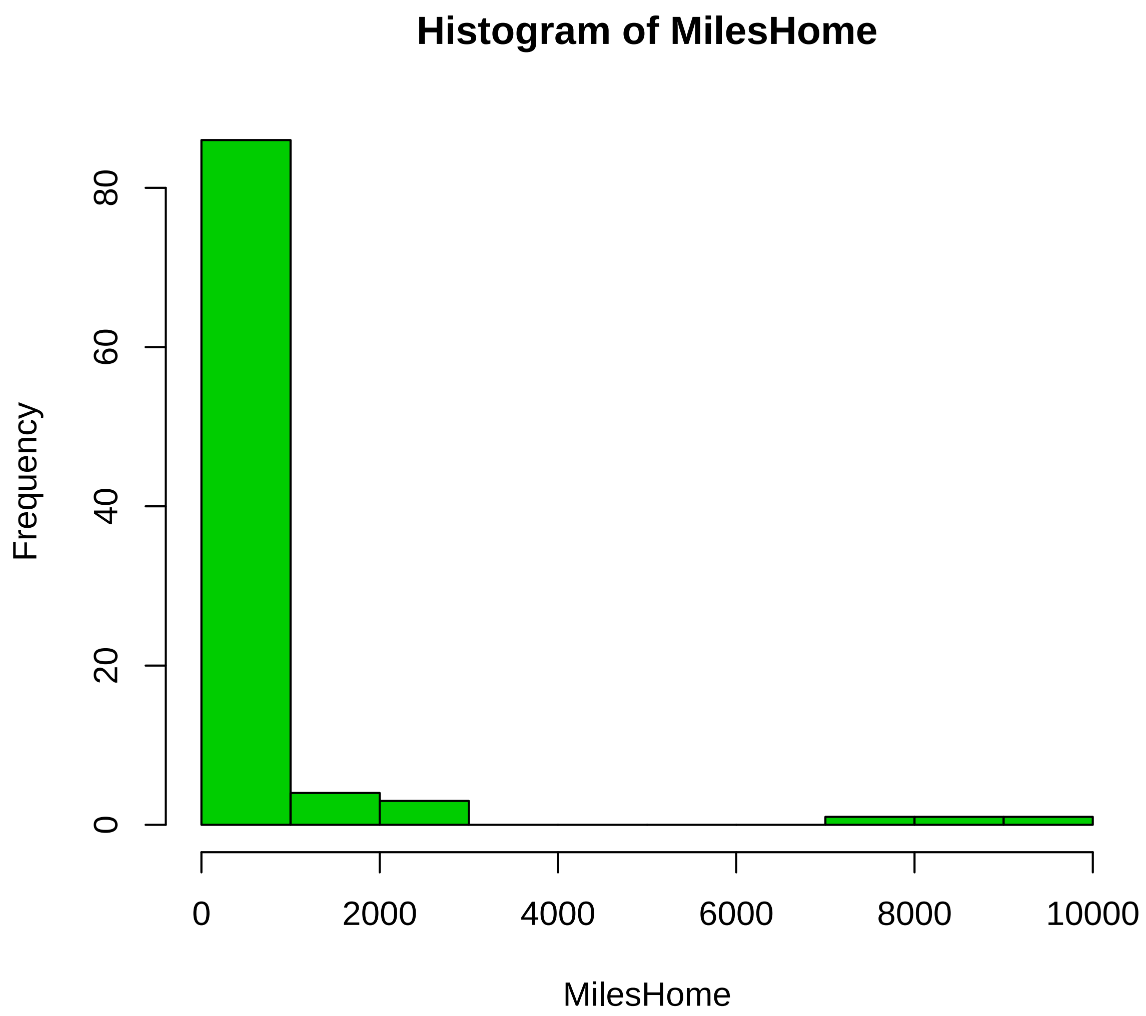
A Dotplot of Hours of Sleep



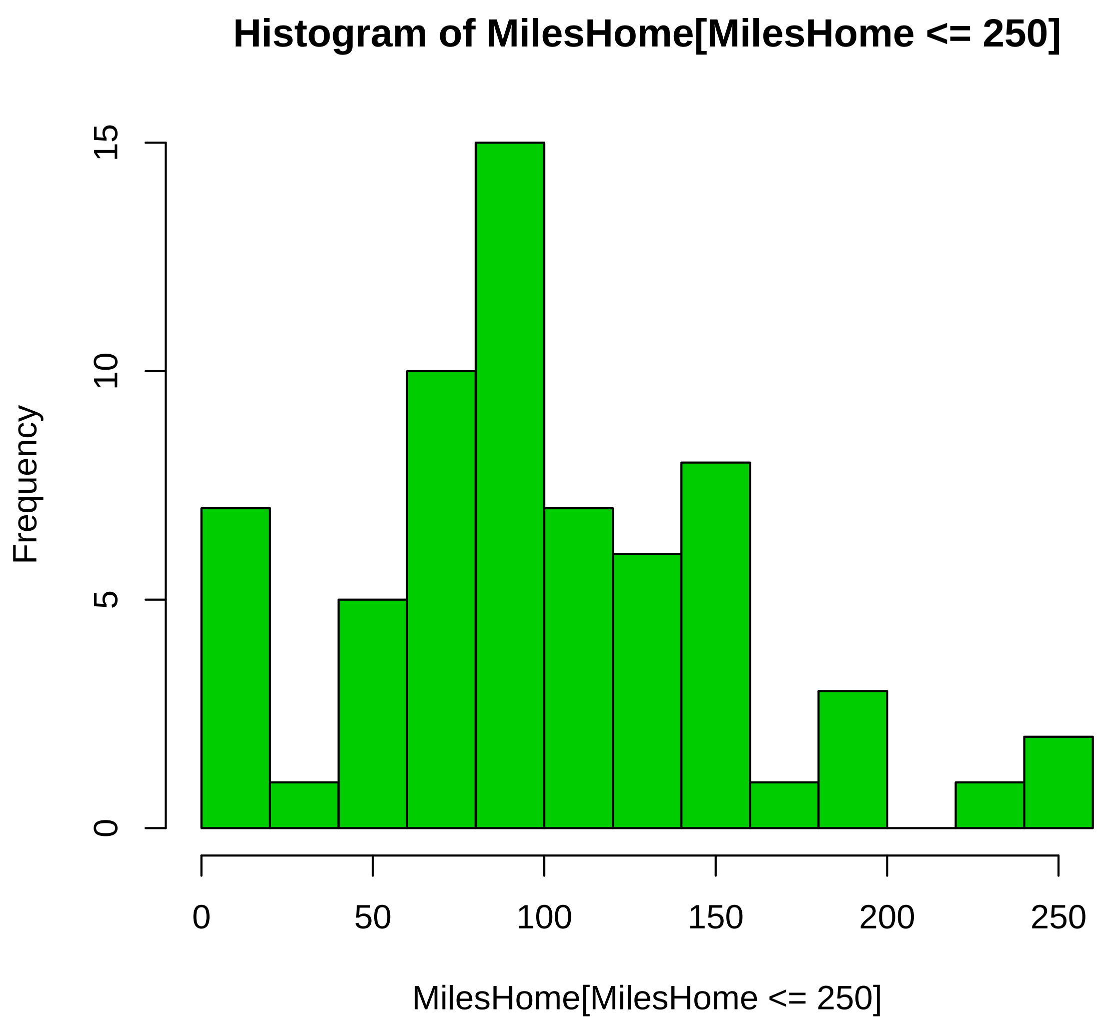
Summary of Miles from MSC



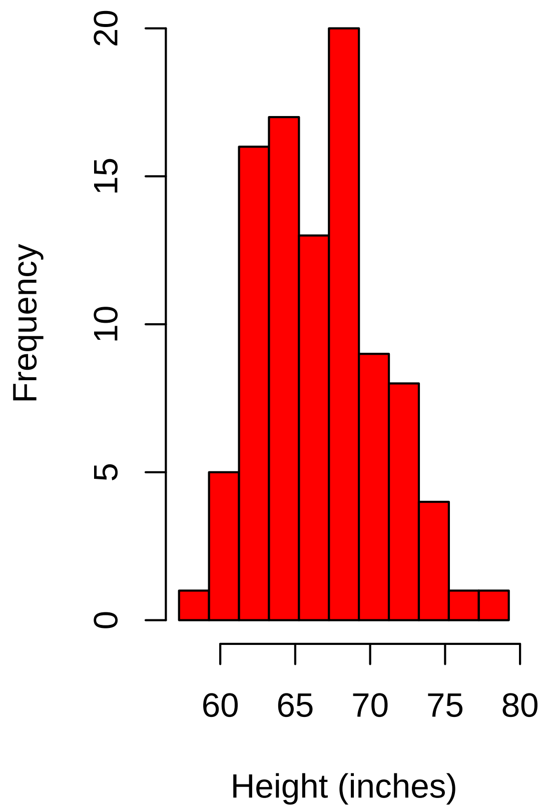
Summary of Miles from Home

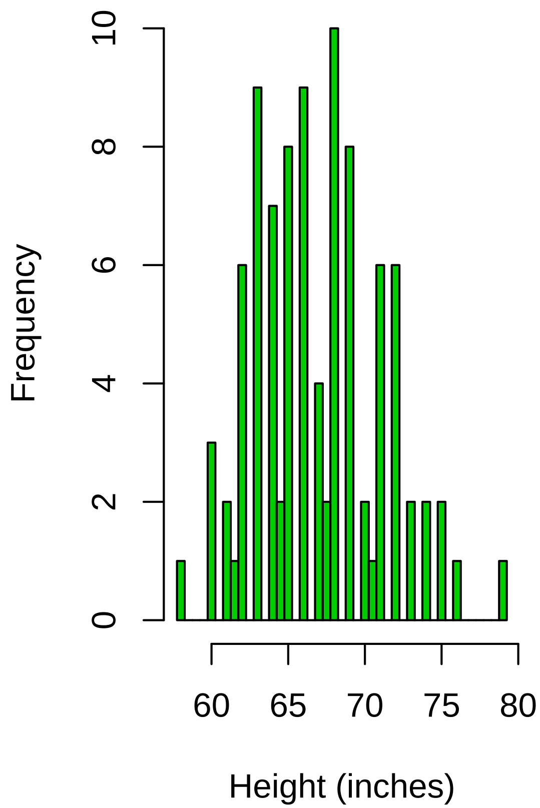


Summary of Miles from Home for Students within 250 miles



Summary of Height





Stem-and-Leaf Diagrams

- Stem-and-Leaf diagrams are useful for showing the shape of the distribution of small data sets without losing any (or much) information.
- Begin by rounding all data to the same precision.
- The last digit is the leaf.
- Anything before the last digit is the stem.
- In a stem-and-leaf diagram, each observation is represented by a single digit to the right of a line.
- Stems are shown only once.
- Show stems to fill gaps!
- Combining or splitting stems can lead to a better picture of the distribution.

Milk Example, Stem-leaf display

Data: 44, 55, 37, 32, 37, 26, 23, 41, 34, 19, 30, 39, 46, 44.

```
> stem(milk)
  The decimal point is 1 digit(s) to the right of the |
  1 | 9
  2 | 36
  3 | 024779
  4 | 1446
  5 | 5
```

- Look for the minimum and maximum, then decide on a precision and **round off** all data at the same precision.
- Last digit: leaf, any digit before: stem. 1 observation=1 leaf
- Leaves may then be ordered.

- Stem-leaf plots on the board

Stem-leaf versus Histograms

- Rotate the histogram: like stem-leaf
- No single histogram. Lots of them! Rules are somewhat arbitrary.
- Histograms are useful with larger datasets, stem-leaf displays with smaller data sets.

Shape of a Distribution, Skewness

- Histograms show several qualitative features of a quantitative variable, such as the number of modes and skewness.
- A distribution is approximately symmetric if the left and right halves are approximately mirror images of each other.
- A distribution is skewed to the right if the right half of the data (the larger values) are more spread out than the left half of the data.
- A distribution is skewed to the left if the left half of the data (the smaller values) are more spread out than the right half of the data.
- It is fairly common for biological data to be skewed to the right. Often times there is a barrier below which there can be no values, but no upper limit.

- Measures of Center (Section 2.4)

Measures of Center

Try to quantify the “center” of “typical value” of the observations in a sample. We consider

1. Mean
2. Median
3. Mode

Mean

Milk yield data: 44, 55, 37, 32, 37, 26, 23, 41, 34, 19, 30, 39, 46, 44.
$y_1 = 44, y_2 = 55, \ldots, y_{14} = 44.$

Sample mean

$$
\begin{aligned}
\bar{y} &= (44 + 55 + \cdots + 44)/14 = (y_1 + y_2 + \cdots + y_{14})/14 \\
&= \frac{1}{n}(y_1 + y_2 + \cdots + y_n) \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i
\end{aligned}
$$

```
> mean(milk)
[1] 36.21429
```

Here $\bar{y} = 36.2$ lbs/day

Median

Median = typical value. Half of observations are below, half are above.

- Sort the data: 19 23 26 30 32 34 37 37 39 41 44 44 46 55

```
> sort(milk)
```

- and find the middle value.If sample size *n* is odd, no problem. If *n* is even, there are 2 middle values. The median is their average.

```
> median(milk)
[1] 37
```

Mode

Mode: most common value.
More interesting for discrete data, with small # possible values and large # observations. Example: # of brothers.

```
0 | 0000000000000000000000000000000
1 | 00000000000000000000000000000000000000000000000000
2 | 000000000000000
3 | 00000
4 |
5 |
6 |
7 | 0
```

Mode = 1 (brother).

Comparing the mean and the median

- Imagine a histogram made of a uniform solid material.
  - The mean is about the point at which the solid would balance.
  - The median is about at a point that would divide the area of the histogram exactly in half.
- The mean and median of a symmetric distribution are the same.
- The median is more resistant to outliers than the mean. For example, the mean and median of the numbers 1, 2, 3 are 2, but for the data set 1, 2, 30, the median is still 2, but the mean is 11, far away from each observation.
- The median can be a better measure of a ‘typical value’ than the mean especially for strongly skewed variables.
- If a variable is skewed to the right, the mean will typically be larger than the median.
- The opposite is true if the variable is skewed to the left.

Examples

Examples: data	median	mean $\bar{y}$
3, 7, 9, 11, 22		10.4
2, 6, 7, 12, 13, 16, 17, 20		11.625
2, 6, 7, 12, 13, 16, 17, 200		

Examples

Examples: data	median	mean $\bar{y}$
3, 7, 9, 11, 22	9	10.4
2, 6, 7, 12, 13, 16, 17, 20		11.625
2, 6, 7, 12, 13, 16, 17, 200		

Examples

Examples: data	median	mean $\bar{y}$
3, 7, 9, 11, 22	9	10.4
2, 6, 7, 12, 13, 16, 17, 20	12.5	11.625
2, 6, 7, 12, 13, 16, 17, 200		

Examples

Examples: data	median	mean $\bar{y}$
3, 7, 9, 11, 22	9	10.4
2, 6, 7, 12, 13, 16, 17, 20	12.5	11.625
2, 6, 7, 12, 13, 16, 17, 200	12.5	34.125

- Boxplots (Section 2.5)

First Step

- Understand quartiles

Quartiles

First quartile Q_1: median of those values below the median
Third quartile Q_3: median of those values above the median

- Note: some authors (software packages) use a slightly different definition for quartiles.

Quartiles - Example

Milk yield:

19 23 26 30 32 34 37 | 37 39 41 44 44 46 55

Quartiles - Example

Milk yield:

19 23 26 30 32 34 37 | 37 39 41 44 44 46 55

Quartiles - Example

Milk yield:

$$19 \quad 23 \quad 26 \quad \underline{30} \quad 32 \quad 34 \quad 37 \;\Big|\; 37 \quad 39 \quad 41 \quad \underline{44} \quad 44 \quad 46 \quad 55$$

Quartiles - Examples for You

p.33 Example 2.20 and Example 2.21

Five-number Summary and Boxplots

- **five-number summary** = minimum, maximum, median, and the quartiles.
- A **boxplot** is a visual representation of the five-number summary

Boxplots

- In a simple boxplot, a box extending from the first to third quartiles represents the middle half of the data. The box is divided at the median, and whiskers extend from each end to the maximum and minimum.
- It is common to draw more sophisticated boxplots in which the whiskers extend to the most extreme observations within upper and lower fences and individual observations outside these fences are labeled with individual points as potential outliers.
- The most common rule defining the fences are that they are 1.5 IQR below the first quartile and 1.5 IQR above the third quartile.

Steps for Making a Boxplot

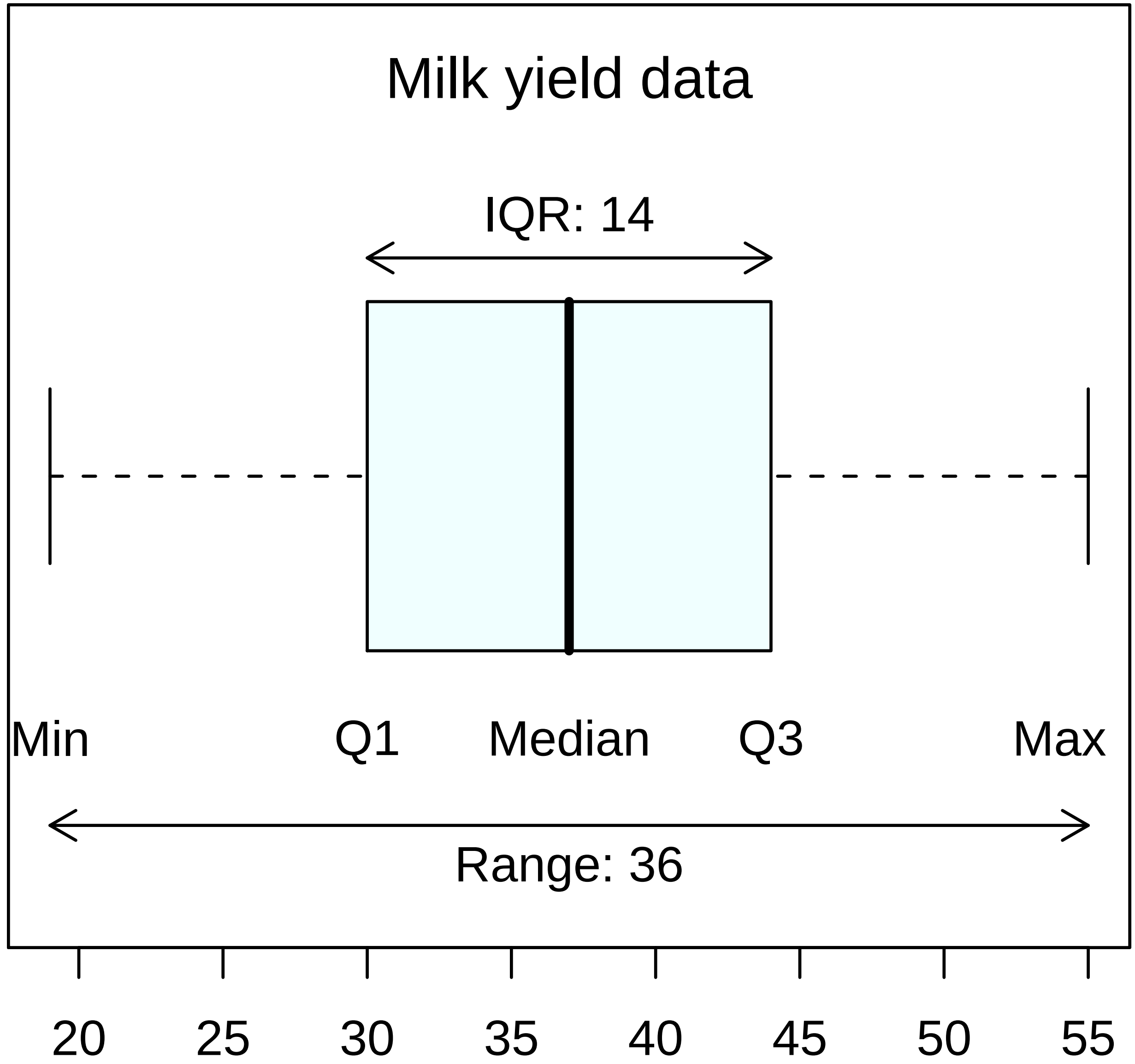
1. Mark the positions for min, Q_1, median, Q_3, max
2. Make a box connecting the quantiles
3. Extend “whiskers” from Q_1 down to the min; and from Q_3 up the max.

Milk Data Example

Chalkboard.

Display: Boxplot

```
> fivenum(milk)
> summary(milk)
> boxplot(milk)
```



Fences

- Sometimes we want to put up a "fence" around our data.
- **lower fence**,

$$\text{lower fence} = Q_1 - 1.5IQR$$

- **upper fence**,

$$\text{upper fence} = Q_3 + 1.5IQR$$

Outliers

- An **outlier** is a data point which differs so much from the rest of the data that it doesn't seem to belong.
- Possible reasons:
  1. typographical error
  2. problem with the experimental protocal (e.g. the lab tech made a mistake)
  3. special circumstances (e.g. an abnormally high value on a medical test might indicate the presence of a disease)

Detecting an Outlier

- An outlier is a data point that falls outside of the fences.
- That is, if

  $$\text{data point} < Q_1 - 1.5IQR$$

  or

  $$\text{data point} > Q_3 + 1.5IQR,$$

  then we call the point an outlier

Display: A Modified Boxplot

- In a boxplot, there are no fences. The whiskers extend to minimum and maximum
- A **modified boxplot** is a boxplot in which any outlier are graphed as separate points.
- A modified boxplot has fences; observations outside fences are drawn as points.
- Modified boxplot - whiskers cannot go beyond fences.
    - Outlier in the upper half of the distribution. Then: extend a whisker from Q_3 up to the largest point that is not an outlier
    - Outlier in the lower half of the distribution. Then: extend a whisker from Q_1 down to the smallest point that is not an outlier
- Often "boxplot" means "modified boxplot". Most software packages draw modified boxplots by default.

Milk Example

- $Q_1 = 30$, $Q_3 = 44$, $IQR = 44 - 30 = 14$
- Fences, $1.5 * 14 = 21$

$$\text{lower fence} = 30 - 21 = 9$$

$$\text{upper fence} = 44 + 21 = 65$$

- Outliers?
  1. smallest data point (min) $= 19 > 9$
  2. largest (max) $= 55 < 65$
  3. no oultier

Example: Female Heights

Height, in females: Min$=$ 53 in, $Q_1 = 64$, median $= 66$, $Q_3 = 68$, max $= 74$. Data: 53, 60, 60, 60.2, 61, ..., 70, 72, 74.

IQR is:
Fences are:
Outliers?

Example: Female Heights

Height, in females: Min$=$ 53 in, $Q_1 = 64$, median $= 66$, $Q_3 = 68$, max $= 74$. Data: 53, 60, 60, 60.2, 61, ..., 70, 72, 74.

IQR is: $68 - 64 = 4$

Fences are:

Outliers?

Example: Female Heights

Height, in females: Min$=$ 53 in, $Q_1 = 64$, median $= 66$, $Q_3 = 68$, max $= 74$. Data: 53, 60, 60, 60.2, 61, ..., 70, 72, 74.

IQR is: $68 - 64 = 4$
Fences are: $64 - 6 = 58$ and $68 + 6 = 74$
Outliers?

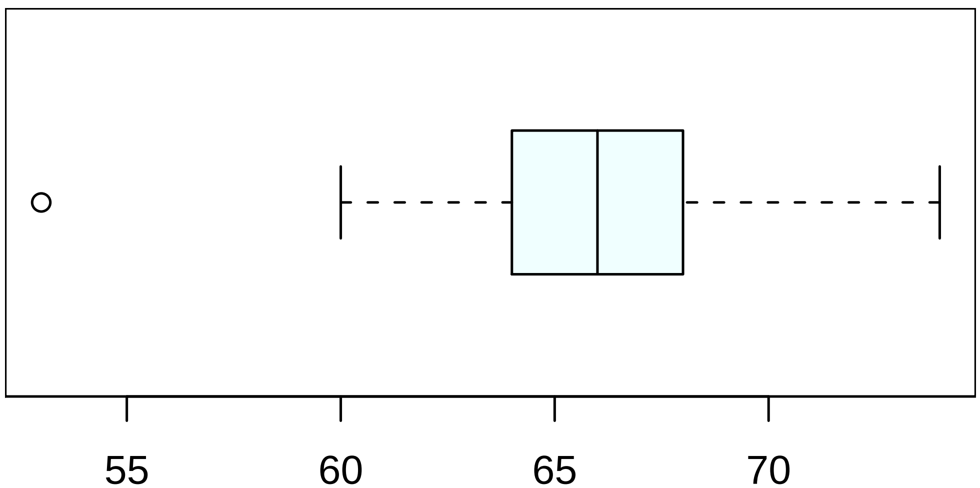
Example: Female Heights

Height, in females: Min$=$ 53 in, $Q_1 = 64$, median $= 66$, $Q_3 = 68$, max $= 74$. Data: 53, 60, 60, 60.2, 61, ..., 70, 72, 74.

IQR is: $68 - 64 = 4$

Fences are: $64 - 6 = 58$ and $68 + 6 = 74$

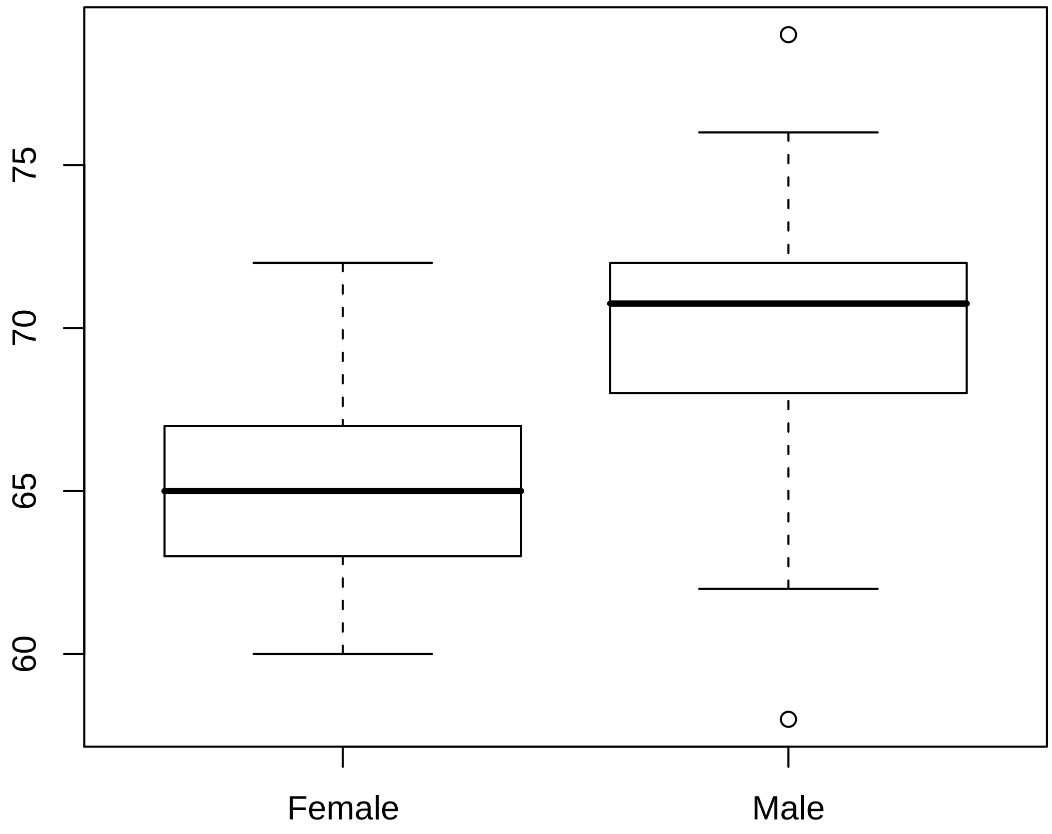
Outliers? 53. The left whisker will extend to 60 only.



A Strength of Boxplots

- Power to give a visual comparison of several distributions

Side-by-side boxplot of height versus sex



Examples for You

- p.36 Example 2.24, 2.25

- Measures of Dispersion (Section 2.6)

Measures of Dispersion

Try to quantify how “spread out” the data is. Consider

1. Range
2. Interquartile range
3. variance
4. standard deviation
5. coefficient of variation

Recall: Milk Data

- Milk data: milk yields (lbs/day) were collected from a particular herd on a given day.
  Data:
  44, 55, 37, 32, 37, 26, 23, 41, 34, 19, 30, 39, 46, 44.
- Sort the data:
  19 23 26 30 32 34 37 37 39 41 44 44 46 55

Range and Interquartile Range

- Range: maximum - minimum
  Milk yield: range is $55 - 19 = 36$
- IQR: Inter Quartile Range $= Q_3 - Q_1$
  Spread in the central “body” of the distribution
  Milk yield: IQR $= 44 - 30 = 14$.

Dispersion as “Deviation from the Mean”

- The variance, standard deviation, and coefficient of variation are all related.
- Based on deviations from the mean.

Deviation

Recall $y_1 =$ first observation, $\ldots$, $y_n =$ last observation.

- A deviation from the mean is the signed distance of an observation from the mean.

$$\text{deviation} = \text{value of observation} - \text{mean}$$

  Observations greater than the mean have positive deviations while those less than the mean have negative deviations.

- Formula (for the i^{th} observation): $y_i - \bar{y}$
- Ex: first cow has deviation $44 - 36.2 = +7.8$,
  cow with data 19 has deviation $19 - 36.2 = -17.2$.

Variance

- Denote s^2
- Formula

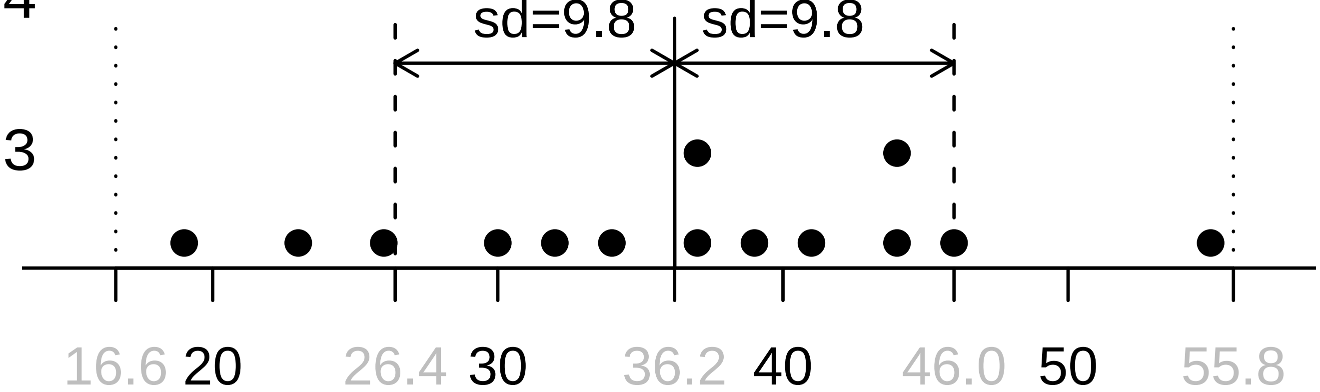
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left((y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2\right)$$

- Note: $s^2 \geq 0$ always!

Standard deviation

- **Standard deviation:** $s = \sqrt{\text{variance}} = \sqrt{s^2}$ is now in original units. s is the typical deviation.
  Here, $s = 9.8$ lbs.

```
> mean(milk)
[1] 95.25824
> sd(milk)
[1] 9.760033
```



The Empirical Rule

For many variables (especially those that are nearly symmetric and bell-shaped), the following empirical rule is often a very good approximation.

- About 68% of the observations are within 1 SD of the mean.
- About 95% of the observations are within 2 SDs of the mean.
- Nearly all observations are within 3 SDs of the mean.

Coefficient of variation

It is the **relative variation** $\text{CV} = \dfrac{s}{\bar{y}}$. It is dimensionless.

Milk data: $\text{CV} = 9.8/36.2 = 0.27$. It means the typical deviation from the mean is about 27% of the mean.

Height in females: $\text{CV} = 3.3 \text{ in } /65.6 \text{ in } = 0.20$.

Examples for You

- p.41 Example 2.28
- p.44 Example 2.32

- Summary

Conclusions

- The first step in a data analysis: exploratory data analysis
- Plot the data and obtain numerical summaries to get a "feel" for your data.