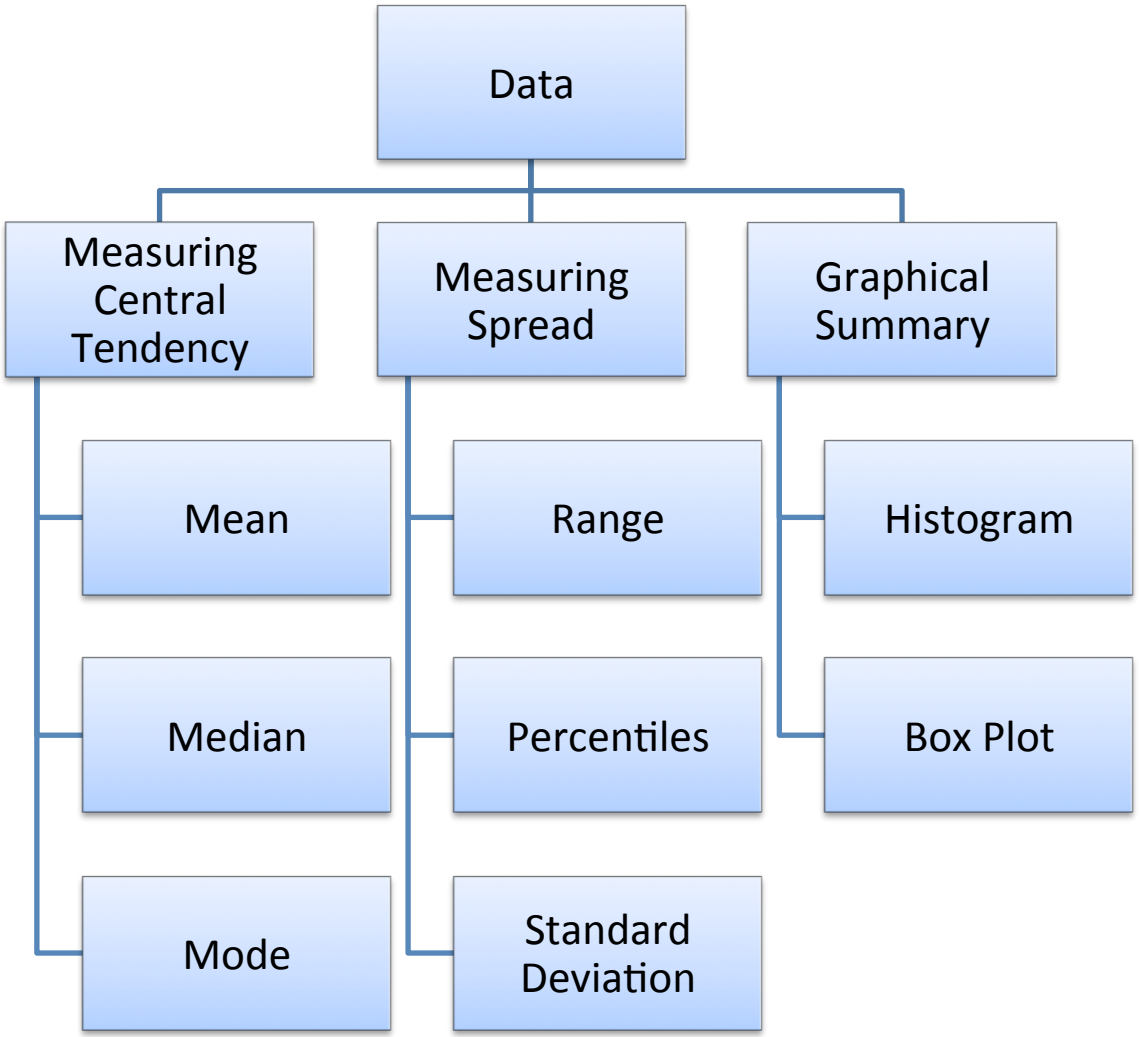
Summarizing Data

Keegan Korthauer

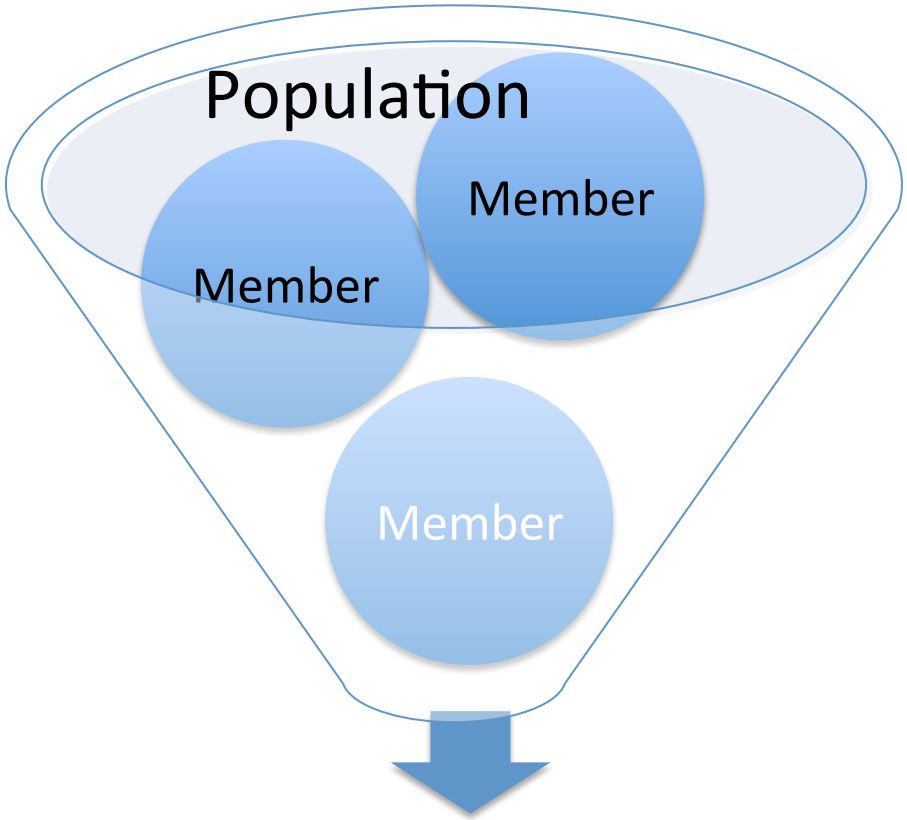
Department of Statistics

UW Madison

Summarizing Data



Algebraic Form of Data



Sample of size n:

$X_1, X_2, X_3, X_4, \ldots, X_n$

MEASURING CENTRAL TENDENCY

Mean

Mode

Median

Mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Sum of all X values divided by the number of elements in the sample
- also known as the average
- Example – a sample of the heights (in inches) of 5 UW Madison students:
  (63.90, 71.45, 68.68, 72.05, 66.27)

  The mean of the sample is :
  (63.90 + 71.45 + 68.68 + 72.05 +66.27)/5 = 68.47

Mode

- Most frequently occurring value(s) in a sample
- There can be more than one mode in a sample
- Examples

  The mode of {1, 2, 3, 3, 4, 5, 6} is 3.

  The modes of {1, 2, 2, 3, 3, 4, 6, 6} are 2, 3 and 6.

Median

A number that divides **sorted data** into exactly **two halves**

ODD n **MIDDLE NUMBER**

EVEN n **M E D + I A N**

$$\frac{D + I}{2}$$

- Example – {3, 2, 4, 6, 7, 5}

  SORT! {2, 3, 4, 5, 6, 7}

  The median is (4+5)/2 = 4.5

- Example – {3, 2, 4, 9, 6, 7, 5}

  SORT! {2, 3, 4, 5, 6, 7, 9}

  The median is 5 (exactly in the middle).

MEASURING SPREAD

Range

Percentiles

Standard deviation

Range

Size of the gap between the smallest and largest values in data

- Example – {2, 3, 4, 6, 1, 9, 10}

  The smallest value is 1 and the largest value is 10.

  So the range of the data is 10 – 1 = 9.

Percentiles

A **position** below which a certain percentage of data lies

X--------X-----X---------X----------X--------X------X---X----X

p^{th} percentile = Data point in the $\frac{p}{100}(n+1)^{th}$ position

* if $\frac{p}{100}(n+1)$ is not an integer, take the average of the two sample values on either side

- Example – Median (50%)
- Example – Quartiles (25%, 50% (median), 75%)

Standard Deviation

- Measures the degree of spread in a sample
- When the spread is large, the sample values will tend to be far from their mean; when the spread is small, they will tend to be close to their mean
- Standard deviation can be viewed as the average deviation from the center/mean
- The square of the standard deviation is called the variance

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)}$$

Equivalent Formulas

Statistic vs. Parameter

- Parameter
  - numerical summary of a **population**
  - usually unknown unless we sample the entire population
- Statistic
  - numerical summary of a **sample**
  - computed from data to estimate a parameter
- Statistic or Parameter?
  - A poll showed that 45% of voters support a certain candidate
  - The average height of UW students is 5'8"

GRAPHICAL SUMMARY

Histogram, box plot

Histogram

- A graphic that gives an idea of the 'shape' of data, indicating regions of concentration and sparseness

- To make: construct a **frequency table**, showing **class intervals** and their corresponding **frequencies**, **relative frequencies** and **densities**

- Plot frequency, relative frequency, or density against class interval

Example - Frequency Table

TABLE 1.2 Particulate matter (PM) emissions (in g/gal) for 62 vehicles driven at high altitude

7.59	6.28	6.07	5.23	5.54	3.46	2.44	3.01	13.63	13.02	23.38	9.24	3.22
2.06	4.04	17.11	12.26	19.91	8.50	7.81	7.18	6.95	18.64	7.10	6.04	5.66
8.86	4.40	3.57	4.35	3.84	2.37	3.81	5.32	5.84	2.89	4.68	1.85	9.14
8.67	9.52	2.68	10.14	9.20	7.31	2.09	6.32	6.53	6.32	2.01	5.91	5.60
5.61	1.50	6.46	5.29	5.64	2.07	1.11	3.32	1.83	7.56			

SORT THE DATA

TABLE 1.4 Frequency table for PM emissions of 62 vehicles driven at high altitude

Class Interval (g/gal)	Frequency	Relative Frequency	Density
1–< 3	12	0.1935	0.0968
3–< 5	11	0.1774	0.0887
5–< 7	18	0.2903	0.1452
7–< 9	9	0.1452	0.0726
9–< 11	5	0.0806	0.0403
11–< 13	1	0.0161	0.0081
13–< 15	2	0.0323	0.0161
15–< 17	0	0.0000	0.0000
17–< 19	2	0.0323	0.0161
19–< 21	1	0.0161	0.0081
21–< 23	0	0.0000	0.0000
23–< 25	1	0.0161	0.0081

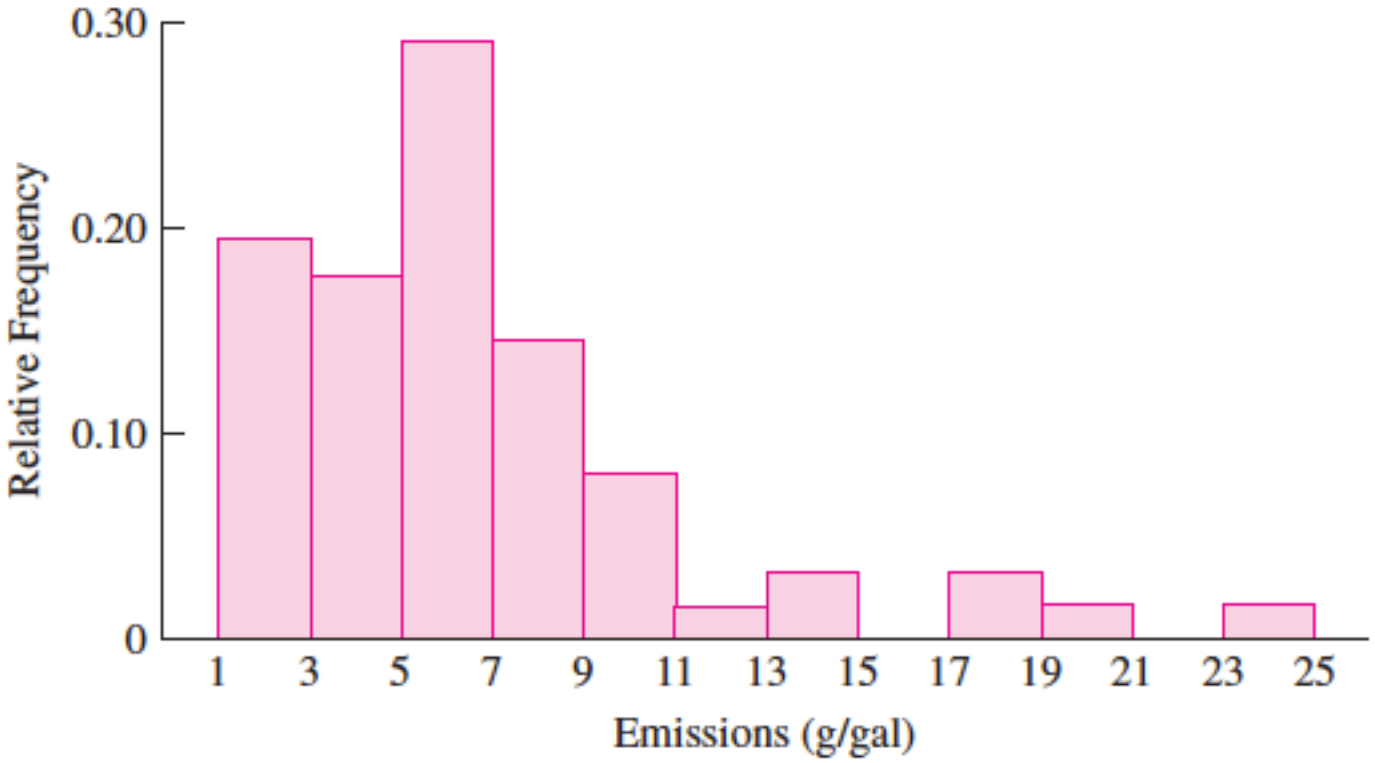
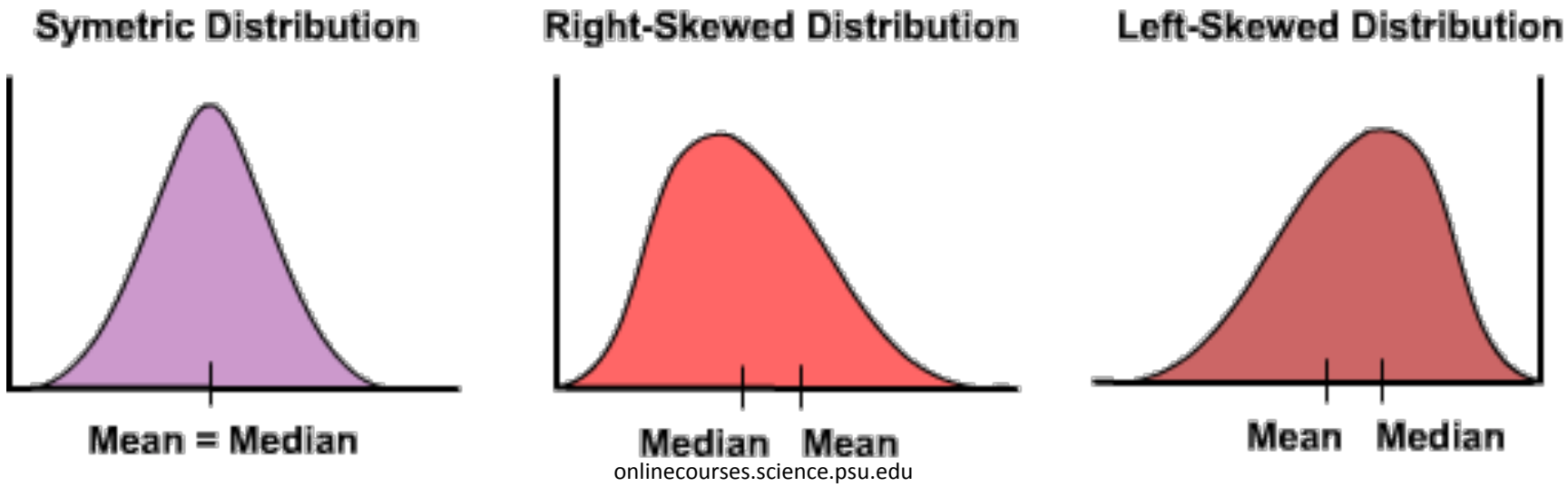
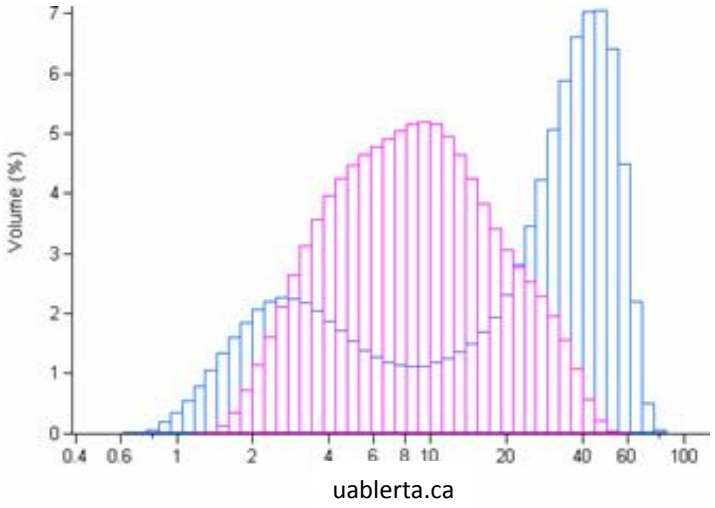


FIGURE 1.8 Histogram for the data in Table 1.4. In this histogram the heights of the rectangles are the relative frequencies. Since the class widths are all the same, the frequencies, relative frequencies, and densities are proportional to one another, so it would have been equally appropriate to set the heights equal to the frequencies or to the densities.

Histograms - Skew and Modality



onlinecourses.science.psu.edu



uablerta.ca

- Skew – tails of the distribution are pulled by extreme values
- Modality – number of prominent peaks
  - one peak: unimodal
  - two peaks: bimodal
  - more than two peaks: multimodal

Histogram example

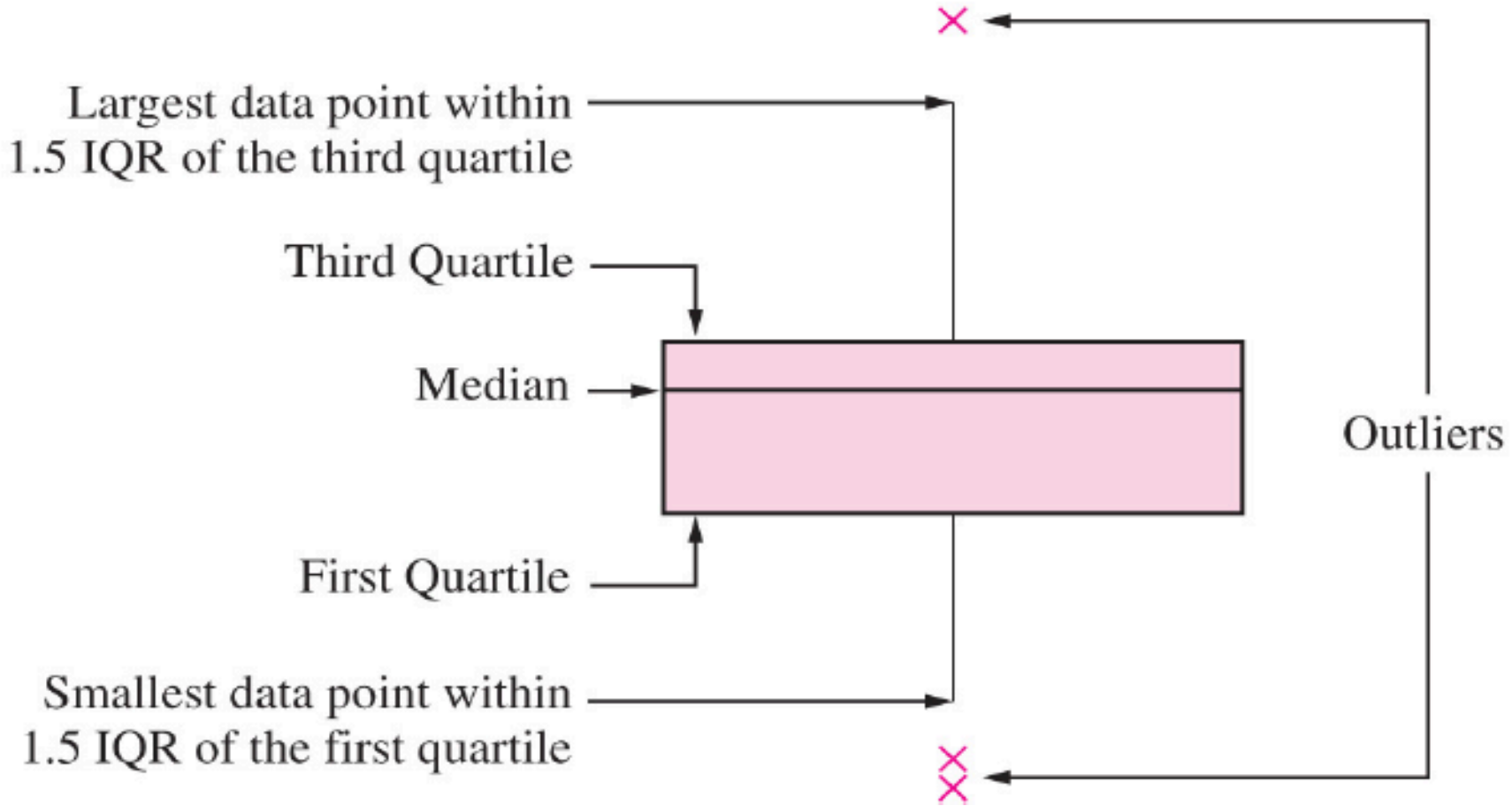
30 Systolic Blood Pressure measurements (already sorted):

92, 94, 97, 99, 105, 108, 108, 109, 111, 114,

115, 115, 119, 122, 125, 127, 127, 127, 128, 128,

128, 129, 129, 130, 132, 135, 138, 140, 141, 150

Box Plot



Example 1.12 – HMA Data

In the article "Evaluation of Low-Temperature Properties of HMA Mixtures", the following values of fracture stress (in megapascals) were measured for a sample of 24 mixtures of hot-mixed asphalt (HMA).

30	75	79	80	80	105	126	138	149
179	179	191	223	232	232	236	240	242
245	247	254	274	384	470			

HMA Example Continued

HMA data:

30 75 79 80 80 105 126 138 149

179 179 191 223 232 232 236 240 242

245 247 254 274 384 470 (outlier)

Q1 = 115.5 Q2 = 207 Q3 = 243.5

IQR = 243.5 – 115.5 = 128 1.5IQR = 192

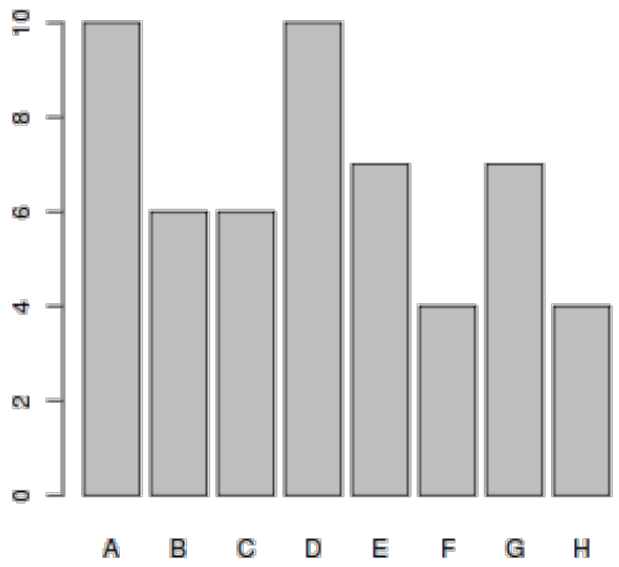
Q3 + 1.5IQR = 435.5 Q1 – 1.5IQR = -76.5

Exercise – HMA Data

- Draw a boxplot
- Draw a histogram

Categorical Data

- Each item assigned a **category** instead of a numerical value
- Summary statistics
  - frequency: how many items are in each category
  - sample proportions: what proportion of the sample is in each category
- Graphical summary: bar chart, pie chart



Next

- Basic probability theory (2.1, 2.2)

- Check Learn@UW for Homework 1 due next Friday 1/31 before lecture