SUMMARIZING DATA: MEASURES OF SPREAD [DEVORE 1.4]

Throughout this page, assume the following discrete numerical sample $x: \ x_1, x_2, \ldots, x_n$

- **RANGE OF A SAMPLE:** Sample **range**, denoted x_R, is the **difference** btw largest & smallest data pt:

$$x_R := x_{(n)} - x_{(1)}$$

$x_{(1)} \equiv$ Smallest Data Point $\qquad$ $x_{(n)} \equiv$ Largest Data Point

---

- **VARIANCE OF A SAMPLE:** Sample **variance**, denoted s^2 or s_x^2, is the following:

$$s^2 := \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2$$

- **STANDARD DEVIATION OF A SAMPLE:** **Standard deviation**, denoted s or s_x, is the square root of variance:

$$s := \sqrt{s^2}$$

- **EASIER FORMULA FOR SAMPLE VARIANCE:**

$$s^2 = \frac{S_{xx}}{n-1} \qquad \text{where} \qquad S_{xx} = \sum_{k=1}^{n} x_k^2 - \frac{1}{n}\left(\sum_{k=1}^{n} x_k\right)^2$$

- **PROPERTIES OF VARIANCE & STD DEV:** Let $c \neq 0$ be a non-zero constant. Then:
  (1) If sample y is defined as follows: $y: \ (x_1 + c), (x_2 + c), \ldots, (x_n + c)$
  Then, $s_y^2 = s_x^2$ and $s_y = s_x$
  (i.e. Uniformly shifting a sample does not change its variance & std dev.)

  (2) If sample z is defined as follows: $z: \ (cx_1), (cx_2), \ldots, (cx_n)$
  Then, $s_z^2 = c^2 s_x^2$ and $s_z = |c| s_x$
  (i.e. Uniformly scaling a sample scales its variance & std dev accordingly.)

- **DEGREES OF FREEDOM:** # of **degrees of freedom** is the # of values that can vary when computing a statistic.
- **INTERQUARTILE RANGE (IQR):** **Interquartile range**, x_{IQR}, is the difference btw the 1^{st} & 3^{rd} quartiles:

$$x_{IQR} := x_{Q3} - x_{Q1}$$

- **INTERHINGE RANGE (IHR):** **Interhinge range**, x_{IHR}, is the difference btw lower & upper hinges:

$$x_{IHR} := x_{UH} - x_{LH}$$

- **MEASURES OF SPREAD & THEIR SENSITIVITY TO OUTLIERS:**
  - The range, x_R, is extremely sensitive to outliers.
  - The variance, s_x^2, is extremely sensitive to outliers.
  - The std dev, s_x, is extremely sensitive to outliers.
  - The interquartile range, x_{IQR}, is almost completely insensitive to outliers.
  - The interhinge range, x_{IHR}, is almost completely insensitive to outliers.

VISUALIZING DATA: BOXPLOTS, COMPARATIVE BOXPLOTS [DEVORE 1.4]

Throughout this page, assume the following discrete numerical sample $x: \; x_1, x_2, \ldots, x_n$

- **CLASSIFYING OUTLIERS:**

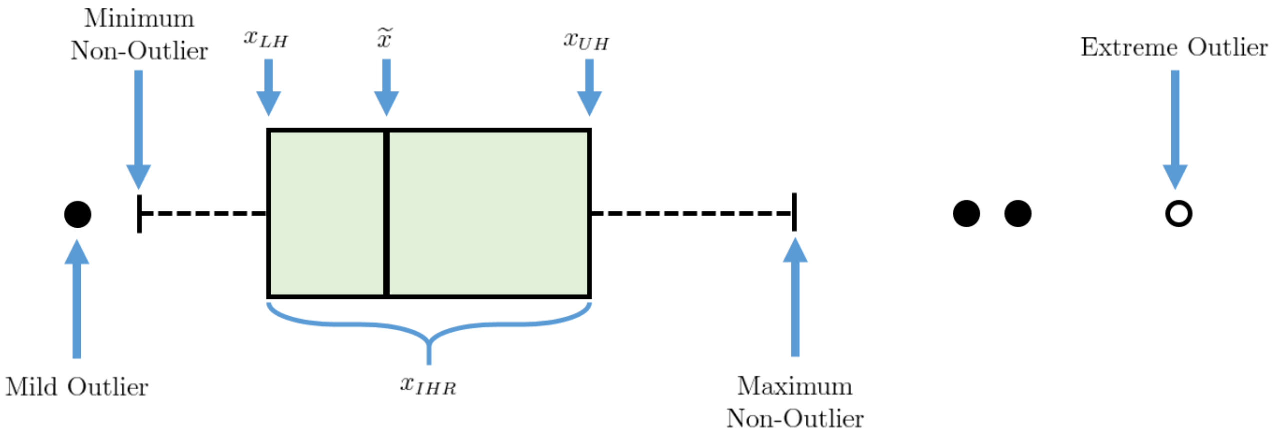
  A data point x_k is an **outlier** if it's farther than $1.5x_{IHR}$ from closest hinge.

  A data point x_k is an **extreme outlier** if it's farther than $3x_{IHR}$ from closest hinge.

  A **mild outlier** is an outlier that's not an extreme outlier.

  $$\text{Outlier: } x_k < x_{LH} - \mathbf{1.5}x_{IHR} \quad \text{OR} \quad x_k > x_{UH} + \mathbf{1.5}x_{IHR}$$
  $$\text{Extreme Outlier: } x_k < x_{LH} - \mathbf{3.0}x_{IHR} \quad \text{OR} \quad x_k > x_{UH} + \mathbf{3.0}x_{IHR}$$

- **BOXPLOTS:** **Boxplots** describe overall skewness, middle 50% skewness, and outliers:



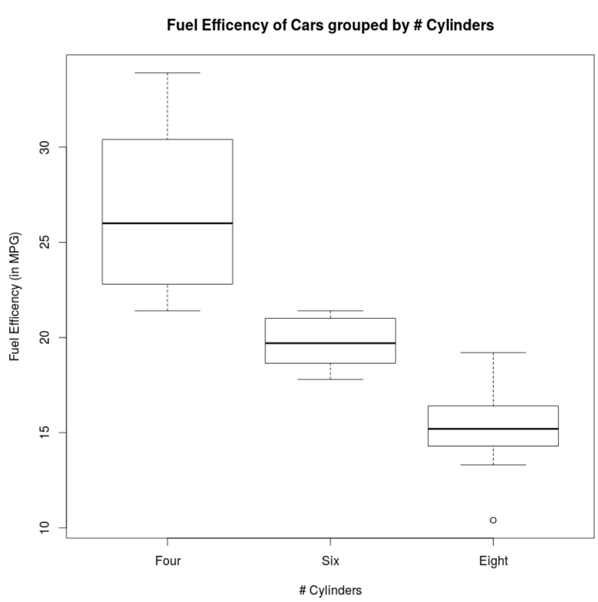
  In the above boxplot:

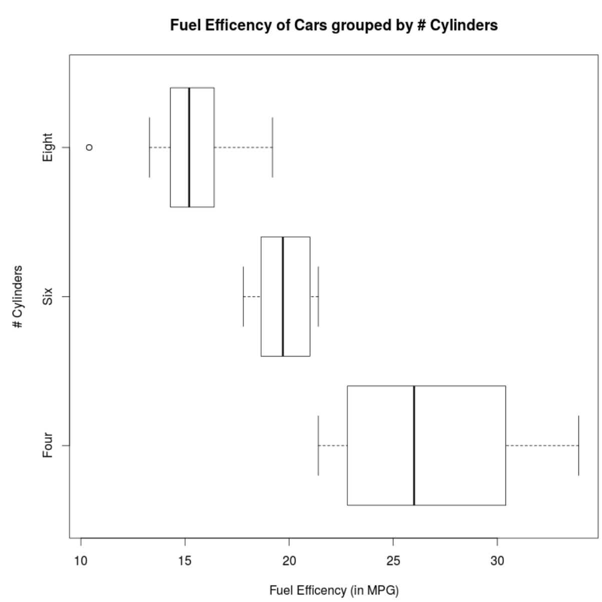
  - The middle 50% of the sample is positively skewed.
    (since median line is closer to left edge of box)
  - The sample is overall positively skewed.
    (since line from upper hinge to max non-outlier is longer than line from lower hinge to min non-outlier)

  WARNING: Software (e.g. MATLAB, R, SPSS, SAS, Minitab) construct boxplots using quartiles instead of hinges.

  WARNING: Software (e.g. MATLAB, R, SPSS, SAS, Minitab) may classify mild & extreme outliers slightly differently.

- **COMPARATIVE BOXPLOTS:** Boxplots are excellent for comparing samples:





EX 1.4.1: Given the following sample of fuel efficiencies of 6-cylinder vehicles (in miles/gallon):

$$x: \ 21.0, \ 15.0, \ 21.0, \ 21.4, \ 18.1, \ 19.2, \ 17.8, \ 19.7, \ 13.0, \ 35.0$$

a) Compute the sample range, x_R.

b) Compute the sample variance, s^2.

c) Compute the sample standard deviation, s.

d) Compute the interquartile range, x_{IQR}.

e) Compute the interhinge range, x_{IHR}.

f) Identify, if any, mild & extreme outliers in the sample.

g) Construct the horizontal boxplot for the sample. (Use hinges, not quartiles!)

h) Use the boxplot to describe the skewness of the sample.