Summarizing Data: Measures of Spread, Boxplots

Engineering Statistics
Section 1.4

Josh Engwer

TTU

29 January 2016

PART I: MEASURES OF SPREAD FOR DISCRETE NUMERICAL DATA

RANGE, VARIANCE, STANDARD DEVIATION
INTERQUARTILE RANGE (IQR), INTERHINGE RANGE (IHR)

The Need for Measures of Spread

Aren't measures of center (mean,median) enough to describe samples? **NO!**

Consider the following three discrete numerical samples:

x : $1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7$
y : $3,\ 4,\ 5$
z : $-1,\ 3,\ 4,\ 5,\ 9$

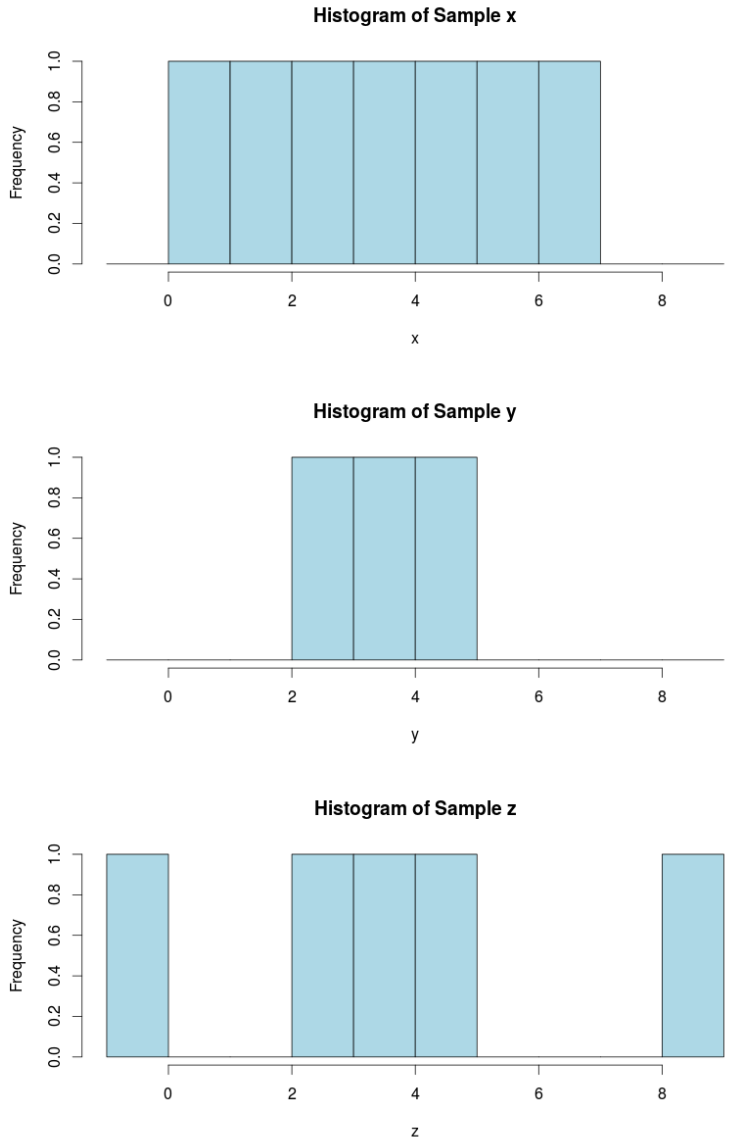
Then their means & medians are:

$\bar{x} = 4, \quad \widetilde{x} = 4$
$\bar{y} = 4, \quad \widetilde{y} = 4$
$\bar{z} = 4, \quad \widetilde{z} = 4$

So, the majority of their measures of center are all equal to each other, yet they're largely non-similar samples!

The Need for Measures of Spread

Aren't measures of center (mean,median) enough to describe samples? **NO!**



Range of a Sample

The **range** is the simplest measure of spread to compute:

Definition

(Range of a Discrete Numerical Sample)

Given a sample with n data points $\quad x: \quad x_1, x_2, \ldots, x_n$
Then its **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

Unfortunately, range conveys very little about the sample.
Therefore, more sophisticated measures of spread are necessary.

Deviations from Mean & Average Deviation from Mean

Consider measuring spread via **deviations from the sample mean**:

Definition

(Deviations from the Mean)

Given a sample with n data points $\quad x: \quad x_1, x_2, \ldots, x_n$

Then its **deviations from the mean** are: $\quad x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$

For this course, it's desired to have a single value for a descriptive statistic.

So it seems sensible to compute the average deviation from the mean:

$$\frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x}) = \frac{1}{n}\sum_{k=1}^{n}x_k - \frac{1}{n}\sum_{k=1}^{n}\bar{x} = \bar{x} - \frac{1}{n}(n\bar{x}) = \bar{x} - \bar{x} = 0 \qquad (!!!!)$$

Mean Absolute Deviation from the Mean

Definition

(Mean Absolute Deviation from the Mean)

Given a sample with n data points $\quad x: \quad x_1, x_2, \ldots, x_n$
Then its **mean absolute deviation from the mean**, denoted $\overline{x}_{MAD(\overline{x})}$, is:

$$\overline{x}_{MAD(\overline{x})} := \frac{1}{n} \sum_{k=1}^{n} |x_k - \overline{x}|$$

Unfortunately, later in the course, many of these statistics encountered so far must be computed using integrals, and integrating functions involving absolute values becomes quite tedious!!

Fortunately, absolute values relate to square roots of squares: $\quad |x| = \sqrt{x^2}$

Variance & Standard Deviation of a Sample

So, instead of using absolute deviations, use **squared deviations** from mean:

Definition

(Variance of a Discrete Numerical Sample)

Given a sample with n data points $\quad x: \quad x_1, x_2, \ldots, x_n$
Then its **variance**, denoted s^2 or s_x^2, is the following:

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

Definition

(Standard Deviation of a Discrete Numerical Sample)

Given a sample with n data points $\quad x: \quad x_1, x_2, \ldots, x_n$
Then its **standard deviation**, denoted s or s_x, is the square root of variance:

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

Why divide by $(n-1)$ instead of n for Variance???

Definition

The # of **degrees of freedom** is the # of values that can freely vary when computing a statistic.

Because a sample with n data points has $(n-1)$ **degrees of freedom**...

...and this follows from the average of deviations from the mean being zero:

$$\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})=0 \implies \sum_{k=1}^{n}(x_k-\bar{x})=0 \implies \sum_{k=1}^{n}d_k=0 \implies d_1+d_2+\cdots+d_n=0$$

where $d_k = x_k - \bar{x}$.

Now, once one freely chooses values for $d_1, d_2, \ldots, d_{n-1}$,

the equation $d_1 + d_2 + \cdots d_{n-1} + d_n = 0$ forces the value for d_n.

(since RHS of equation is a constant)

A stronger justification for this will be encountered in Ch6.

Easier Formula to Compute Variance

Computing $\sum_{k=1}^{n}(x_k - \bar{x})^2$ is tedious! (and requires finding the mean first!)

Fortunately, there's an easier way to compute s^2 without needing the mean $\bar{x}$:

Proposition

(Variance of a Discrete Numerical Sample)

Given a sample with n data points $\quad x: \quad x_1, x_2, \ldots, x_n$

Then its ***variance****, denoted s^2 or s_x^2, can be easily computed as follows:*

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

S_{xx} is called the ***sum of squared deviations from the mean****.*

PROOF:

$$\begin{aligned} S_{xx} &:= \textstyle\sum_{k=1}^{n}(x_k-\bar{x})^2 \overset{FOIL}{=} \sum_{k=1}^{n}(x_k^2 - 2\bar{x}x_k + \bar{x}^2) = \sum_{k=1}^{n} x_k^2 - 2\bar{x}\sum_{k=1}^{n} x_k + \sum_{k=1}^{n}(\bar{x})^2 \\ &= \textstyle\sum_{k=1}^{n} x_k^2 - 2\bar{x}(n\bar{x}) + n(\bar{x})^2 = \sum_{k=1}^{n} x_k^2 - 2n(\bar{x})^2 + n(\bar{x})^2 = \sum_{k=1}^{n} x_k^2 - n(\bar{x})^2 \\ &= \textstyle\sum_{k=1}^{n} x_k^2 - n\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)^2 = \sum x_k^2 - n\left[\frac{1}{n^2}\left(\sum x_k\right)^2\right] = \sum x_k^2 - \frac{1}{n}\left(\sum x_k\right)^2 \end{aligned}$$

□

Properties of Sample Variance & Standard Deviation

Proposition

(Properties of Sample Variance & Standard Deviation)

Let $c \neq 0$ be a non-zero constant.
Given a sample with n data points $x: \quad x_1, x_2, \ldots, x_n$ *Then:*

(1) If sample y is defined as follows: $y: \quad (x_1 + c), (x_2 + c), \ldots, (x_n + c)$
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: $z: \quad (cx_1), (cx_2), \ldots, (cx_n)$
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.)

PROOF:

$$\bar{y} := \frac{1}{n}\sum_{k=1}^{n} y_k = \frac{1}{n}\sum_{k=1}^{n}(x_k + c) = \frac{1}{n}\sum_{k=1}^{n} x_k + \frac{1}{n}\sum_{k=1}^{n} c = \bar{x} + \frac{1}{n}(nc) = \bar{x} + c$$

(1) $$s_y^2 := \frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2 = \frac{1}{n-1}\sum_{k=1}^{n}[(x_k + c) - (\bar{x} + c)]^2$$
$$= \frac{1}{n-1}\sum_{k=1}^{n}[(x_k - \bar{x}) + (c - c)]^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2 := s_x^2$$

Properties of Sample Variance & Standard Deviation

Proposition

(Properties of Sample Variance & Standard Deviation)

Let $c \neq 0$ *be a non-zero constant.*
Given a sample with n *data points* $x: \quad x_1, x_2, \ldots, x_n$ *Then:*

(1) If sample y *is defined as follows:* $y: \quad (x_1 + c), (x_2 + c), \ldots, (x_n + c)$
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:* $z: \quad (cx_1), (cx_2), \ldots, (cx_n)$
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.)

PROOF:

$$\bar{z} := \tfrac{1}{n}\sum_{k=1}^{n} z_k = \tfrac{1}{n}\sum_{k=1}^{n}(cx_k) = c\left(\tfrac{1}{n}\sum_{k=1}^{n} x_k\right) = c\bar{x}$$

(2) $$s_z^2 := \tfrac{1}{n-1}\sum_{k=1}^{n}(z_k - \bar{z})^2 = \tfrac{1}{n-1}\sum_{k=1}^{n}[(cx_k) - (c\bar{x})]^2$$
$$= \tfrac{1}{n-1}\sum_{k=1}^{n}[c(x_k - \bar{x})]^2 = c^2\left[\tfrac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2\right] := c^2 s_x^2 \qquad \square$$

Variance & Standard Deviation of a Population

The variance and standard deviation can be computed for populations.

Definition

(Notation for Variance & Standard Deviation of a Population)

The **population variance** is denoted by σ^2. ("sigma squared")
The **population standard deviation** is denoted by σ. ("sigma")

REMARKS:

σ is the lower-case Greek letter sigma.

Computing σ^2 & σ for finite populations is not practical due to their enormity.

Computing σ^2 & σ for infinite populations will be encountered in Chapter 4.

Interquartile Range (IQR) & Interhinge Range (IHR)

Fortunately, there are some useful measures of spread that are "ranges":

Definition

(Interquartile Range of a Discrete Numerical Sample)

Given a sample with n data points $\quad x: \quad x_1, x_2, \ldots, x_n$
Then its **interquartile range**, x_{IQR}, is the difference btw the 1^{st} & 3^{rd} quartiles:

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

Definition

(Interhinge Range of a Discrete Numerical Sample)

Given a sample with n data points $\quad x: \quad x_1, x_2, \ldots, x_n$
Then its **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 sample range, variance, std dev, IQR and IHR are examples of **statistics**.
The sample range, variance, std dev, IQR & IHR are all **measures of spread**.
This means they indicate how the sample is ”spread out” in unique ways.

A question that's important for many situations is:

HOW SENSITIVE ARE THESE MEASURES OF SPREAD TO **OUTLIERS**??

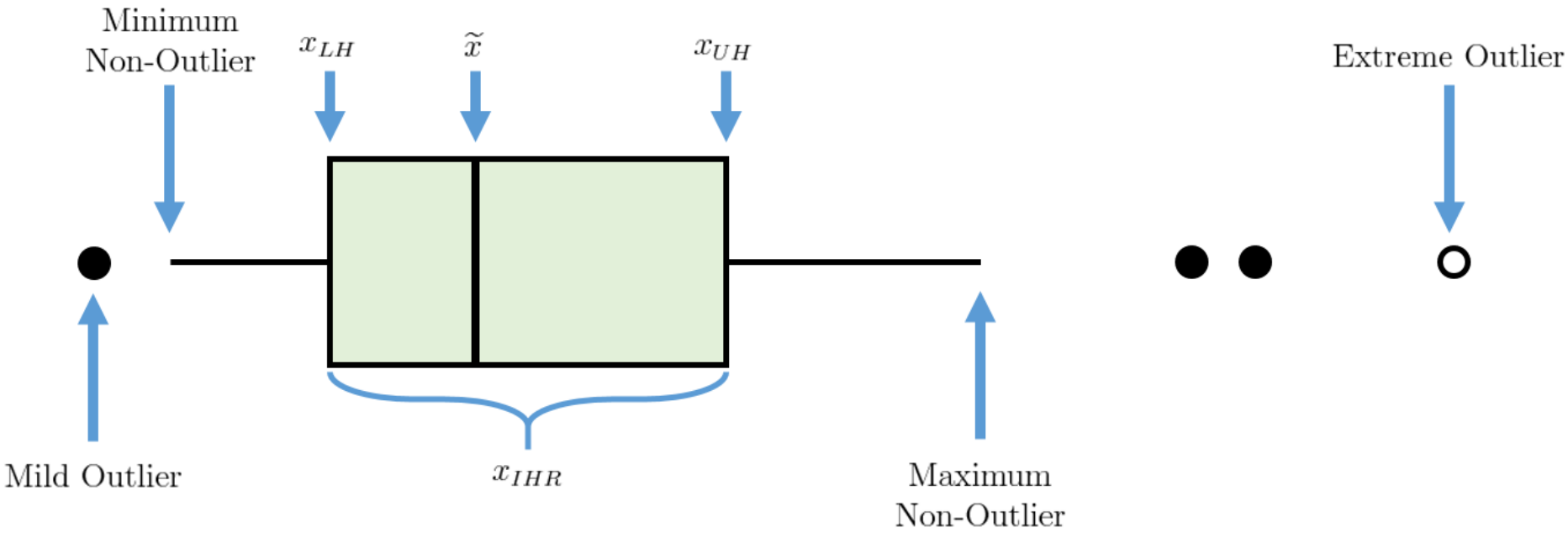
- The range, x_R, is extremely sensitive to outliers.
- The variance, s_x^2, is extremely sensitive to outliers.
- The std dev, s_x, is extremely sensitive to outliers.
- The interquartile range, x_{IQR}, is almost completely insensitive to outliers.
- The interhinge range, x_{IHR}, is almost completely insensitive to outliers.

PART II: VISUALIZATION OF DATA (REVISITED)

BOXPLOTS, COMPARATIVE BOXPLOTS
CLASSIFICATION OF OUTLIERS

Boxplots

Boxplots describe overall skewness, middle 50% skewness, and outliers:

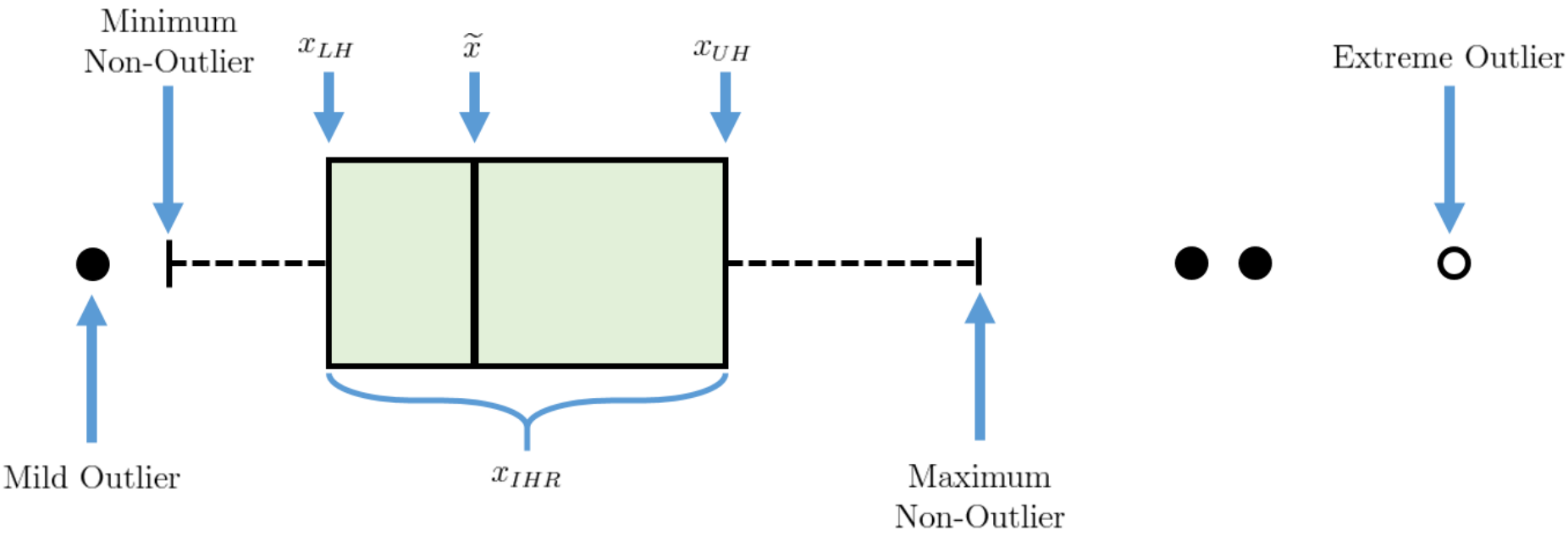


In the above boxplot:

- 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)

Boxplots

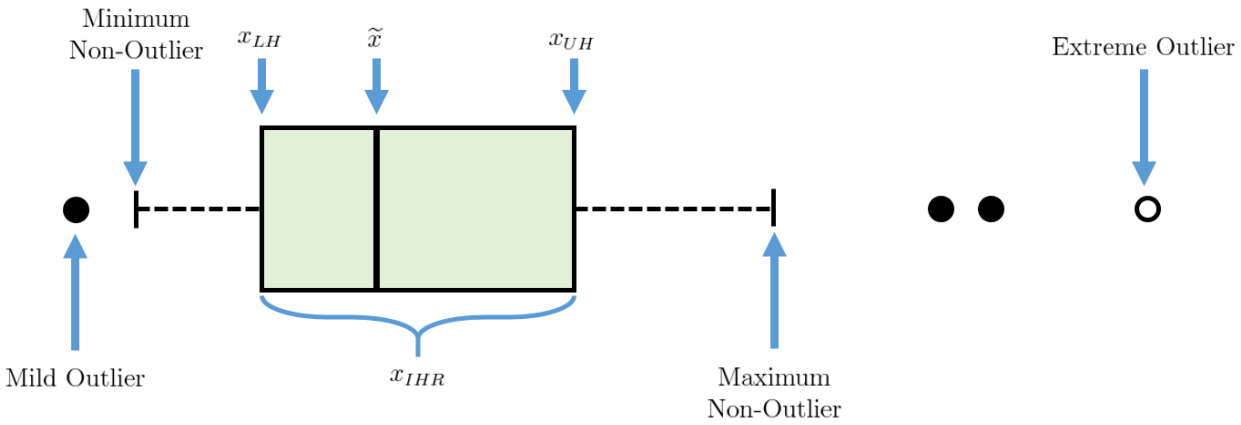
Boxplots describe overall skewness, middle 50% skewness, and outliers:



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)

Numerical Classification of Outliers



Definition

(Numerical Classification of Outliers)

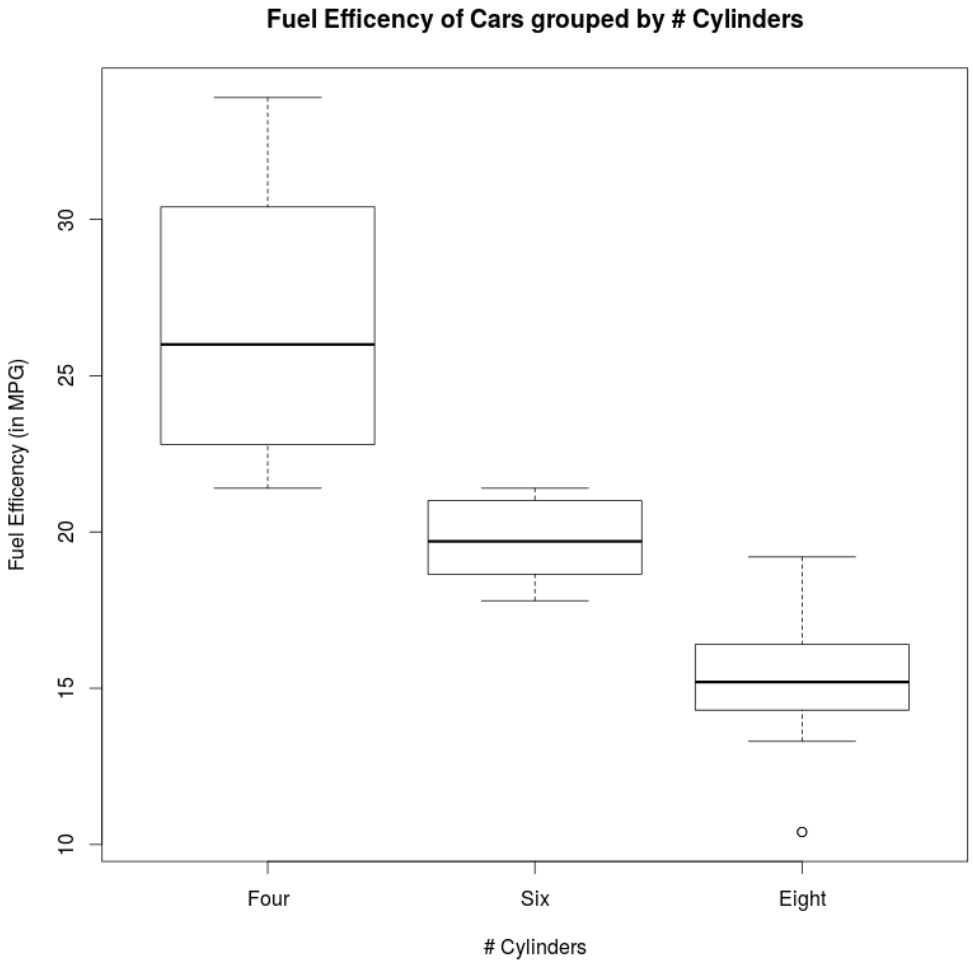
Given a sample with n data points $\quad x: \quad x_1, x_2, \ldots, x_n$

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

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

Comparative Boxplots

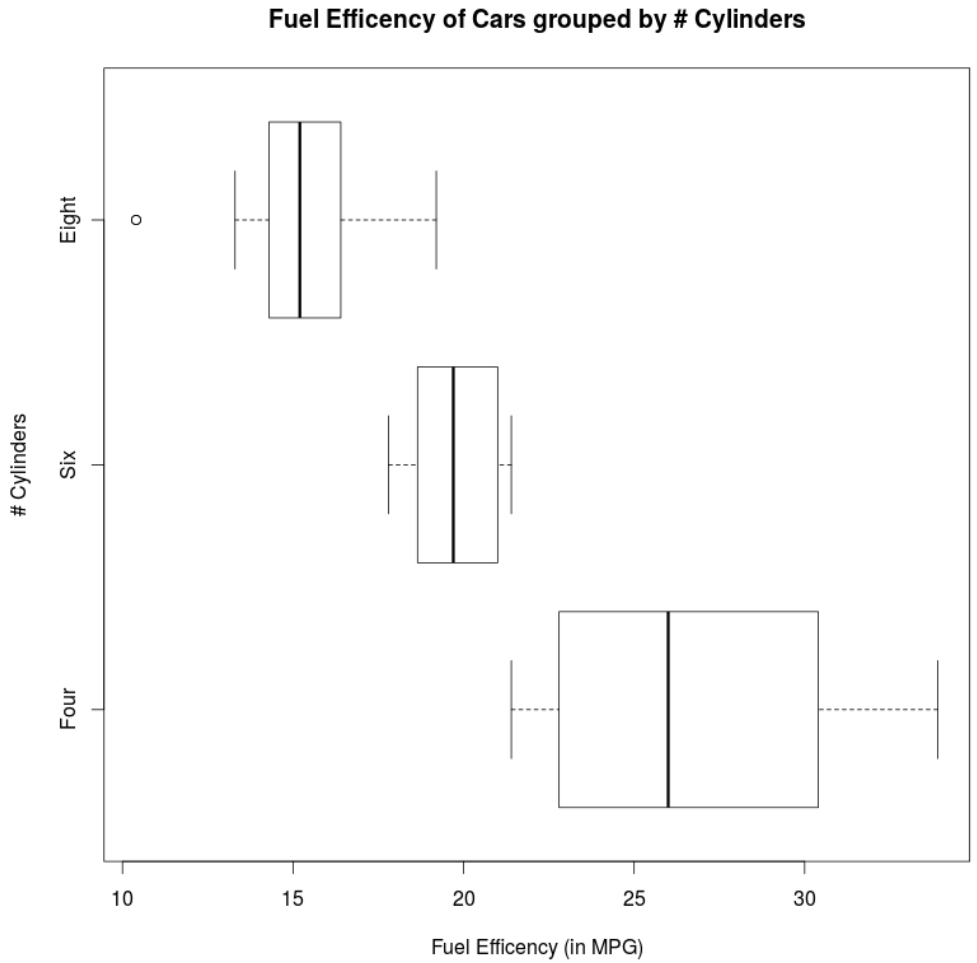
Boxplots are excellent for comparing samples:



Source of above samples: `mtcars` dataset from R.

Comparative Boxplots

Boxplots are excellent for comparing samples:



Source of above samples: `mtcars` dataset from R.

WARNING about Constructing Boxplots via Software

Most statistics packages (e.g. MATLAB, R, SPSS, SAS, Minitab, ...) construct boxplots using quartiles instead of hinges!

Also, these stats packages may define the cutoffs for mild & extreme outliers slightly differently.

Textbook Logistics for Section 1.4

- Difference(s) in Terminology:

TEXTBOOK TERMINOLOGY	**SLIDES/OUTLINE TERMINOLOGY**
Measures of **Variability**	Measures of **Spread**
Measures of **Dispersion**	Measures of **Spread**
Measures of **Scale**	Measures of **Spread**
Lower Fourth	Lower Hinge
Upper Fourth	Upper Hinge
Fourth Spread	Interhinge Range
Box-and-whiskers Plot	Boxplot

- Difference(s) in Notation:

CONCEPT	**TEXTBOOK NOTATION**	**SLIDES/OUTLINE NOTATION**
Interhinge Range	f_s	x_{IHR}

Textbook Logistics for Section 1.4

- Ignore the formula for variance of a finite population: (pg 38)

$$\sigma^2 := \frac{1}{N}\sum_{k=1}^{N}(x_k - \mu)^2$$

  - This will be encountered in Ch6.

- Ignore the brief mention of **standard error** (`SE Mean`) (pg 41)
  - Standard errors are a crucial piece for Inference.
  - Hence, standard error will be encountered in Ch6.

Fin

Fin.