

1.2 Describing Distribution with Numbers

I. The mean

If the n observations are x_1, x_2, \dots, x_n , their mean is

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum x_i$$

II. The median

The median M is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger.

Arrange all the observations in order such that:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

Then

$$M = x_{\left(\frac{n+1}{2}\right)} = \begin{cases} x_{(m+1)}, & \text{if } n = 2m + 1, \text{ odd} \\ x_{\left(m+\frac{1}{2}\right)} = \frac{x_{(m)} + x_{(m+1)}}{2}, & \text{if } n = 2m, \text{ even} \end{cases}$$

Section 1.2

1

Table 1.8 Fuel economy (miles per gallon) for model year 2001 cars

Minicompact cars			Two-seater cars		
Model	City	Highway	Model	City	Highway
Audi TT Coupe	22	31	Acura NSX	17	24
BMW 325Ci Convertible	19	27	Audi TT Roadster	22	30
BMW 330Ci Convertible	20	28	BMW Z3 Coupe	21	28
BMW M3 convertible	20	28	BMW Z3 Roadster	20	27
Jaguar XK8 Convertible	17	24	BMW Z8	13	21
Jaguar XKR Convertible	16	22	Chevrolet Corvette	18	26
Mercedes-Benz CLK320	20	28	Dodge Viper	11	21
Mercedes-Benz CLK430	18	24	Ferrari Modena	11	16
Mitsubishi Eclipse	22	30	Ferrari Maranello	8	13
Porsche 011 Carrera	17	25	Honda Insight	61	68
Porsche 911 Turbo	15	22	Honda S2000	20	26
			Lamborghini Diablo	10	13
			Mazda Miata	22	28
			Mercedes-Benz SL500	16	23
			Mercedes-Benz SL600	13	19
			Mercedes-Benz SLK320	21	27
			Plymouth Prowler	17	23
			Porsche Boxster	19	27
			Toyota MR2	25	30

Section 1.2

2

III. Mean vs. Median

- Mean is not a resistant measure of center, but median is.
- Two strategies to deal with outliers
 - Investigate causes
 - Correct them for wrong record
 - Delete them for good reason
 - Give them special attention
 - Use resistant methods, like median instead of mean for center measurement

IV. The quartiles

- 1). P_{th} percentile = the value such that p percent of the observations fall at or below it.
 - First quartile Q_1 : the median of the observations whose position in the ordered list is to the left of the location of the overall median.
 - Third quartile Q_3 : the median of the observations whose position in the ordered list is to the right of the location of the overall median.

- 2). The five-number summary:

Minimum, Q_1 , M (median), Q_3 , Maximum

- 3). The interquartile range (IQR)

$$IQR = Q_3 - Q_1$$

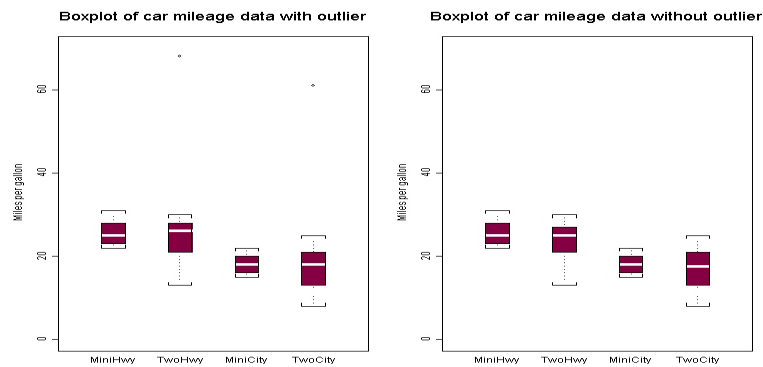
- The $1.5 \cdot IQR$ criterion for outliers:

An observation x_i is a suspected outlier if

$$X_i < Q_1 - 1.5 \cdot IQR \text{ or } X_i > Q_3 + 1.5 \cdot IQR$$

- 4). The boxplots

- A boxplot is a graph of the five-number summary.
 - A central box spans the quartiles Q_1 and Q_3 .
 - A line in the box marks the median M .
 - Lines extend from the box out to the smallest and largest observations.
- Modified boxplot (That is the boxplot we use often)



Section 1.2

5

V. The standard deviation

1). The variance s^2 of a set of observations is the average of the squares of the deviations of the observations from their mean. The variance of n observations x_1, x_2, \dots, x_n is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

2). The standard deviation s is the square root of the variance s^2 :

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

3). Properties

- Using s to measure spread (about) mean only when using mean to measure center.
 - Center --- Spread
 - Mean --- Deviation
 - Median --- Quartiles
- $S=0$ means $x_1 = x_2 = \dots = x_n$. No spread.
- S is not resistant. I.e a few outliers can make s very large.

Section 1.2

6

Summary of a distribution of a quantitative variable

Graphs	<ul style="list-style-type: none"> ▪Stemplots(small data set) ▪Histogram(large data set) <ul style="list-style-type: none"> •Providing a clear display of a single distribution (back-to-back stemplots can compare two distributions) 	<ul style="list-style-type: none"> ▪Boxplots <ul style="list-style-type: none"> •Comparing several distributions
Numerical measures of center and spread	<ul style="list-style-type: none"> ▪Mean ▪Standard deviation <p>Note: for reasonably symmetric distributions which are free of outliers</p>	<ul style="list-style-type: none"> ▪Median ▪Quartiles <ul style="list-style-type: none"> •Five number summary: Min Q1 Median Q3 Max <p>Note: better describing a skewed distribution or a distribution with strong outliers</p>

Example 1.19

A central principle in the study of investments is that taking bigger risks is rewarded by higher returns, at least on the average over long periods of time. Let's compare the approximate mean and standard deviation of the annual percent returns on American common stocks and U.S. Treasury bills over the period from 1950 to 2000:

Investment	Mean return	Standard deviation
Common stocks	13.3%	17.1%
Treasury bills	5.2%	2.9%

Stemplot of annual returns for stocks

```

-2 | 8
-1 | 911000
-0 | 9643
 0 | 000123899
 1 | 1334466678
 2 | 0112344457799
 3 | 113467
 4 | 5
 5 | 0

```

Leaf unit=1

Stemplot of annual returns for Treasury bills

```

 0 | 9
 1 | 255668
 2 | 15779
 3 | 01155899
 4 | 24778
 5 | 1122256678
 6 | 24569
 7 | 278
 8 | 048
 9 | 8
10 | 45
11 | 3
12 |
13 |
14 | 7

```

Leaf unit=0.1

VI. Change the unit of measurement

- A linear transformation that changes x into x_{new} can be denoted by

$$x_{\text{new}} = a + bx \quad \text{where } a \text{ and } b \text{ are constants.}$$

- Properties:

- Linear transformations do not change the shape of a distribution (for b non zero).

Symmetric ----- symmetric

(right) skewed ----- (still right) skewed if $b > 0$, left skewed otherwise

Unimodal ----- unimodal

-
-

$$\bar{x}_{\text{new}} = a + b \bar{x} \quad \text{mean}$$

$$M_{x_{\text{new}}} = a + b M_x \quad \text{median}$$

-
-
- $S_{x_{\text{new}}} = |b|s_x$ (standard deviation)

- Standard deviation of a new variable is affected only by multiplier b , not shift a .