Section 1.2 - Numerical Summaries

Statistics 104

Autumn 2004

Numerical Summaries

- Center: Mean, Median

  Example: Soybean protein

  A sample of 6 soybean plants was collected and the protein content of the leaves was measured.

11.7	16.1	14.0	6.1	5.1	4.9
x_1	x_2	x_3	x_4	x_5	x_6

$$\begin{aligned} \text{Mean} &= \text{Average} = \bar{x} \\ &= \frac{1}{n}(x_1 + x_2 + \ldots + x_n) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i \end{aligned}$$

For the example

$$\bar{x} = \frac{11.7 + 16.1 + 14.0 + 6.1 + 5.1 + 4.9}{6} = 9.65$$

Median = Middle value (of the ordered data set)

Data: 11.7 16.1 14.0 6.1 5.1 4.9

$$4.9 \quad 5.1 \quad 6.1 \quad 11.7 \quad 14.0 \quad 16.1$$

$$\text{Med} = \frac{6.1 + 11.7}{2} = 8.9$$

Now add 20 to the dataset

$$4.9 \quad 5.1 \quad 6.1 \quad 11.7 \quad 14.0 \quad 16.1 \quad 20.0$$

$$\text{Med} = 11.7$$

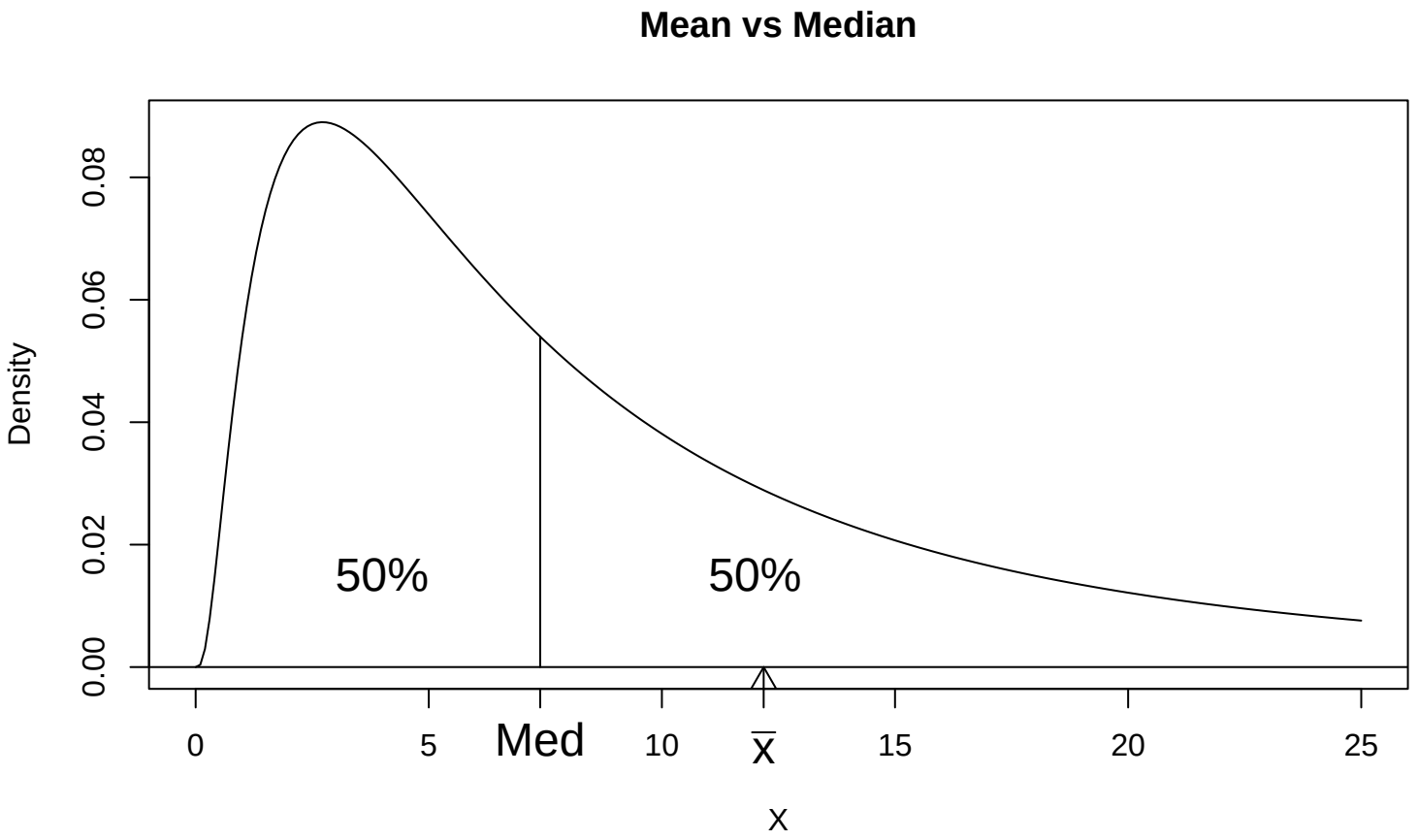
- Odd number of observations

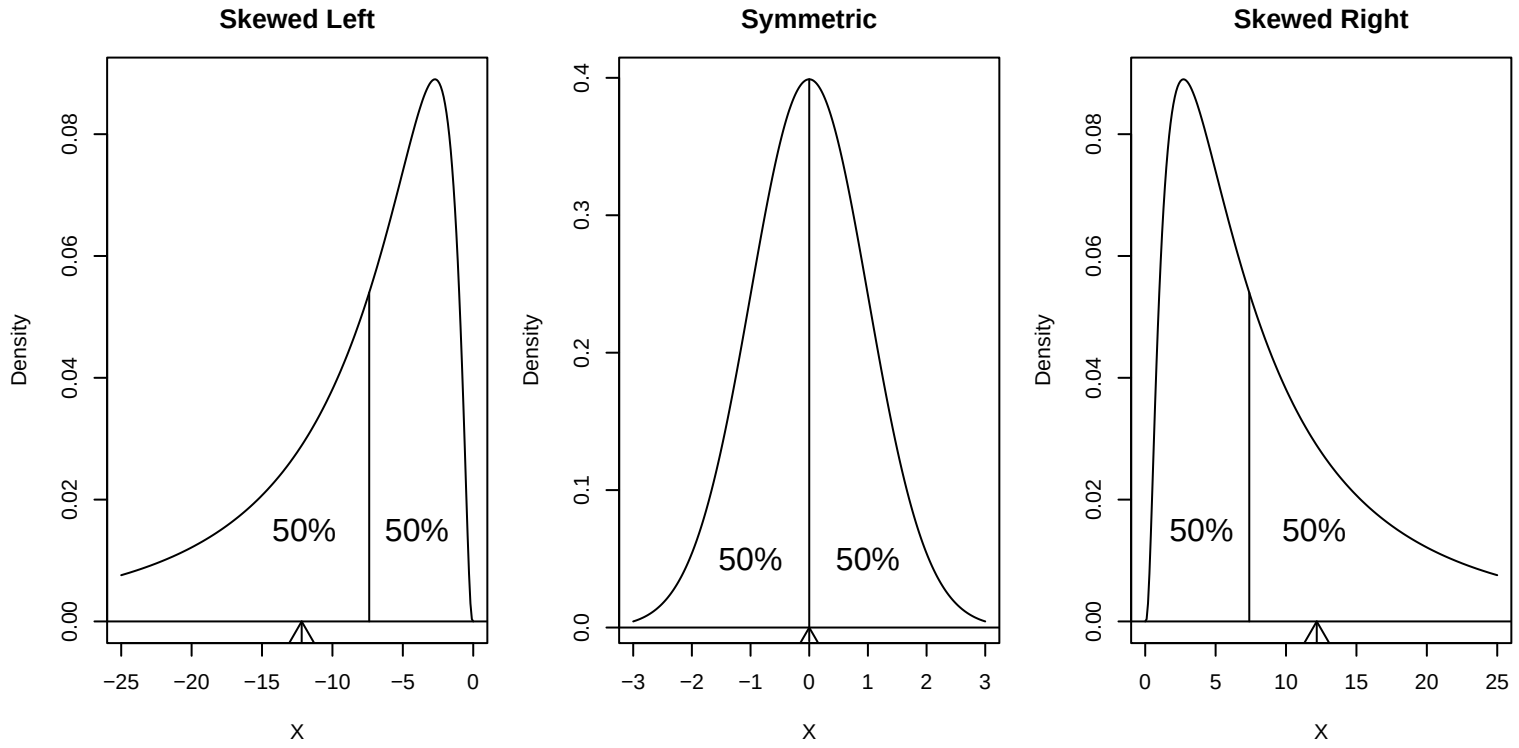
  Median = middle value

- Even number of observations

  Median = average of the 2 middle values

Comparing the two





Skewed Left	Symmetric	Skewed Right
$\bar{x} < \text{Median}$	$\bar{x} \approx \text{Median}$	$\bar{x} > \text{Median}$

Back to Soybean Example

Data: 4.9 5.1 6.1 11.7 14.0 16.1

$$\bar{x} = 9.65 \qquad \text{Med} = 8.9$$

Add one more observation at 50

$$\bar{x} = 15.41 \qquad \text{Med} = 11.7$$

Instead of 50, add one more observation at 100

$$\bar{x} = 22.56 \qquad \text{Med} = 11.7$$

The mean may be strongly influenced by 1 or 2 observations.

The median is not.

A **Resistant** measure is one that is not unduly influenced by a few values in the data set.

– The mean is not resistant
– The median is resistant

Which should you use: Mean or Median?

General rule of thumb:

> If the data is roughly symmetric, use the mean; if skewed, use the median.

However it depends on the question of interest.

Outliers can also influence things

Example: Splitting the check at a restaurant

13 people at dinner. A couple orders appetizers and a bottle of wine in addition to their entrees. So their contribution to the total bill is much larger than the rest, leading the a right skewed distribution. How to split the check?

$$\begin{aligned} \text{Median:} \quad & 13 * \text{Median} < \text{Total Bill} \\ \text{Mean:} \quad & 13 * \text{Mean} = \text{Total Bill} \end{aligned}$$

So if you are interested in totals, the mean will often be a better description of the distribution.

Example: Fast Food Restaurant Hourly Wages

\$5.00 \$5.00 \$5.00 \$5.00 \$5.25
\$5.25 \$5.25 \$8.25 \$120.00

$$\bar{x} = 18.22 \qquad \text{Med} = 5.25$$

Which to use: ???

Example: Insurance company payouts

\$0.00 \$0.00 \$0.00 \$0.00 \$10,000

$$\bar{x} = 2000 \qquad \text{Med} = 0$$

Which to use: ???

Example: Fast Food Restaurant Hourly Wages

\$5.00 \$5.00 \$5.00 \$5.00 \$5.25
\$5.25 \$5.25 \$8.25 \$120.00

$$\bar{x} = 18.22 \qquad \text{Med} = 5.25$$

Which to use: Median

Probably of more interest is the typical wage. The mean in this case does not give this information as it is greater than all but one of the observations. However \$5.25 is a reasonable summary of the typical hourly wage paid.

Example: Insurance company payouts

\$0.00 \$0.00 \$0.00 \$0.00 \$10,000

$$\bar{x} = 2000 \qquad \text{Med} = 0$$

Which to use: ???

Example: Fast Food Restaurant Hourly Wages

\$5.00 \$5.00 \$5.00 \$5.00 \$5.25
\$5.25 \$5.25 \$8.25 \$120.00

$$\bar{x} = 18.22 \qquad \text{Med} = 5.25$$

Which to use: Median

Example: Insurance company payouts

\$0.00 \$0.00 \$0.00 \$0.00 \$10,000

$$\bar{x} = 2000 \qquad \text{Med} = 0$$

Which to use: Mean

The total amount of payouts is probably more important, which can only be derived from the mean. Note that the actual situation is more complex as the relationship between premiums and claims payed out is probably important.

Example: Boston Area House Prices

Distribution is skewed right, with a very long tail.

If you are moving into the area and are looking to buy a new house, the median is probably a better descripion.

However if you are trying to figure out how much money will be coming in from property taxes, the mean will probably be a more useful summary.

- Spread: Standard Deviation, Interquartile Range

  Standard Deviation:

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

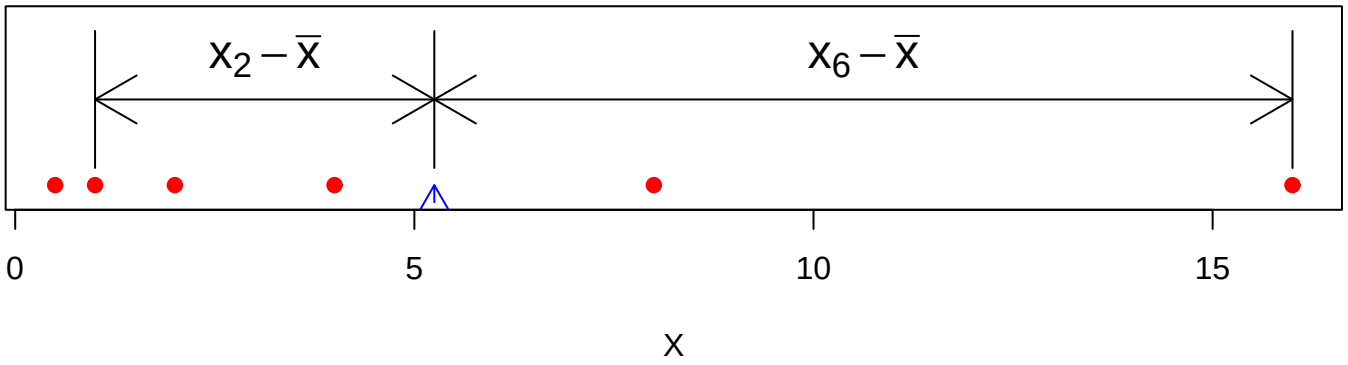
  This is based on the variance

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

  which is an average squared deviation

Interpretation:

$x_i - \bar{x}$ = deviation from the mean



"Big" deviation $\Rightarrow$ "Big" $(\text{deviation})^2$

$\Rightarrow$ "Big" spread

Can think of s as a "typical" deviation

Can think of s^2 as a "typical" squared deviation

Example:

Data: 3,2,1,0,-1

$$\bar{x} = 1; \text{Med} = 1; n = 5$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
3		
2		
1		
0		
-1		
Total		

Example:

Data: 3,2,1,0,-1

$$\bar{x} = 1; \text{Med} = 1; n = 5$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
3	$3 - 1 = 2$	4
2	$2 - 1 = 1$	1
1	$1 - 1 = 0$	0
0	$0 - 1 = -1$	1
-1	$-1 - 1 = -2$	4
Total	0	10

$$\begin{aligned} s^2 &= \frac{10}{5-1} = 2.5 \\ s &= \sqrt{2.5} = 1.58 \end{aligned}$$

Soybean example

$\bar{x} = 9.65$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
4.9	4.9 - 9.65 = -4.75	22.5625
5.1	5.1 - 9.65 = -4.55	20.7025
6.1	-3.55	12.6025
11.7	2.05	4.2025
14.0	4.35	18.6025
16.1	6.45	41.6025
Total	0	120.595

$$\begin{aligned} s^2 &= \frac{120.595}{5} = 24.119 \\ s &= \sqrt{24.119} = 4.911 \end{aligned}$$

$n - 1 =$ Degrees of freedom

Why $n - 1$ and not n?

In the two previous examples, the sum of $x_i - \bar{x}$ is 0. In fact this always has to hold since

$$\sum (x_i - \bar{x}) = \sum x_i - n\bar{x} = \sum x_i - \sum x_i = 0$$

One consequence of this, is if you know $n - 1$ of the deviations, then you know the remaining one

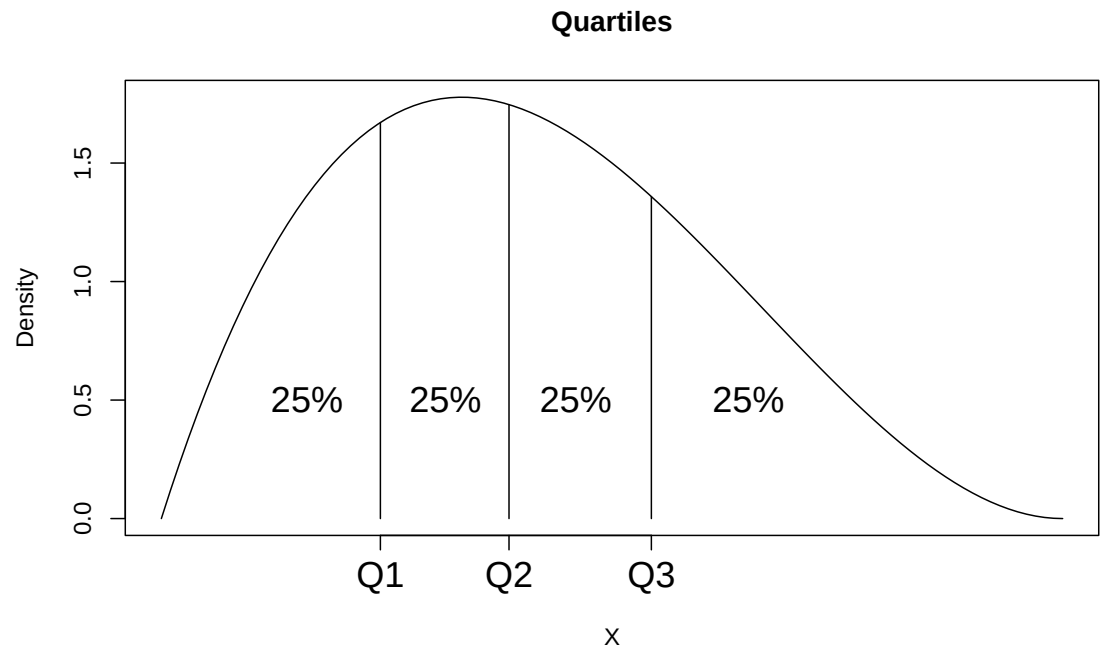
or

If you know all but one observations and $\bar{x}$, then you know the last observation.

Also, it can be shown that the estimators based on $n - 1$ have better properties in terms of describing the population being sampled from.

Quartiles

- The median splits the data into halves
- Can look into other splits
- Quartiles split data into quarters
- Other common splits are deciles (10ths) and percentiles



Q1 = 1st quartile; Q2 = 2nd quartile = median; Q3 = 3rd quartile

How to calculate

1. Order the data
2. Find the median
3. Q1 = median of the observations below the location of the median
   Q3 = median of the observations above the location of the median

Soybean data

$$4.9 \quad 5.1 \quad 6.1 \quad 11.7 \quad 14.0 \quad 16.1$$

$$\begin{aligned} \text{Med} &= \frac{6.1 + 11.7}{2} = 8.9 \\ Q1 &= \text{Median}(4.9, 5.1, 6.1) = 5.1 \\ Q3 &= \text{Median}(11.7, 14.0, 16.1) = 5.1 \end{aligned}$$

What if n is odd (meaning that the median is one of the observations in the data set)

Data: $\underbrace{-1 \quad 4 \quad 5}$ $\quad 6 \quad$ $\underbrace{6 \quad 7 \quad 10}$

Med $= 6$

If n is odd, don't include the median observation in determining the quartiles

If the median value is repeated, as in this example, use the repeats

$$\begin{aligned} Q1 &= \text{Median}(-1, 4, 5) = 4 \\ Q3 &= \text{Median}(6, 7, 10) = 7 \end{aligned}$$

Note about computer software:

Many statistics package use slightly different algorithms for calculating the quartiles.

For example, Stata may give slightly different answers than this algorithm.

Don't worry about the differences from the different algorithms as they usually are small.

Data: $\underbrace{-1 \quad 4 \quad 5} \quad 6 \quad \underbrace{6 \quad 7 \quad 10}$

By book's algorithm

$$Q1 = 4 \qquad \text{Med} = 6 \qquad Q3 = 7$$

Stata gives the same answer.

However with the data

Data: $\underbrace{-1 \;\; 4 \;\; 5 \;\; 6}\;\; 6 \;\; \underbrace{7 \;\; 10 \;\; 15 \;\; 20}$ $\qquad (n = 9)$

By the book's algorithm

$$Q1 = 4.5 \qquad \text{Med} = 6 \qquad Q3 = 12.5$$

var5

Percentiles		Smallest		
1%	-1	-1		
5%	-1	4		
10%	-1	5	Obs	9
25%	5	6	Sum of Wgt.	9
50%	6		Mean	8
		Largest	Std. Dev.	6.244998
75%	10	7		
90%	20	10	Variance	39
95%	20	15	Skewness	.6761
99%	20	20	Kurtosis	2.781435

Can be different if $n = 4m + 1$ for some integer m.

Interquartile Range (IQR): $IQR = Q3 - Q1$

Length of the range of the middle 50% of the data

For the Soybean data

$$Q1 = 5.1 \quad Q3 = 14.0 \quad IQR = 14.0 - 5.1 = 8.9$$

Properties of s and IQR

1. $s \geq 0, IQR \geq 0$

2. All observations are the same implies

$$s = IQR = 0$$

3. Suppose that $s = 0$. Then all observations must be the same.

4. Suppose that $IQR = 0$. This does **not** imply that all the observations are the same. (Try to create a dataset where $IQR = 0, \text{Med} = 5, \text{Min} = 1,$ and $\text{Max} = 10$. Show that s and s^2 are both bigger than 0.)

5. IQR is a resistant measure, s is not.

When to use s or IQR?

Rule of Thumb

If $\bar{x}$ is a reasonable center of center, use s.

If Median is a reasonable summary of center, use IQR.

- Outliers

  A useful technique for searching for outlier can be based on the quartiles and the IQR.

  1.5 IQR Criterion

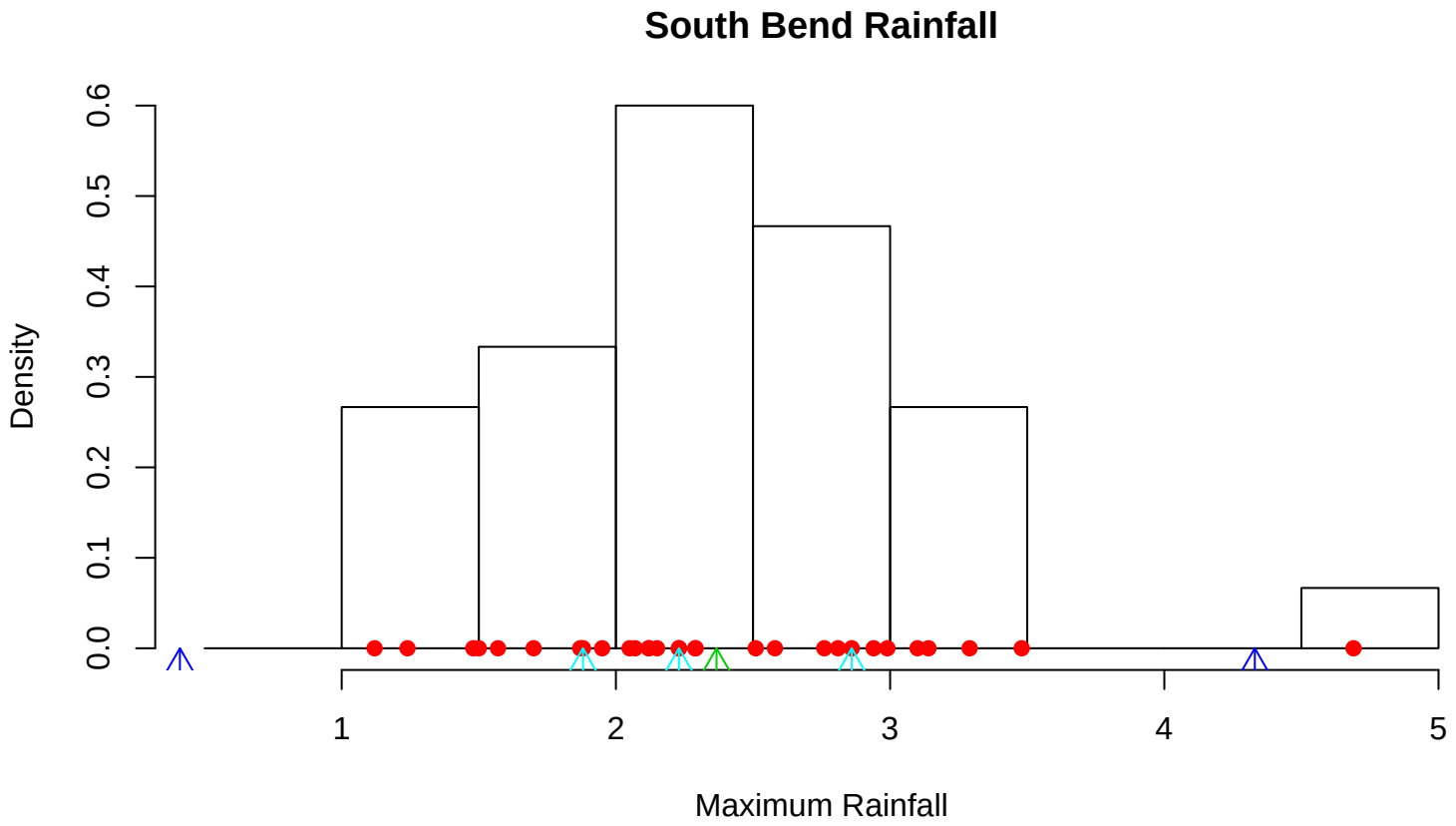
  Look for observations

$$x_i > Q3 + 1.5IQR \quad \text{or} \quad x_i < Q1 - 1.5IQR$$

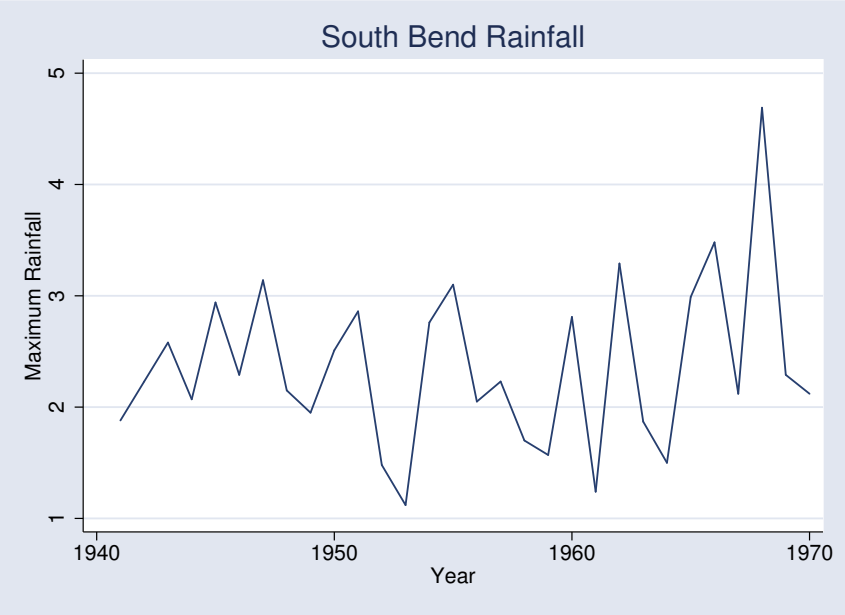
South Bend Rainfall Example

$$Q1 = 1.88, Q3 = 2.86 \quad \Rightarrow \quad IQR = 0.98$$

$$Q1 - 1.5IQR = 0.41 \quad Q3 + 1.5IQR = 4.33$$



One **possible** outlier by this rule (which happens to be from 1968)



Nothing in this data set suggests what might be happening here.

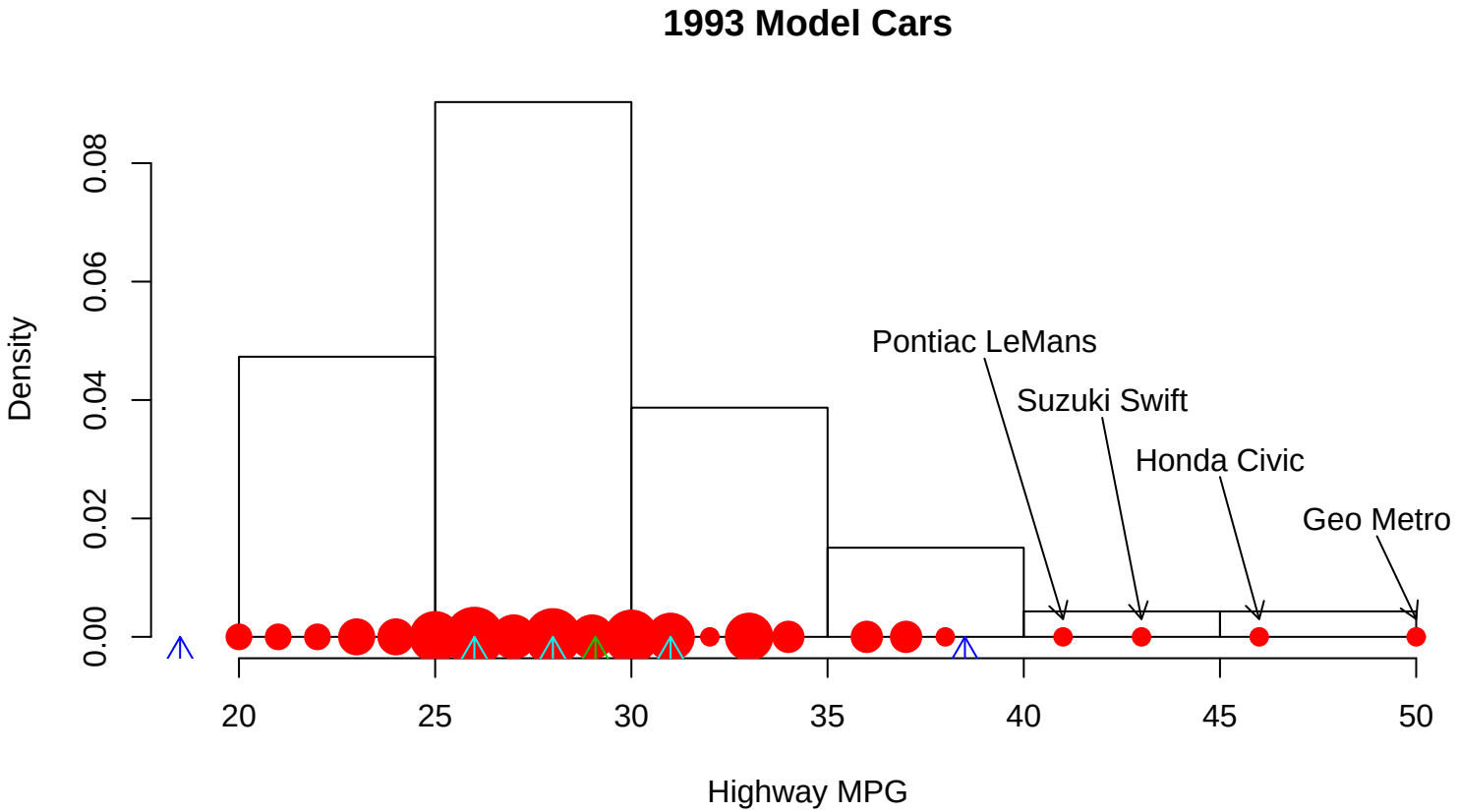
Data sets looking at maxima, such as here or with measures of strengths of earthquakes will often have an extreme observation like this.

This procedure is based on assumptions about the underlying distribution. There are processes that lead to very long tailed histograms. This may be an example of one.

Highway MPG Example

$$Q1 = 26, Q3 = 31 \quad \Rightarrow \quad IQR = 5$$

$$Q1 - 1.5IQR = 18.5 \quad Q3 + 1.5IQR = 38.5$$



Four **possible** outliers by this rule

Model	Geo Metro	Honda Civic	Pontiac LeMans	Suzuki Swift
HighMPG	50	46	41	43
CityMPG	46	42	31	39
Cylinder	3	4	4	3
Engine Size (l)	1.0	1.5	1.6	1.3
Horsepower	55	102	74	70
Length (in)	151	173	177	161
Weight (lbs)	1695	2350	2350	1965

These are some of the lightest cars in the data set. If we adjust our analysis for the weight of the car, the only one of these cars that really stands out is the Honda Civic.

Boxplots

5 figure summary

Minimum Q1 Median Q3 Maximum

Useful for

- Simple numerical summary of the data

- Gives information for another graphical summary

Basic Boxplot

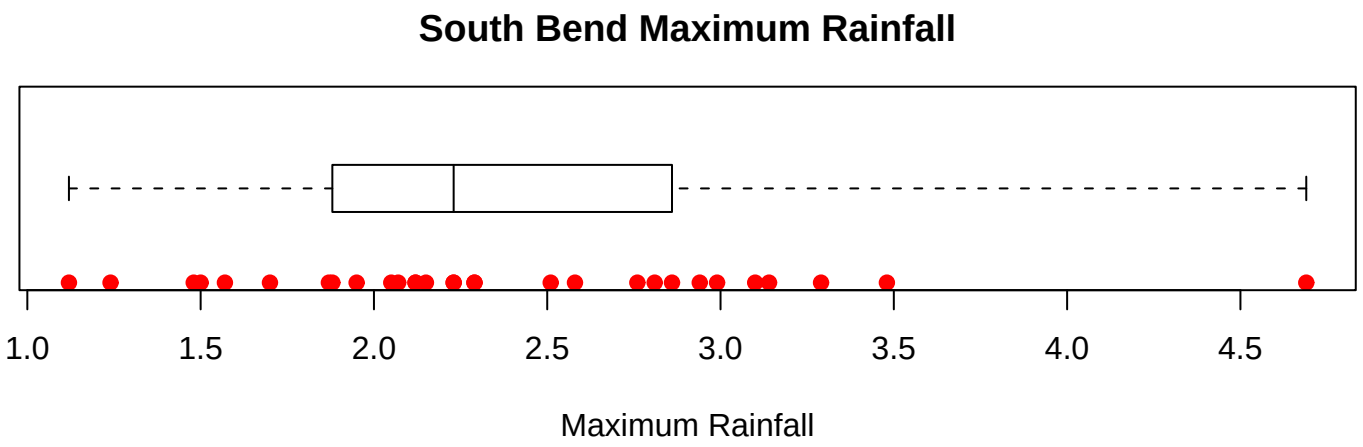
1. Ends of the box are at the quartiles

2. Median is indicated by a line across the box

3. The two lines outside the box extend to the largest and smallest observations. (Whiskers)

Example: South Bend Rainfall

Min	Q1	Med	Q3	Max
1.12	1.88	2.23	2.86	4.69



Most programs, when asked for a box plot, do not give you this.

Instead they give, . . .

Modified Boxplot

1. Same as before, ends of the box are at the quartiles

2. Same as before, median is indicated by a line across the box

3. Find 1.5 IQR limits

$$\text{Lower fence} = Q1 - 1.5IQR \qquad \text{Upper fence} = Q3 + 1.5IQR$$

If $Max < \text{Upper fence}$ (no big potential outlier)

Draw whisker from the box to Max

Otherwise $Max > \text{Upper fence}$ (big potential outlier)

Draw whisker from the box to largest value $< \text{Upper fence}$ and plot extreme values individually.

If $Min >$ Lower fence (no small potential outlier)

Draw whisker from box to Min

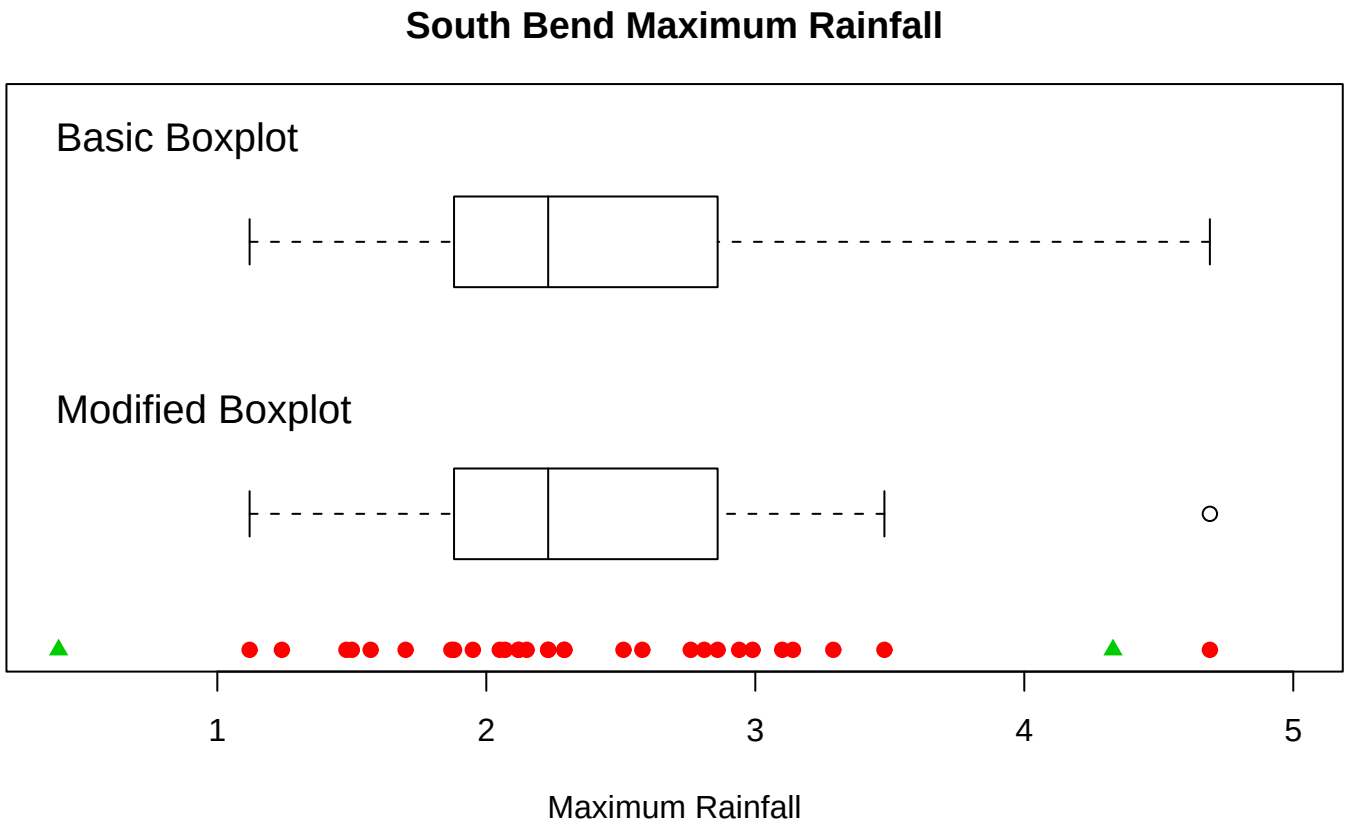
Otherwise $Min <$ Lower fence (small potential outlier)

Draw whisker from box to smallest value $>$ Lower fence and plot extreme values individually.

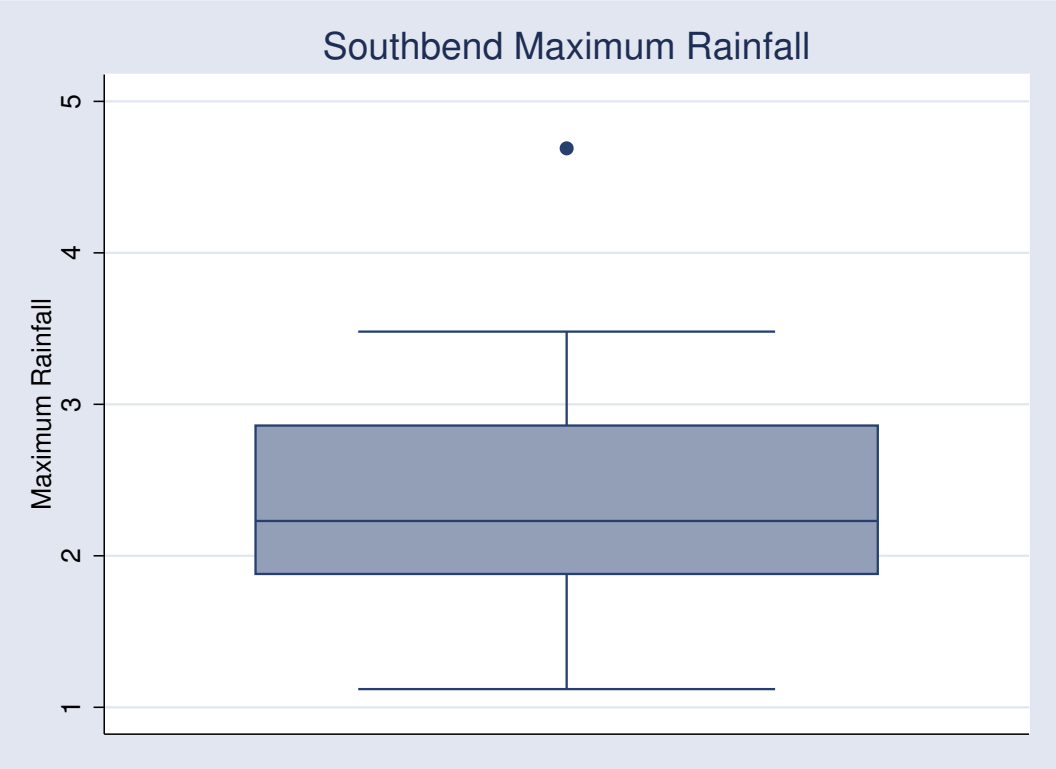
Example: South Bend Rainfall

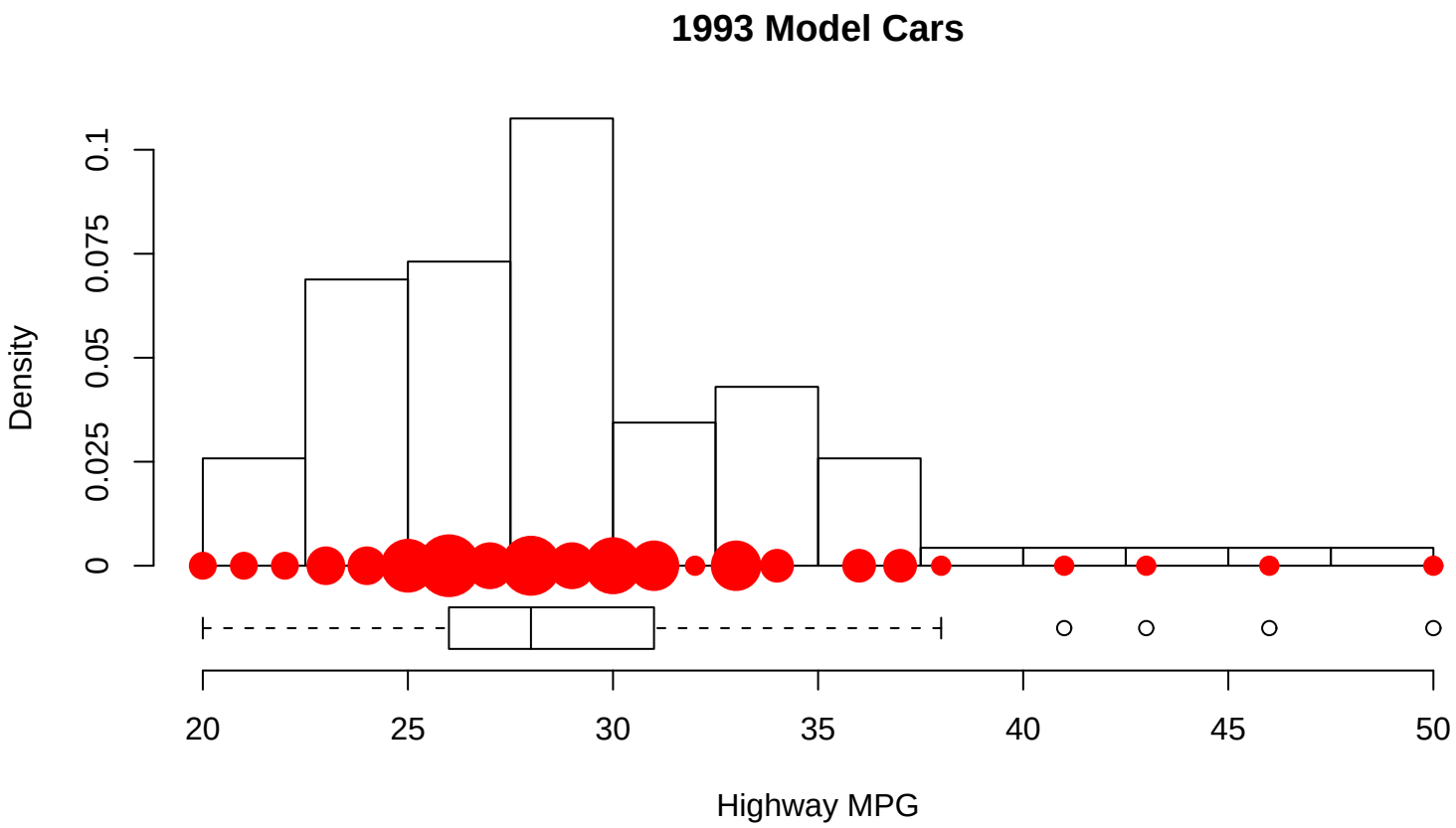
Min	Q1	Med	Q3	Max
1.12	1.88	2.23	2.86	4.69

$$Q1 - 1.5IQR = 0.41 \quad Q3 + 1.5IQR = 4.33$$



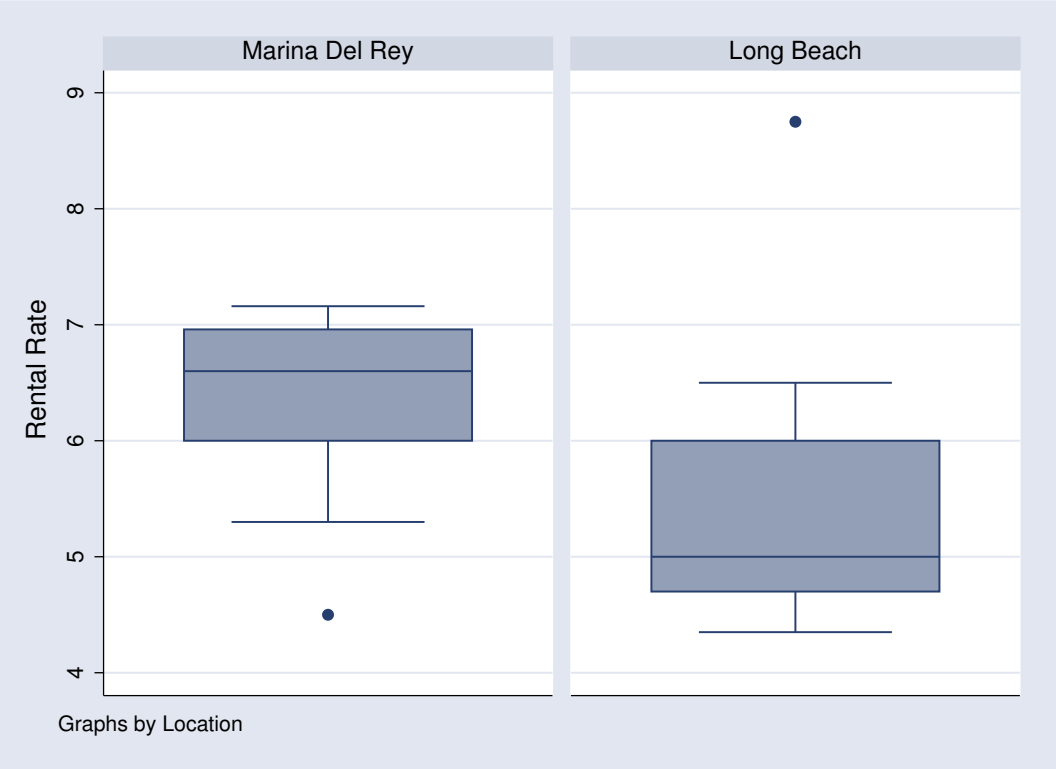
Boxplots can be done with horizontally, or vertically (as below). For most software, the default is vertical bars.



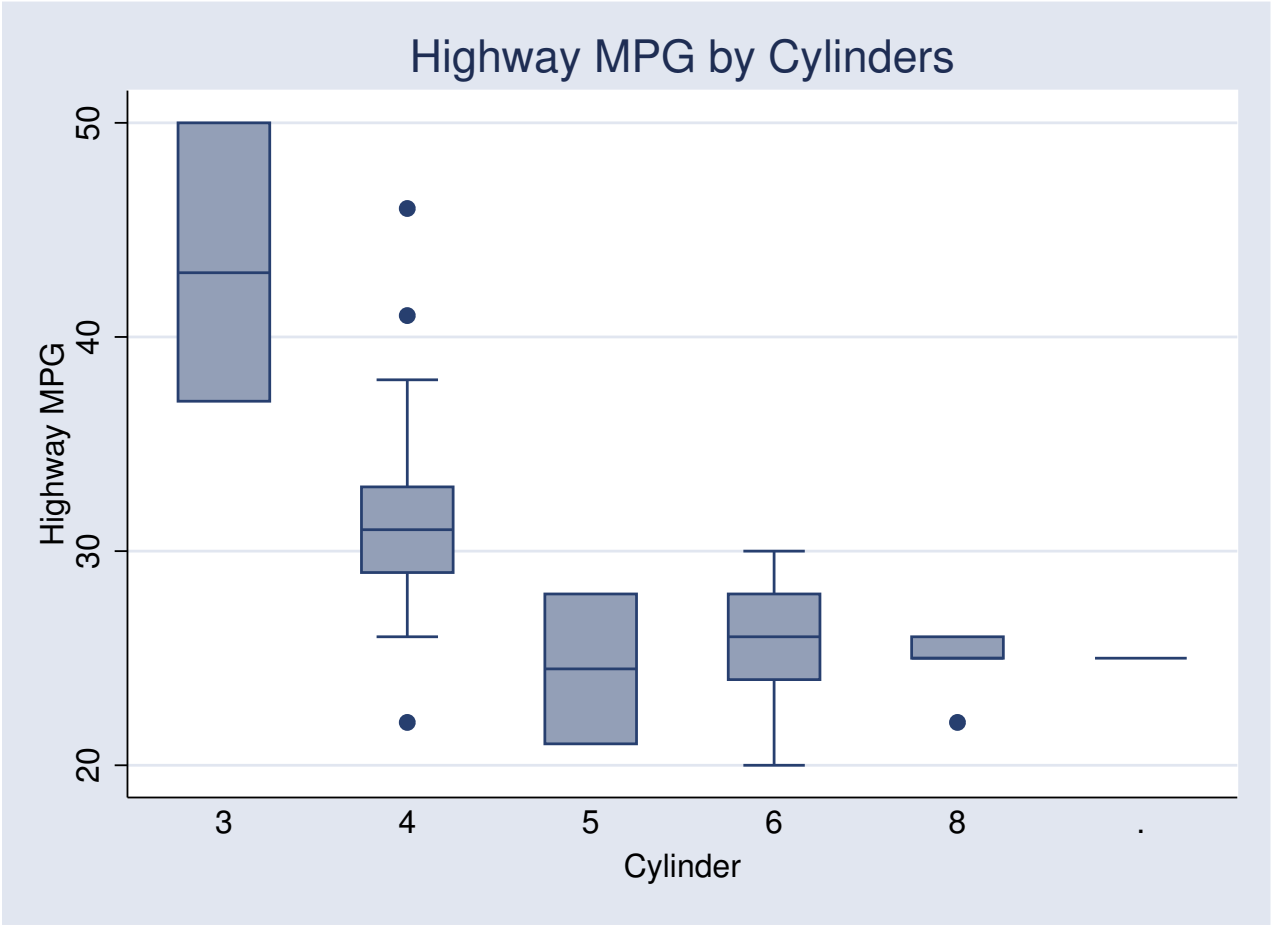

1993 Model Cars
Density
0
0.025
0.05
0.075
0.1
20
25
30
35
40
45
50
Highway MPG

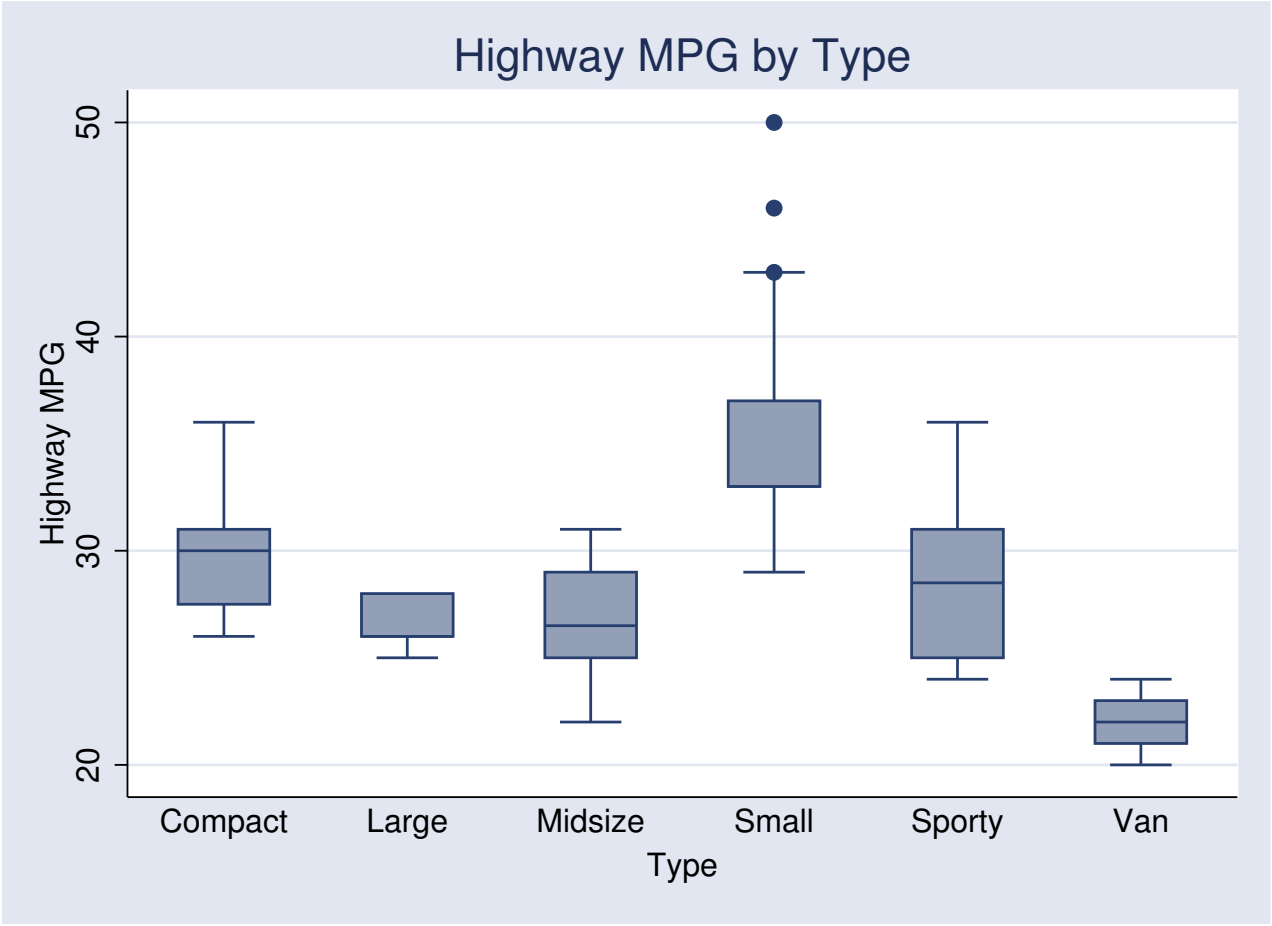
Uses for boxplots

- Basic structure of distributions - location, spread, shape, outliers
- Comparisons - side by side boxplots.



These are particularly useful when you want to compare more than two groups.




Highway MPG by Type
Highway MPG
50
40
30
20
Compact
Large
Midsize
Small
Sporty
Van
Type

Changing Units

MPG to km/l or °F to °C

$$km/l = 0.425MPG \qquad MPG = 2.35km/l$$

$$^\circ C = (^\circ F - 32) \times \frac{5}{9} = \frac{5^\circ}{9} F - \frac{160}{9}$$
$$^\circ F = \frac{9^\circ}{5} C + 32$$

In general, we are interested in transformations of the form

$$Y = a + bX$$

where

$$Y : \text{New Units} \qquad X : \text{Old Units}$$

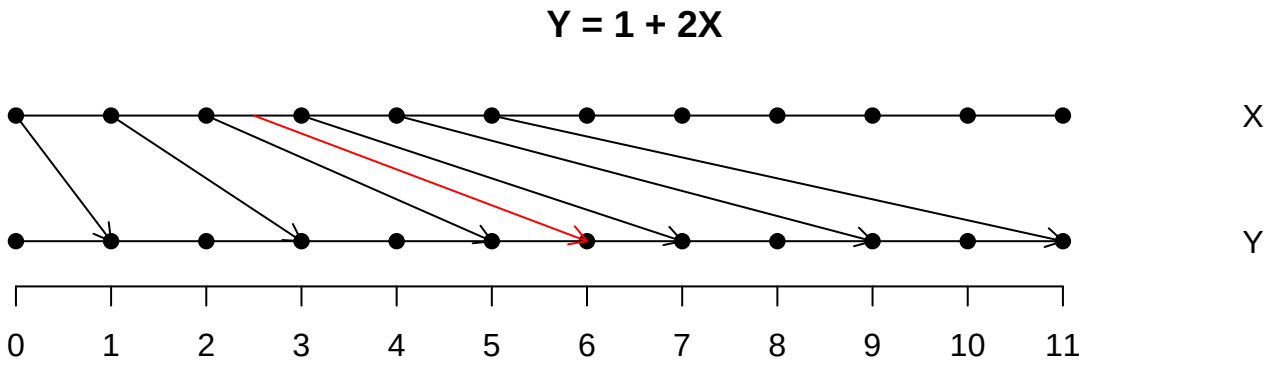
What happens to $\bar{x}, \text{Med}, Q1, Q3, s, IQR$?

Could transform all the data and then calculate the desired summaries with the transformed data, or ...

$$\bar{y} = a + b\bar{x} \qquad \text{Med}_y = a + b\text{Med}_x$$

$$Q1_y = \begin{cases} a + bQ1_x & b > 0 \\ a + bQ3_x & b < 0 \end{cases} \qquad Q3_y = \begin{cases} a + bQ3_x & b > 0 \\ a + bQ1_x & b < 0 \end{cases}$$

$$s_y^2 = b^2 S_x^2 \qquad s_y = |b| s_x \qquad IQR_y = |b| IQR_x$$



	MPG	km/l
$\bar{x}$	29.09	$0.425 \times 29.09 = 12.36$
Med	28	$0.425 \times 28 = 11.90$
s	5.33	$0.425 \times 5.33 = 2.265$
IQR	5	$0.425 \times 5 = 2.125$

HighKML

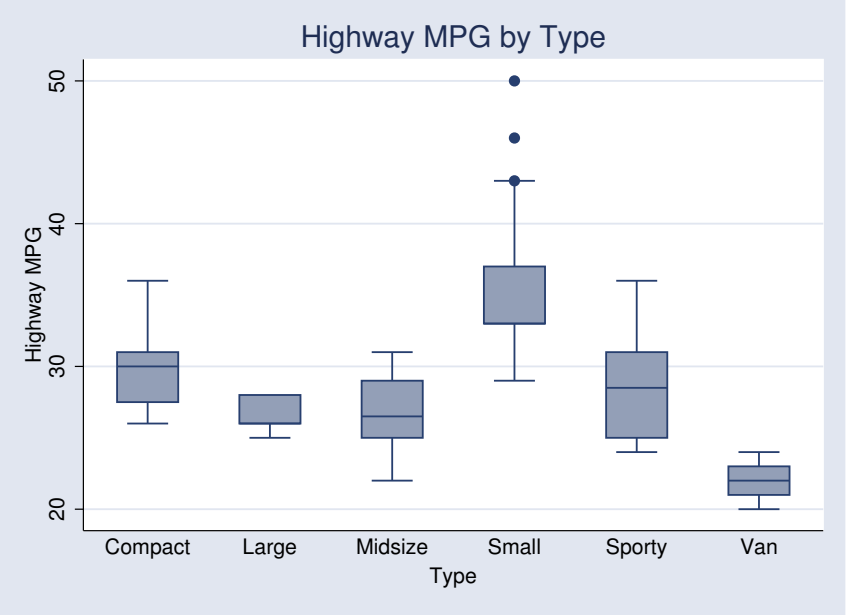
	Percentiles	Smallest		
1%	8.5	8.5		
5%	9.35	8.5		
10%	9.775	8.925	Obs	93
25%	11.05	8.925	Sum of Wgt.	93
50%	11.9		Mean	12.36156
		Largest	Std. Dev.	2.265984
75%	13.175	17.425		
90%	15.3	18.275	Variance	5.134681
95%	16.15	19.55	Skewness	1.20997
99%	21.25	21.25	Kurtosis	5.411922

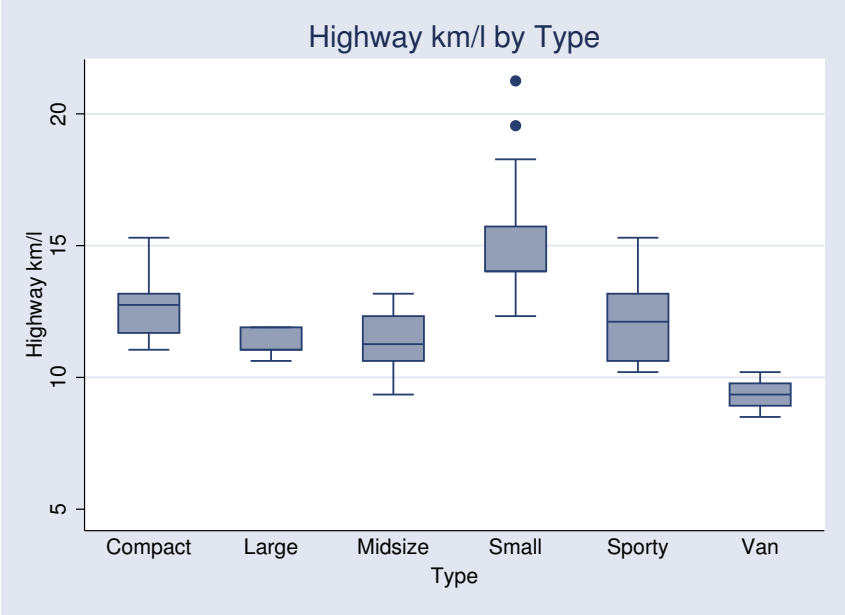
$$\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) \\
&= a + b\frac{1}{n}\sum_{i=1}^{n} x_i = a + b\bar{x}
\end{aligned}$$

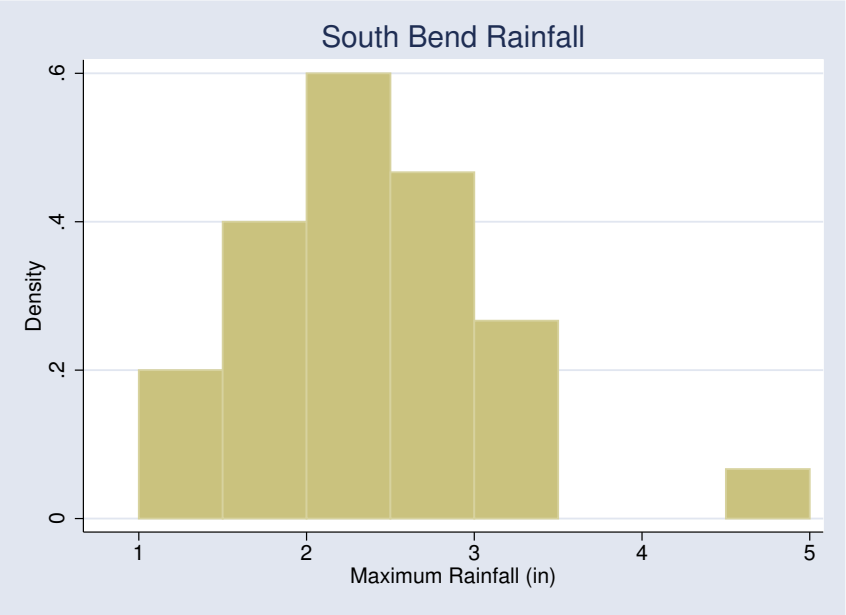
$$\begin{aligned}
S_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}((a + bx_i) - (a + b\bar{x}))^2 \\
&= b^2\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= b^2 s_x^2
\end{aligned}$$

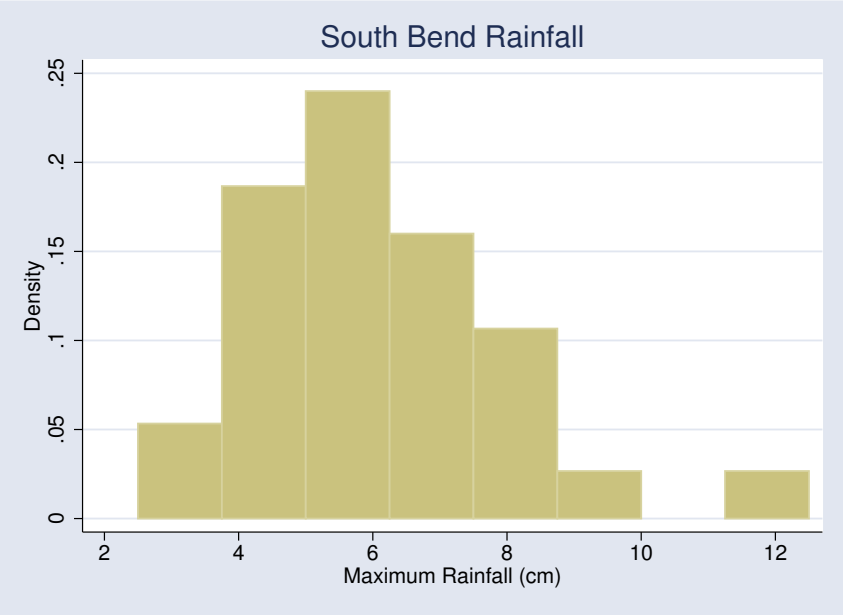
Note that linear transformations like these do not effectively change the distribution, but just relabels values.

However, when graphical summaries are used, you can get slightly different pictures.









In this example the bins are set to be approximately equivalent

- Width: 1.25 cm vs 0.5 in
- Bin start: 2.5 cm vs 1 in