Statistical Science 1024

Chapter 2

Describing Distributions with Numbers

Henry Cavendish (1731 – 1810)

- British scientist
- carried out wide-ranging scientific studies that included
  - chemistry
  - electricity
  - physics
  - astronomy
- Cavendish Laboratory for physics at Cambridge University named for him

Cavendish's Density of the Earth Data

In 1798 Cavendish made 23 measurements of the density of the earth relative to the density of water:

5.10 5.27 5.29 5.29 5.30 5.34 5.34
5.36 5.39 5.42 5.44 5.46 5.47 5.53
5.57 5.58 5.62 5.63 5.65 5.68 5.75
5.79 5.85

THE MEAN $\overline{x}$

To find the **mean** of a set of observations, add their values and divide by the number of observations. If the n observations are $x_1, x_2, \ldots, x_n$, their mean is

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

or, in more compact notation,

$$\overline{x} = \frac{1}{n} \sum x_i$$

Measure of centre: the mean

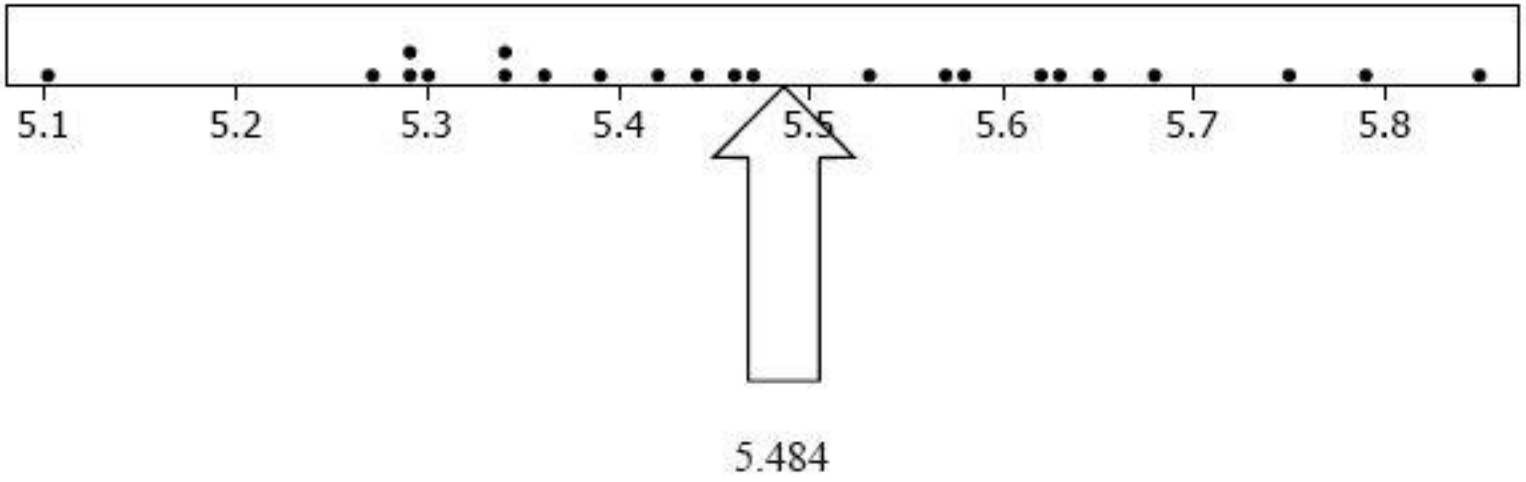
The mean or arithmetic average

To calculate the *average,* or *mean,* add all values, then divide by the number of individuals. It is the “center of mass.”

Sum of densities is 126.12
Divided by 23 measurements = 5.4835

density	
5.10	5.47
5.27	5.53
5.29	5.57
5.29	5.58
5.30	5.62
5.34	5.63
5.34	5.65
5.36	5.68
5.39	5.75
5.42	5.79
5.44	5.85
5.46	

The Mean as Centre of Gravity



THE MEDIAN *M*

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

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations n is odd, the median M is the center observation in the ordered list. Find the location of the median by counting $(n + 1)/2$ observations up from the bottom of the list.
3. If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. The location of the median is again $(n + 1)/2$ from the bottom of the list.

THE QUARTILES Q_1 and Q_3

To calculate the **quartiles:**

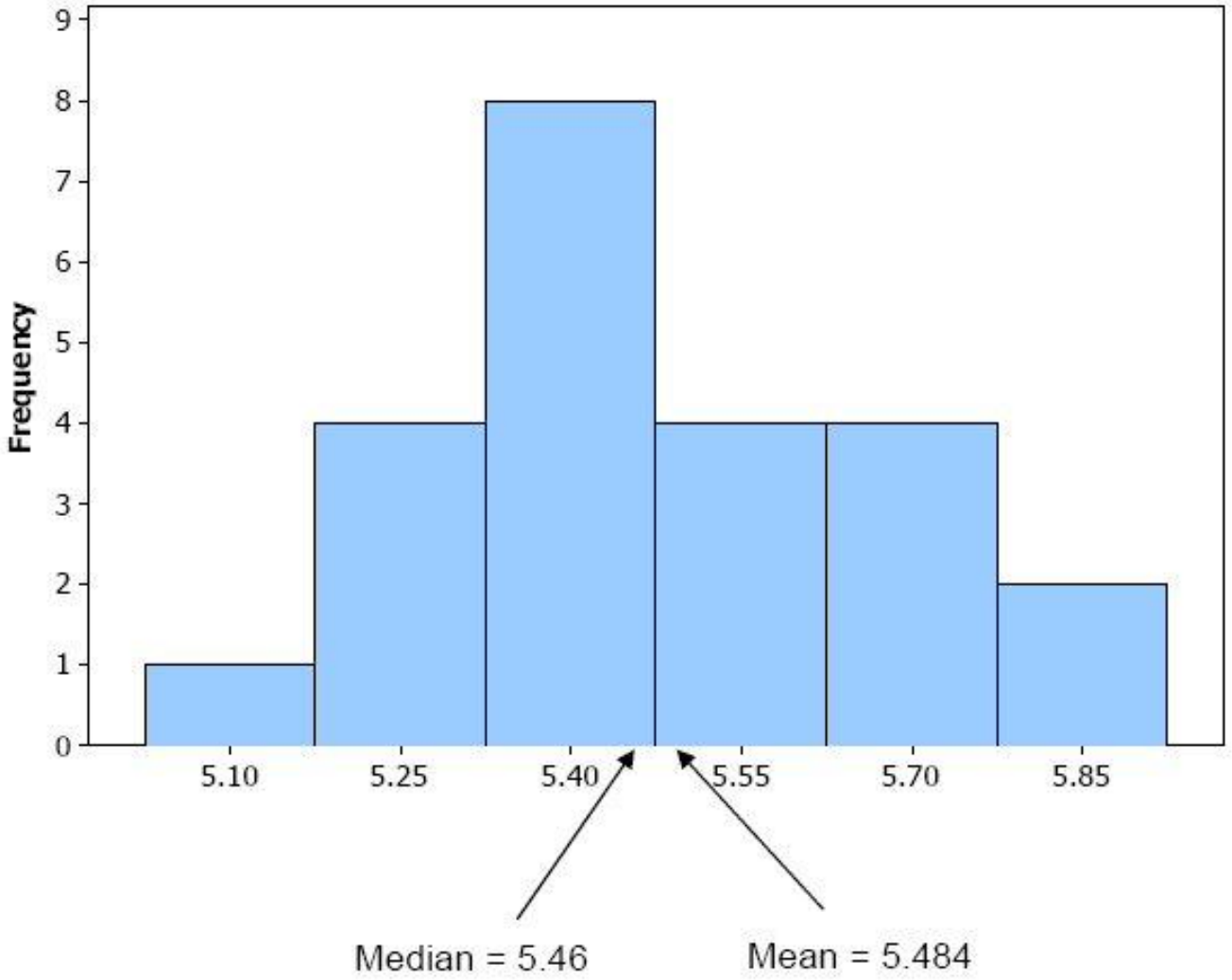
1. Arrange the observations in increasing order and locate the median M in the ordered list of observations.
2. The **first quartile** Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
3. The **third quartile** Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

Ordered Data			
5.10	First Half	5.10	
5.27	Ordered	5.27	
5.29	Data	5.29	
5.29		5.29	
5.30		5.30	
5.34		5.34	Q1
5.34		5.34	
5.36		5.36	
5.39		5.39	
5.42		5.42	
5.44		5.44	
5.46		M	Q2
5.47	Second	5.47	
5.53	Half	5.53	
5.57	Ordered	5.57	
5.58	Data	5.58	
5.62		5.62	
5.63		5.63	Q3
5.65		5.65	
5.68		5.68	
5.75		5.75	
5.79		5.79	
5.85		5.85	

COMPARING THE MEAN AND THE MEDIAN

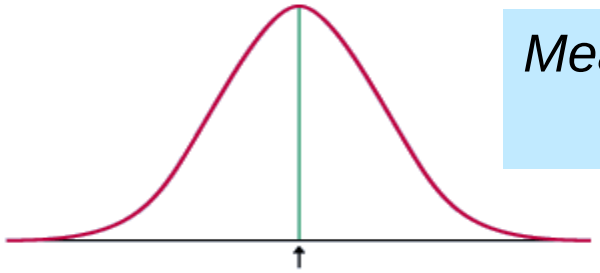
The mean and median of a roughly symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution, the mean is usually farther out in the long tail than is the median.

Cavendish Data on Density of the Earth

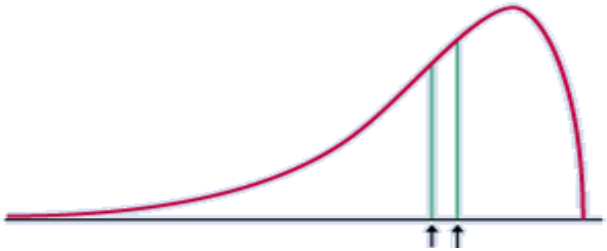


Comparing the Mean and the Median

The mean and the median are the same only if the distribution is symmetrical. The median is a measure of center that is resistant to skew and outliers. The mean is not.

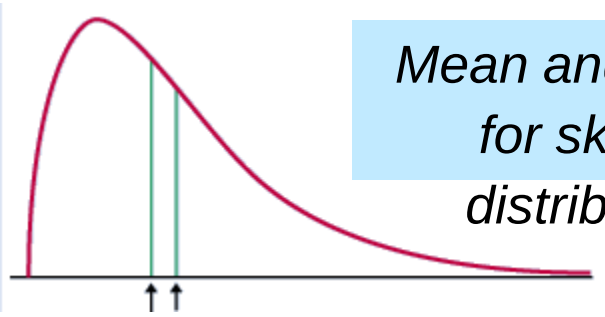


Mean and median for a symmetric distribution



Left skew

Mean
Median



Mean and median for skewed distributions

Right skew

THE FIVE-NUMBER SUMMARY

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

$$\text{Minimum} \quad Q_1 \quad M \quad Q_3 \quad \text{Maximum}$$

Cavendish Data

Ordered Data			
5.10	First Half	5.10	
5.27	Ordered	5.27	
5.29	Data	5.29	
5.29		5.29	
5.30		5.30	
5.34		5.34	Q1
5.34		5.34	
5.36		5.36	
5.39		5.39	
5.42		5.42	
5.44		5.44	
5.46		M	Q2
5.47	Second	5.47	
5.53	Half	5.53	
5.57	Ordered	5.57	
5.58	Data	5.58	
5.62		5.62	
5.63		5.63	Q3
5.65		5.65	
5.68		5.68	
5.75		5.75	
5.79		5.79	
5.85		5.85	

Minimum = 5.10

First quartile = 5.34

Median = 5.46

Third quartile = 5.63

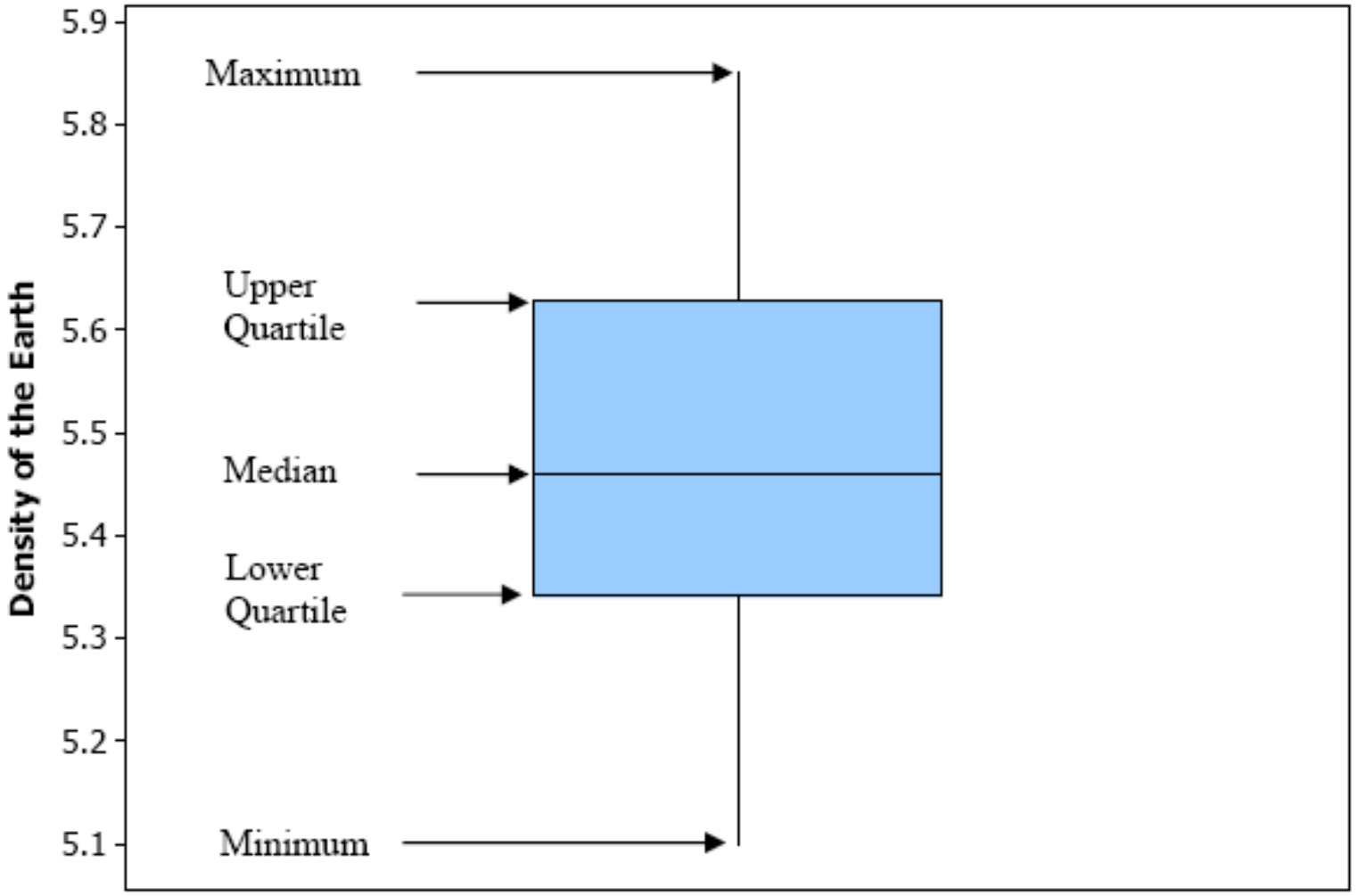
Maximum = 5.85

BOXPLOT

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.

Boxplot: Cavendish Data



THE STANDARD DEVIATION s

The **variance** s^2 of a set of observations is an average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations $x_1, x_2, \ldots, x_n$ is

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

or, more compactly,

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

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

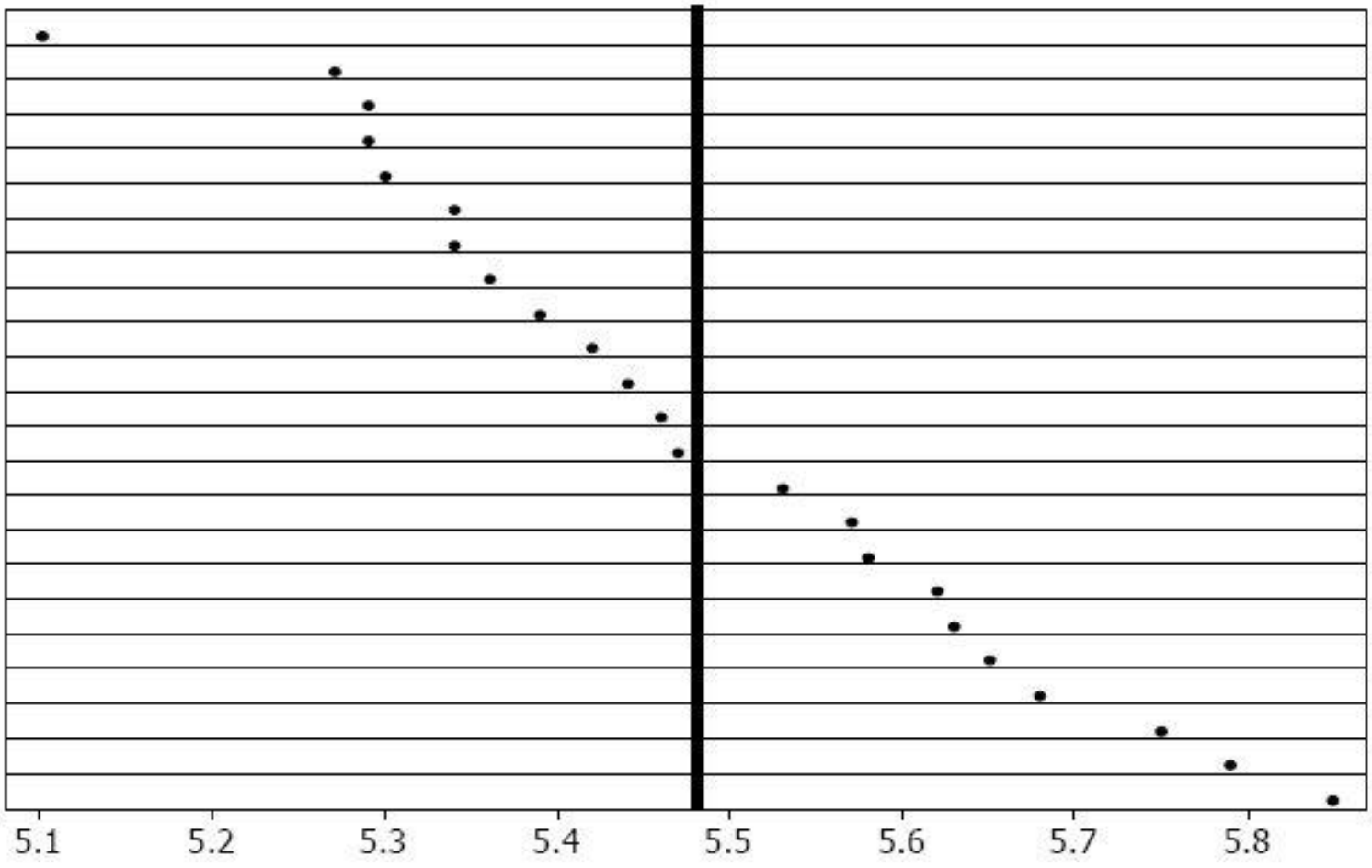
$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

Standard Deviation

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

measures the spread of the data in the same unit of measurement as the data.

Deviations from the Mean: Cavendish Data



Standard Deviation Calculations

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 =$$

$$(5.10-5.4835)^2 +$$

$$(5.27-5.4835)^2 +$$

$$\cdots +$$

$$(5.79-5.4835)^2 +$$

$$(5.85-5.4835)^2$$

$$=0.797722$$

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

$$= \frac{0.797722}{23-1}$$

$$=0.0362609$$

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

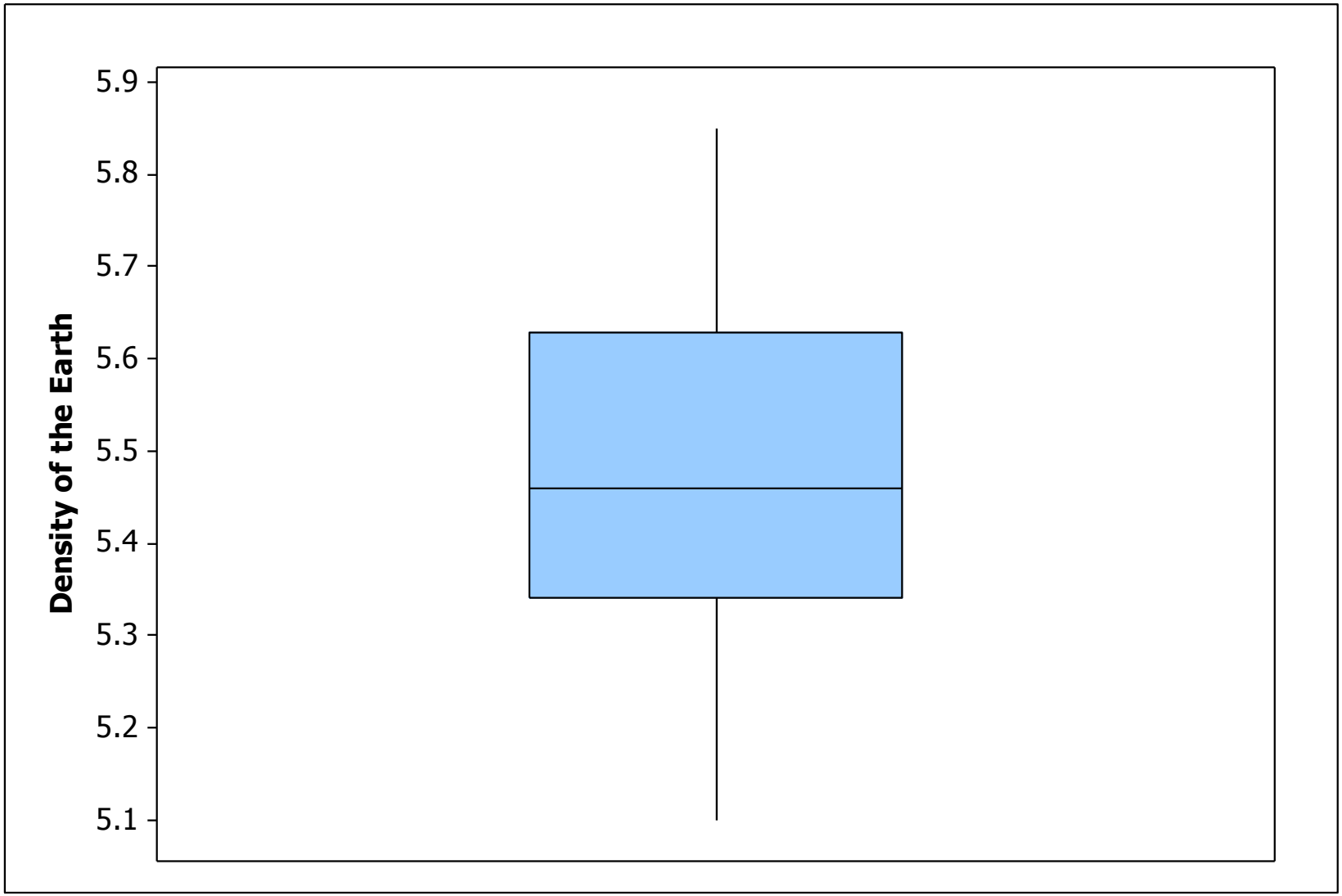
$$= \sqrt{0.0362609}$$

$$=0.1904$$

CHOOSING A SUMMARY

The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use $\bar{x}$ and s only for reasonably symmetric distributions that are free of outliers.

Cavendish Data



Groups of Observations

Observations

- are subjected to different treatments

or

- are taken at different sets of times or in different places

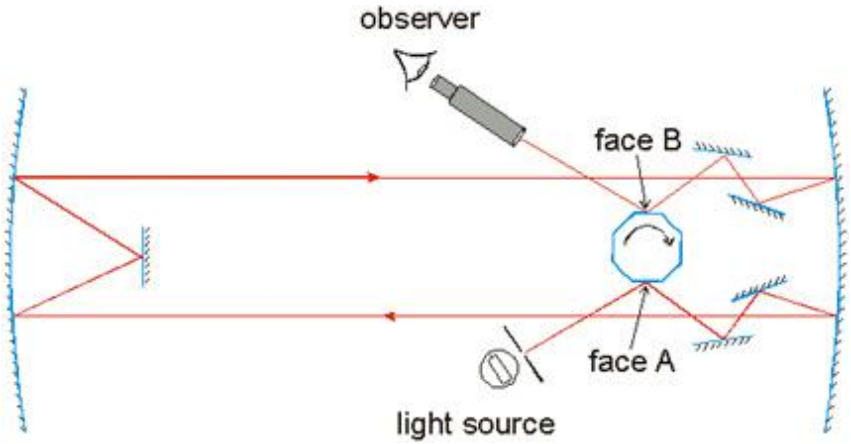
or

- occur naturally in a variety of categories (eg. male, female)

or …

Numerical Example: Speed of Light

In 1879, Albert Michelson made 100 determinations of the velocity of light in air using a modification of a method proposed by the French physicist Foucault (20 observations in each of 5 trials). The data are given are in km/sec, and have had 299,000 subtracted from them. The currently accepted "true" velocity of light in vacuum is 299,792.5 km/sec. With the corrections used by Michelson the "true" value appropriate for comparison to these measurements is 734.5.



Michelson's Data and Summary Statistics

Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
850	960	880	890	890
740	940	880	810	840
900	960	880	810	780
1070	940	860	820	810
930	880	720	800	760
850	800	720	770	810
950	850	620	760	790
980	880	860	740	810
980	900	970	750	820
880	840	950	760	850
1000	830	880	910	870
980	790	910	920	870
930	810	850	890	810
650	880	870	860	740
760	880	840	880	810
810	830	840	720	940
1000	800	850	840	950
1000	790	840	850	800
960	760	840	850	810
960	800	840	780	870

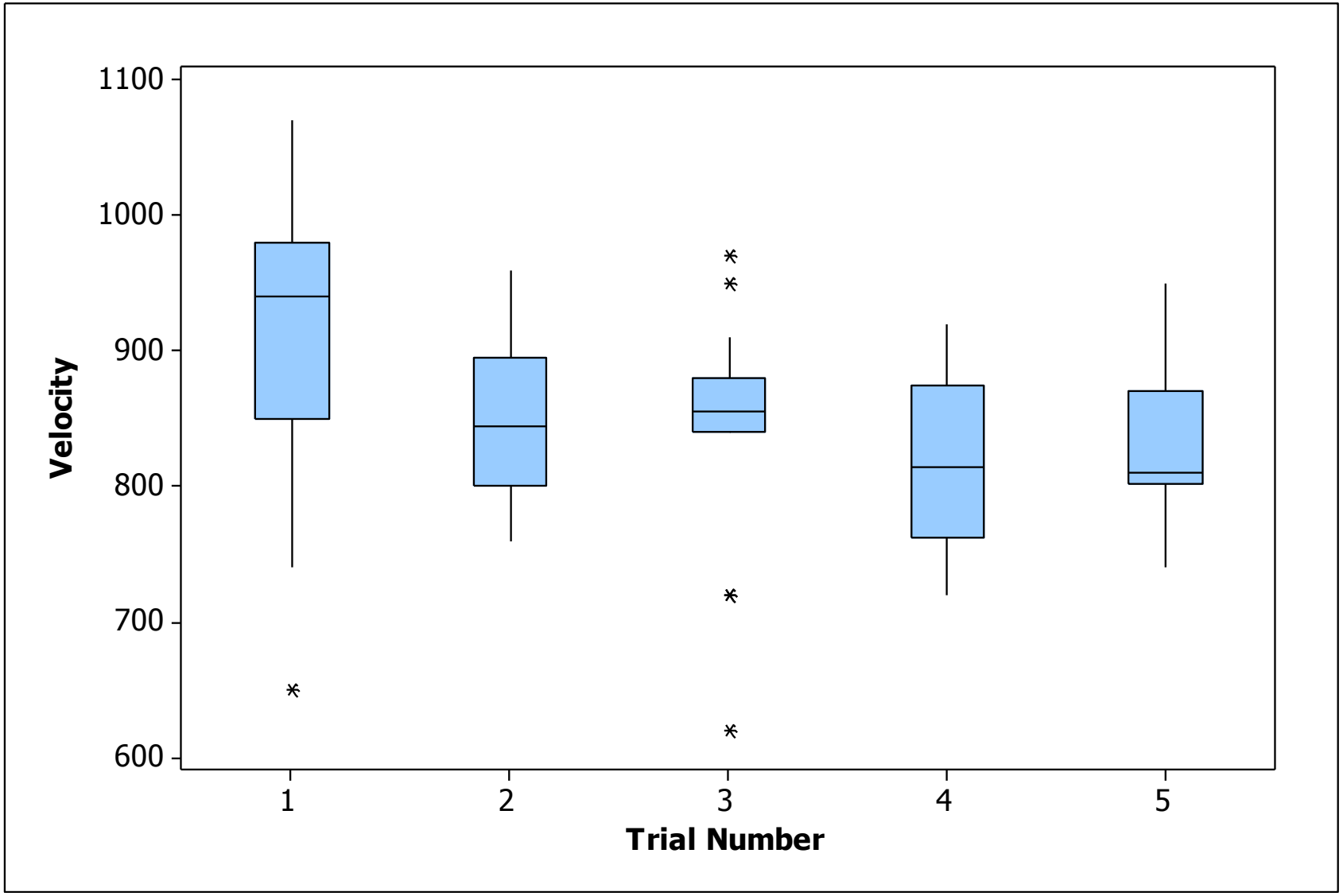
Means

- Trial 1: 909.0
- Trial 2: 856.0
- Trial 3: 845.9
- Trial 4: 820.5
- Trial 5: 831.5

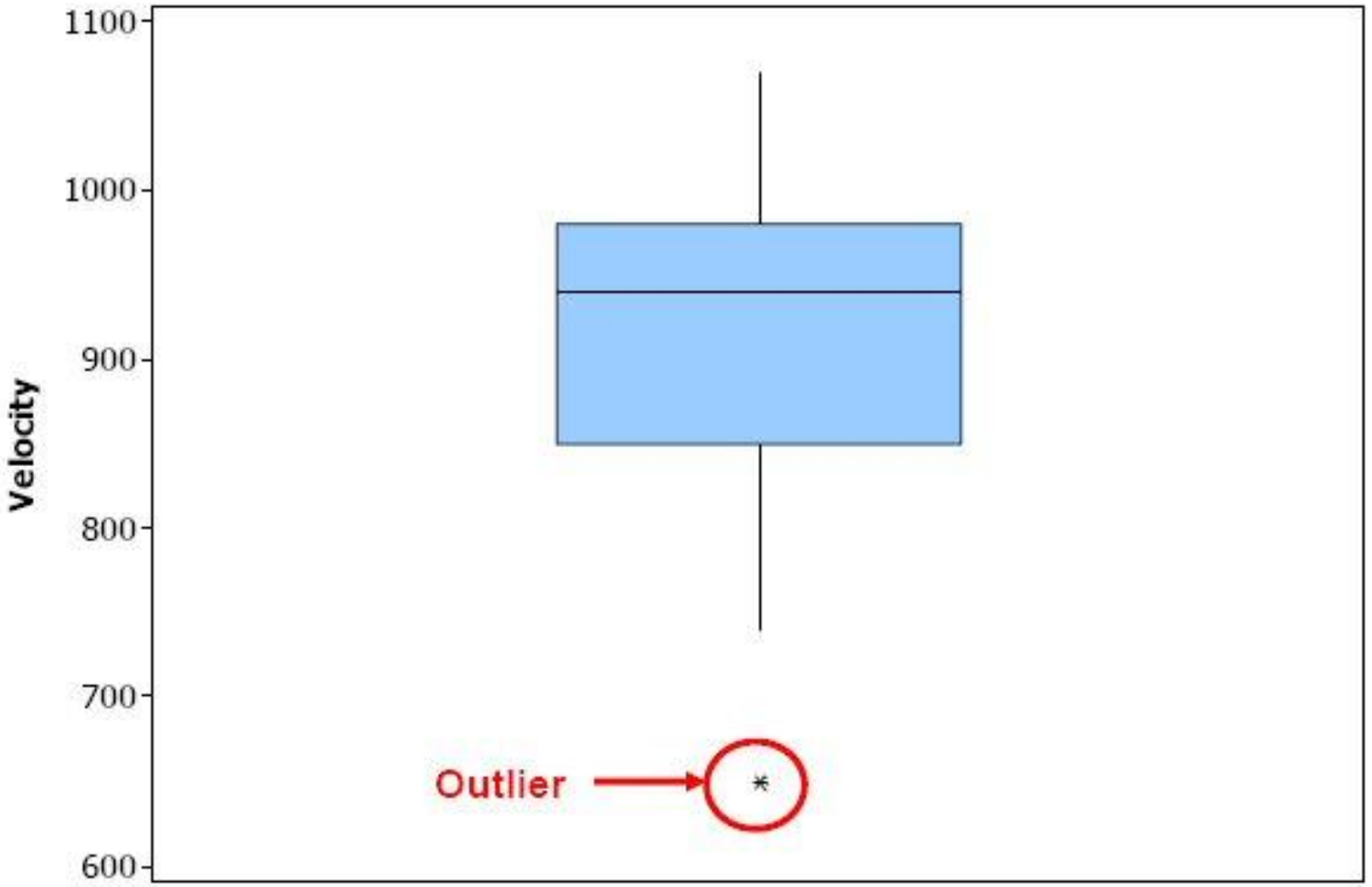
Standard Deviations

- Trial 1: 104.9
- Trial 2: 61.2
- Trial 3: 79.1
- Trial 4: 60.0
- Trial 5: 54.2

Speed of Light: All Five Trials



Michelson Data Trial 1



THE INTERQUARTILE RANGE *IQR*

The **interquartile range** IQR is the distance between the first and third quartiles:

$$IQR = Q_3 - Q_1$$

Trial 1: Five Number Summary

Minimum:

650

First Quartile:

850

Median:

940

Third Quartile:

980

Maximum:

1070

IQR:

980 – 850 = 130

THE 1.5 × *IQR* RULE FOR OUTLIERS

Call an observation a suspected outlier if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile.

Outlier Calculation: Trial 1

1.5 × IQR

- 1.5 × 130 = 195

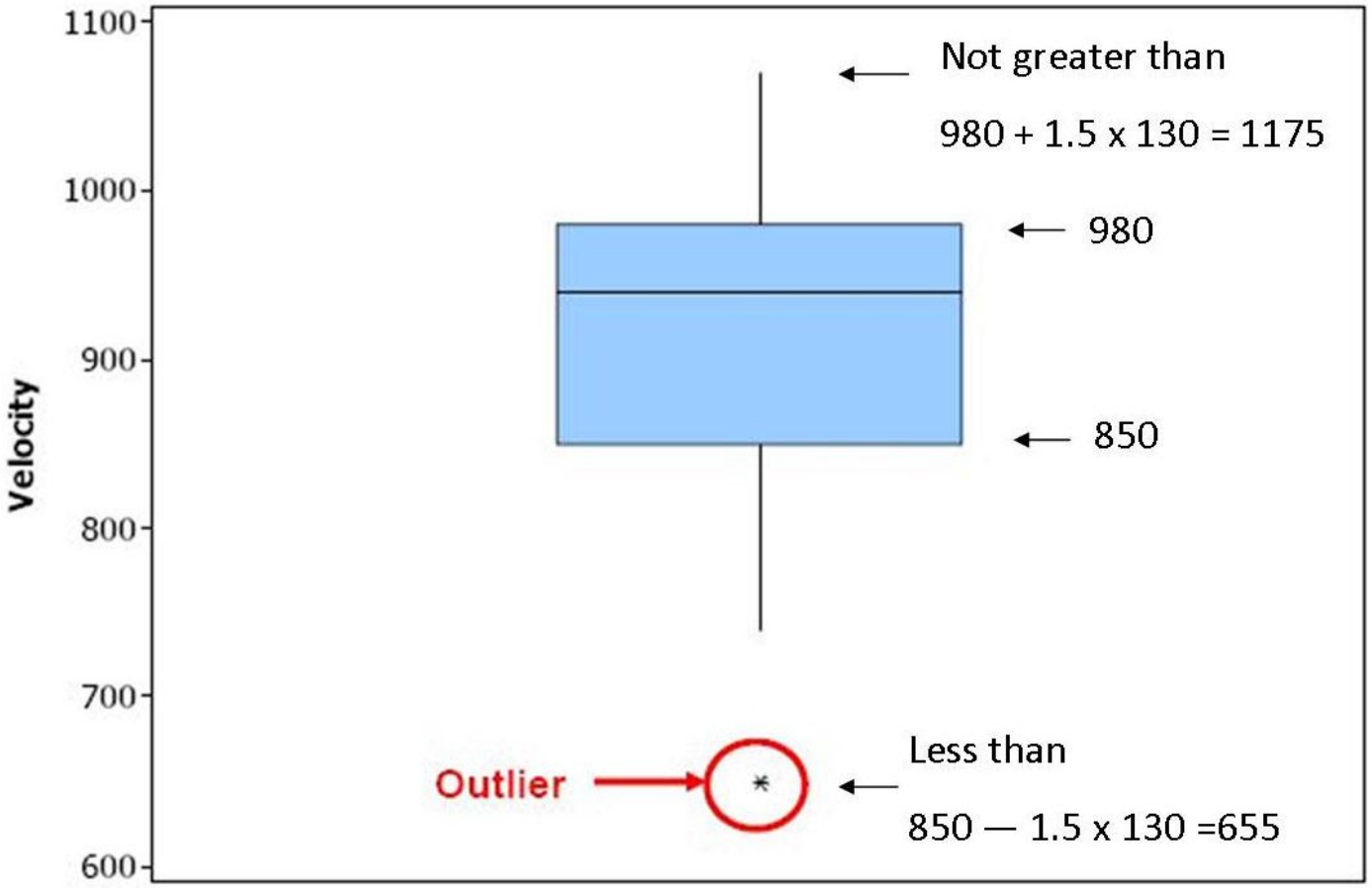
First Quartile: 850

- outlier if observation is less than 850 – 195 = 655 (Minimum = 650)

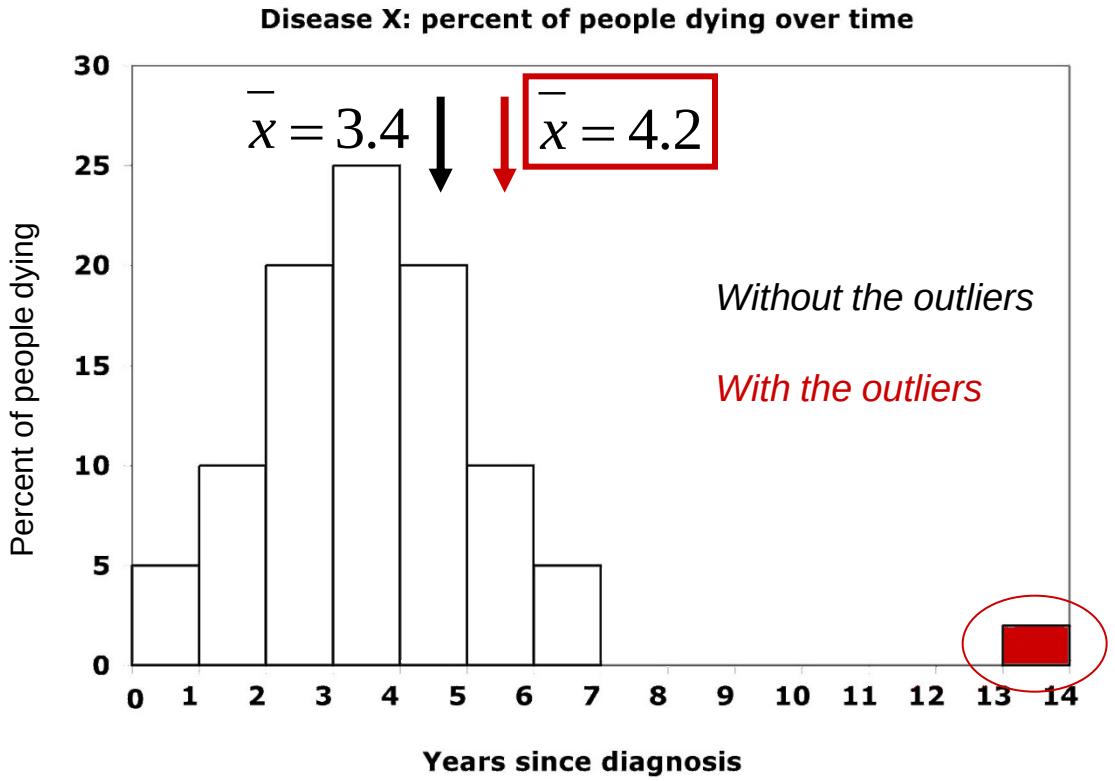
Third Quartile: 980

- outlier if observation is greater than 980 + 195 = 1175 (Maximum = 1070)

Michelson Data Trial 1 Again



Mean and median of a distribution with outliers



The mean is pulled to the right a lot by the outliers (from 3.4 to 4.2).

The median, on the other hand, is only slightly pulled to the right by the outliers (from 3.4 to 3.6).

Resistant Measures

- A **resistant measure** of any aspect of the distribution is relatively unaffected by changes in the numerical value of a small proportion of the total number of observations, no matter how large these changes are.
- The median and quartiles are resistant measures. The mean and standard deviation are NOT resistant measures.

ORGANIZING A STATISTICAL PROBLEM: A FOUR-STEP PROCESS

STATE: What is the practical question, in the context of the real-world setting?

FORMULATE: What specific statistical operations does this problem call for?

SOLVE: Make the graphs and carry out the calculations needed for this problem.

CONCLUDE: Give your practical conclusion in the setting of the real-world problem.

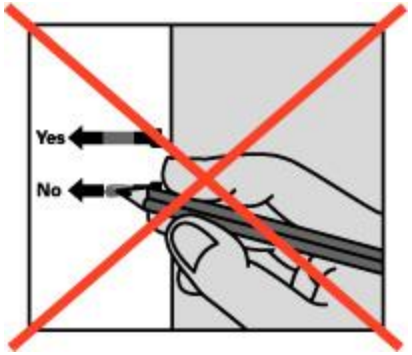
Undervotes and Overvotes in U.S. Elections

Undervotes

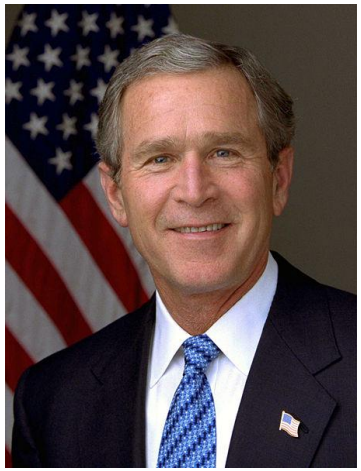
- If you do not cast a vote on a certain race or measure it is simply considered an undervote. An undervote does not invalidate the rest of your ballot. Those races and measures you do vote on will still count.

Overvotes

- If you vote for more candidates than allowed, or if you vote **both** Yes **and** No on a measure, it is called an overvote. **Your vote will not count for that candidate or measure.**



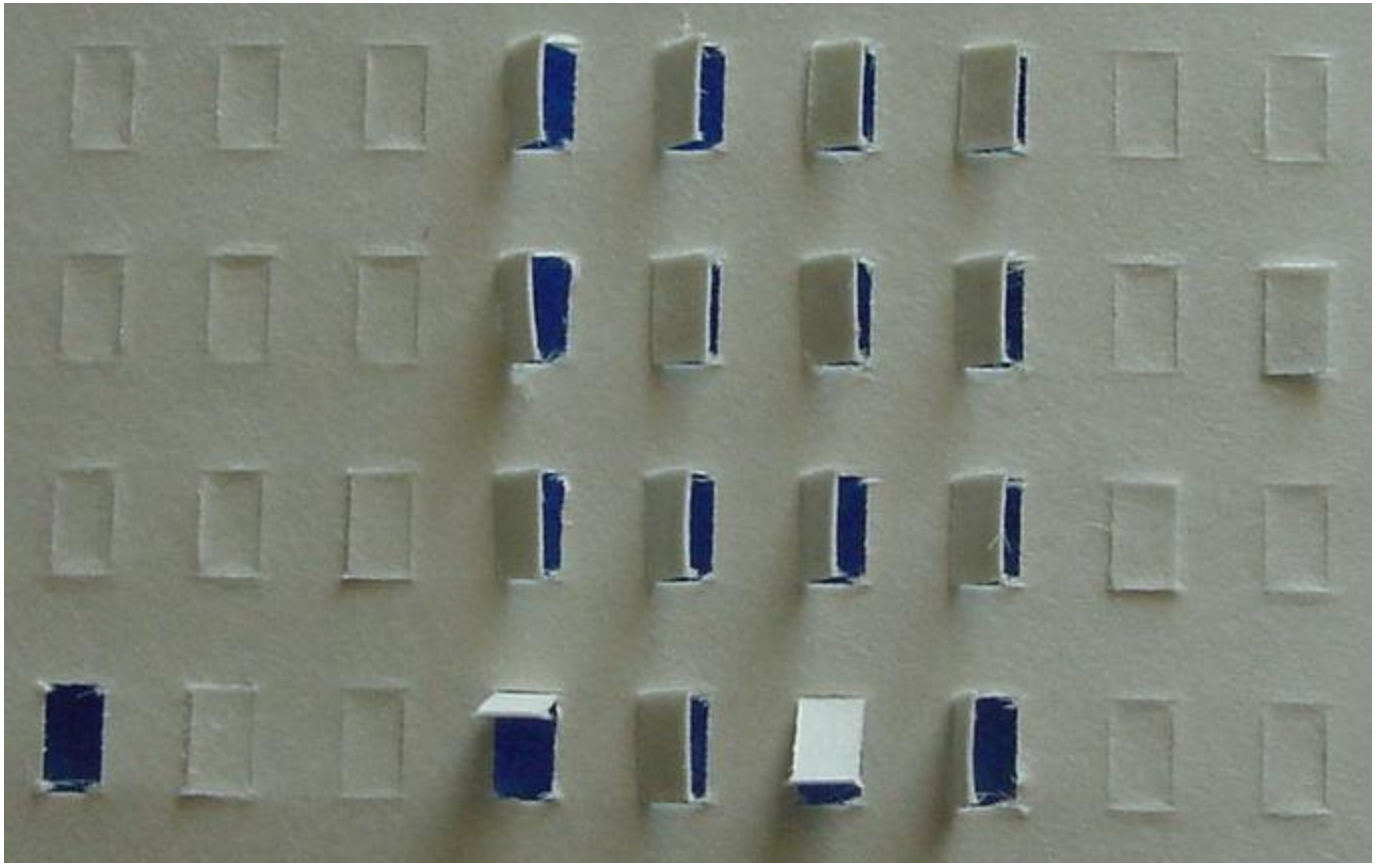
Florida 2000 Bush vs. Gore

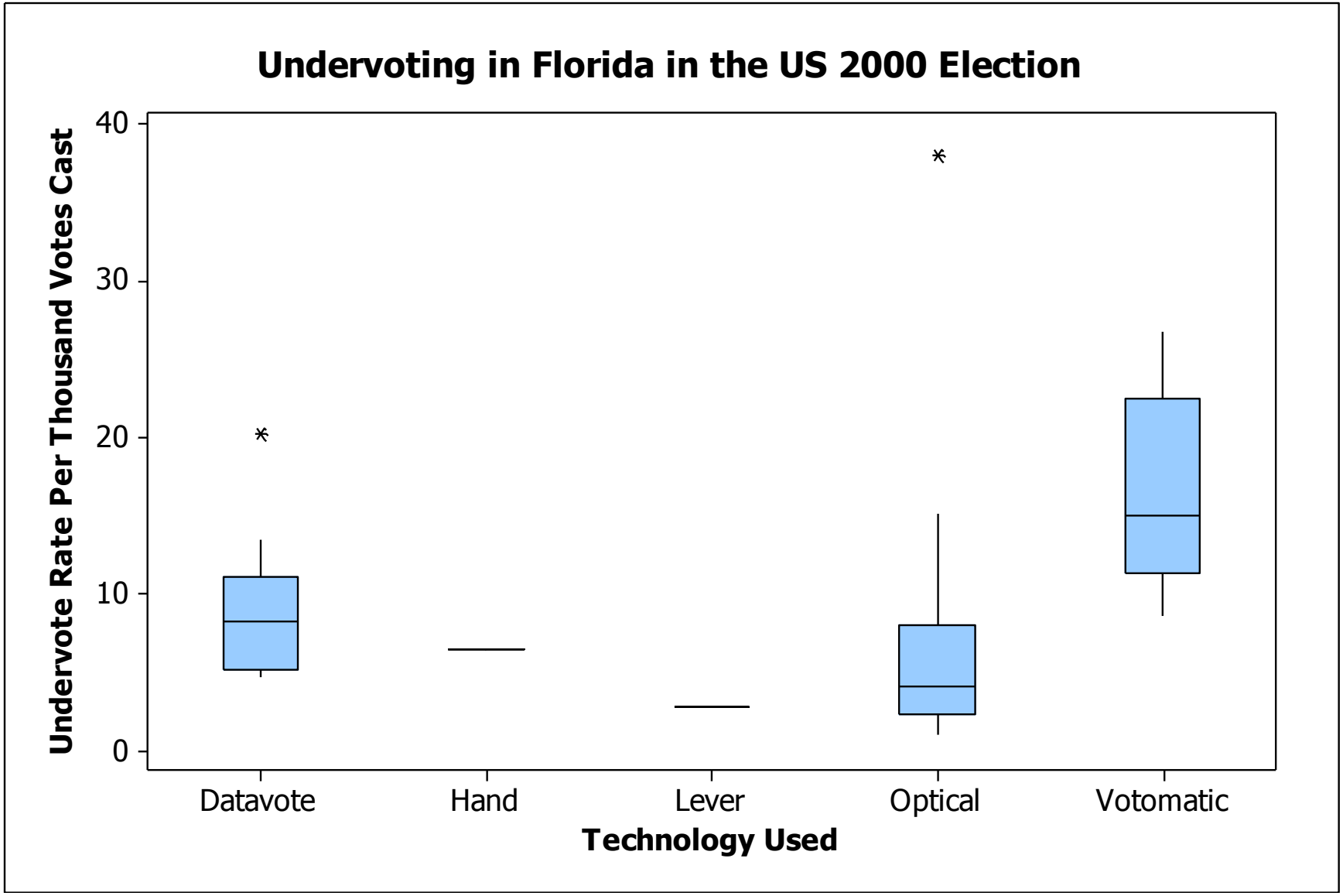


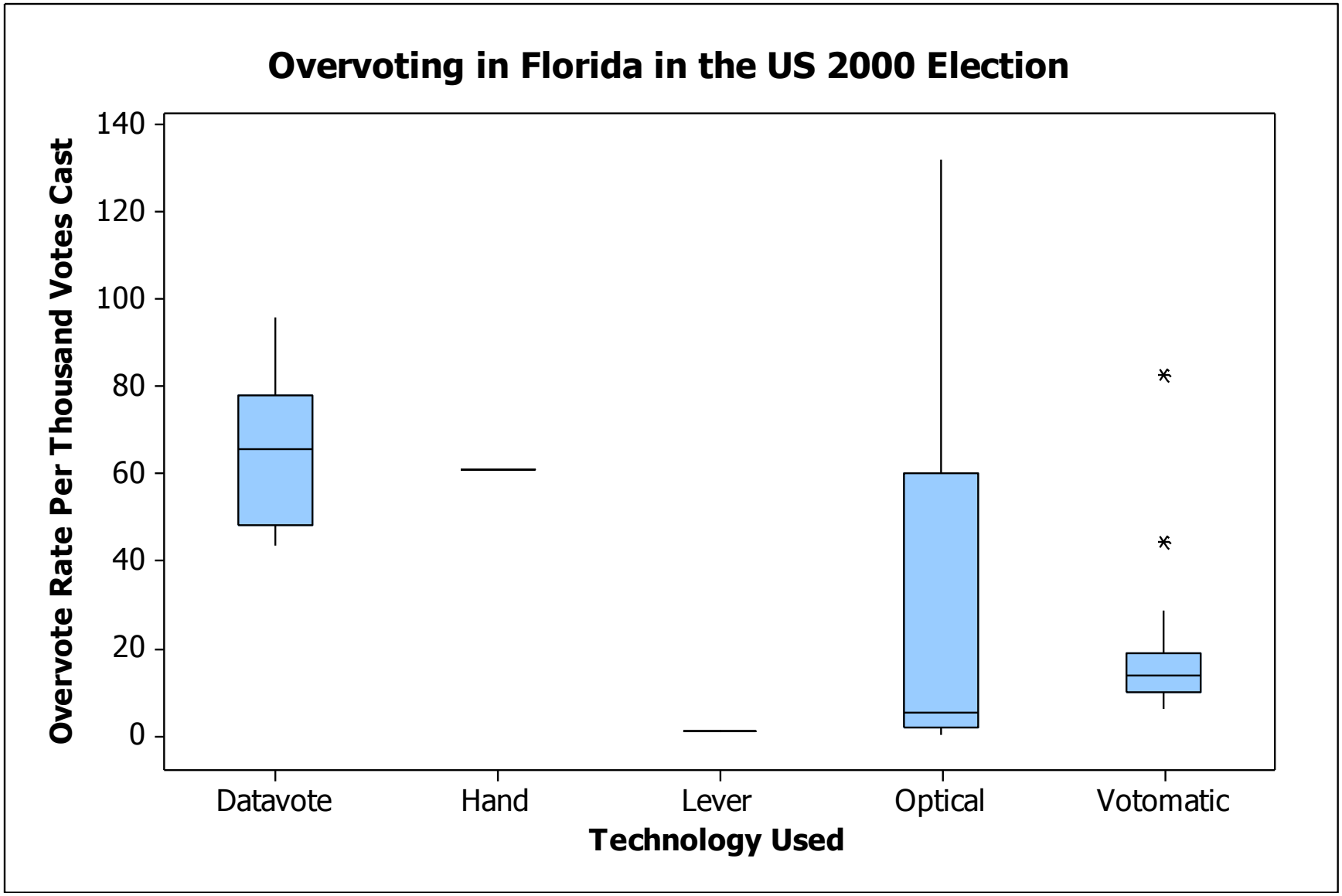
67 counties in the state
Four kinds of methods of voting

- hand ballot: 1 county
- lever method: 1 county
- optical scanning device: 41 counties
- punched cards
  - Votomatic: 15 counties
  - Datavote: 9 counties

Voting Problems: The Hanging Chad




Undervoting in Florida in the US 2000 Election
Undervote Rate Per Thousand Votes Cast
40
30
20
10
0
Datavote
Hand
Lever
Optical
Votomatic
Technology Used


Overvoting in Florida in the US 2000 Election
Overvote Rate Per Thousand Votes Cast
140
120
100
80
60
40
20
0
Datavote
Hand
Lever
Optical
Votomatic
Technology Used