Fundamentals of Business Statistics

Descriptive Statistics

Data versus Information

When managers are bewildered by plethora of data, which do not make any sense on the surface of it, they are looking for methods to classify data that would convey meaning. The idea here is to help them draw the right conclusion. This session provides the nitty-gritty of arranging data into information.

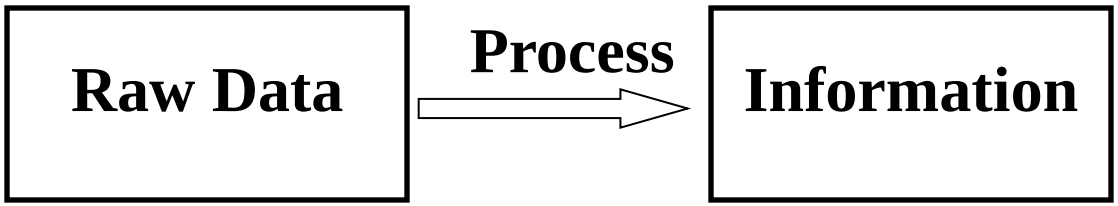
Raw Data

Meaning of Raw Data:

Raw Data represent numbers and facts in the original format in which the data have been Collected. You need to convert the raw data into information for managerial decision Making.

Information is Key

Large and massive raw data tend to bewilder you so much that the overall patterns are obscured. You cannot see the wood for the trees. This implies that the raw data must be processed to give you useful information.



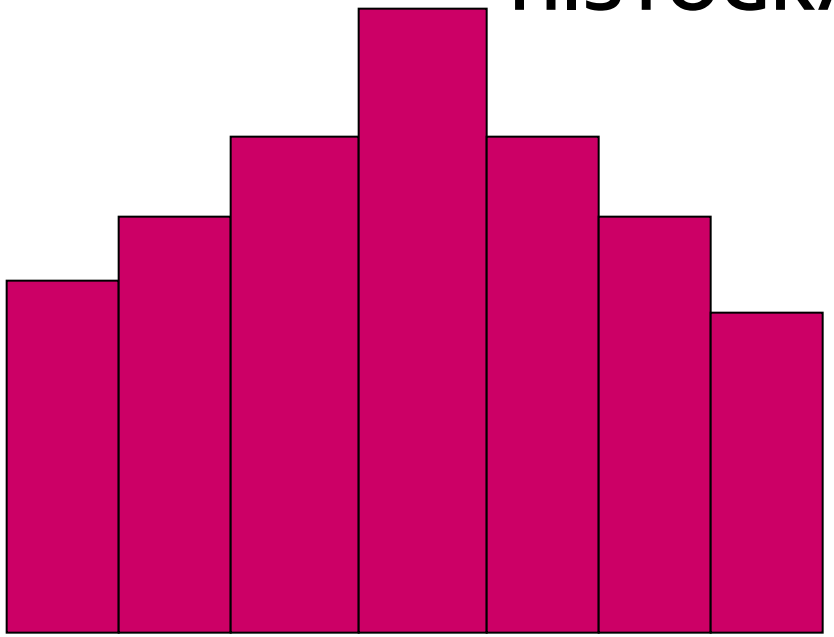
Frequency Distribution

In simple terms, frequency distribution is a summarized table in which raw data are arranged into classes and frequencies.

Frequency distribution focuses on classifying raw data into information. It is the most widely used data reduction technique in descriptive statistics.

When you are looking for pattern that would help you understand the characteristic you measure in a problem situation, frequency distribution comes to your rescue.

HISTOGRAM



Histogram (also known as frequency histogram) is a snap shot of the frequency distribution.

Histogram is a graphical representation of the frequency distribution in which the X-axis represents the classes and the Y-axis represents the frequencies in bars

Histogram depicts the pattern of the distribution emerging from the characteristic being measured.

Role of Histogram in Practice

Histogram - Uses

- Assessing material strengths
- Estimating process capabilities
- Indicating the necessity for corrective action
- Measuring the effects of corrective action
- Estimating machine capability
- Comparing operators, materials, vendors, and products

What is Central Tendency?

Whenever you measure things of the same kind, a fairly large number of such measurements will tend to cluster around the middle value. Such a value is called a measure of "Central Tendency". The other terms that are used synonymously are "Measures of Location", or "Statistical Averages".

Measures of Central Tendency

As a manager, You need the summary measures of central tendency to draw meaningful conclusions in your functional area of operation. The most widely used measures of central tendency are **Arithmetic Mean**, **Median**, and **Mode**.

Arithmetic Mean

Arithmetic Mean (called mean) is defined as the sum of all observations in a data set divided by the total number of observations. For example, consider a data set containing the following observations:

In symbolic form mean is given by $\overline{X} = \frac{\sum X}{n}$

$\overline{X}$ = Arithmetic Mean

$\sum X$ = Indicates sum all X values in the data set

n = Total number of observations(Sample Size)

Arithmetic Mean -Example

The inner diameter of a particular grade of tire based on 5 sample measurements are as follows: (figures in millimeters)

565, 570, 572, 568, 585

Applying the formula $\overline{X} = \frac{\sum X}{n}$

We get mean = (565+570+572+568+585)/5 =572

Caution: Arithmetic Mean is affected by extreme values or fluctuations in sampling. It is not the best average to use when the data set contains extreme values (Very high or very low values).

Median

Median is the middle most observation when you arrange data in ascending order of magnitude. Median is such that 50% of the observations are above the median and 50% of the observations are below the median.

Median is a very useful measure for ranked data in the context of consumer preferences and rating. It is not affected by extreme values (greater resistance to outliers)

$$\text{Median} = \frac{n+1}{2} \text{ th value of ranked data}$$

n = Number of observations in the sample

Median - Example

Marks obtained by 7 students in Computer Science Exam are given below: Compute the median.

45 40 60 80 90 65 55

Arranging the data after ranking gives

90 80 65 60 55 45 40

Median = (n+1)/2 th value in this set = (7+1)/2 th observation= 4^{th} observation=60

Hence Median = 60 for this problem.

Mode

Mode is that value which occurs most often. It has the maximum frequency of occurrence. Mode also has resistance to outliers.

Mode is a very useful measure when you want to keep in the inventory, the most popular shirt in terms of collar size during festival season.

Caution: In a few problems in real life, there will be more than one mode such as bimodal and multi-modal values. In these cases mode cannot be uniquely determined.

Mode -Example

The life in number of hours of 10 flashlight batteries are as follows: Find the mode.

340 350 340 340 320 340 330 330
340 350

340 occurs five times. Hence, mode=340.

Comparison of Mean, Median, Mode

Mean	Median	Mode
Defined as the arithmetic average of all observations in the data set.	Defined as the middle value in the data set arranged in ascending or descending order.	Defined as the most frequently occurring value in the distribution; it has the largest frequency.
Requires measurement on all observations.	Does not require measurement on all observations	Does not require measurement on all observations
Uniquely and comprehensively defined.	Cannot be determined under all conditions.	Not uniquely defined for multi-modal situations.

Comparison of Mean, Median, Mode Cont.

Mean	Median	Mode
Affected by extreme values. Can be treated algebraically. That is, Means of several groups can be combined.	Not affected by extreme values. Cannot be treated algebraically. That is, Medians of several groups cannot be combined.	Not affected by extreme values. Cannot be treated algebraically. That is, Modes of several groups cannot be combined.

Measures of Dispersion

In simple terms, measures of dispersion indicate how large the spread of the distribution is around the central tendency. It answers unambiguously the question " What is the magnitude of departure from the average value for different groups having identical averages?".

Range

Range is the simplest of all measures of dispersion. It is calculated as the difference between maximum and minimum value in the data set.

Range =

$$X_{\text{Maximum}} - X_{\text{Minimum}}$$

Range-Example

Example for Computing Range

The following data represent the percentage return on investment for 10 mutual funds per annum. Calculate Range.

12, 14, 11, 18, 10.5, 11.3, 12, 14, 11, 9

Range = $X_{Maximum} - X_{Minimum}$ = 18-9=9

Caution: If one of the components of range namely the maximum value or minimum value becomes an extreme value, then range should not be used.

Inter-Quartile Range(IQR)

IQR= Range computed on middle 50% of the observations after eliminating the highest and lowest 25% of observations in a data set that is arranged in ascending order. IQR is less affected by outliers.

$IQR = Q_3 - Q_1$

Interquartile Range-Example

The following data represent the percentage return on investment for 9 mutual funds per annum. Calculate interquartile range.

Data Set: 12, 14, 11, 18, 10.5, 12, 14, 11, 9

Arranging in ascending order, the data set becomes

9, 10.5, 11, 11, 12, 12, 14, 14, 18

$IQR=Q_3-Q_1=14-10.75=3.25$

Standard Deviation

Standard deviation forms the cornerstone for Inferential Statistics.

To define standard deviation, you need to define another term called variance. In simple terms, standard deviation is the square root of variance.

Key Formulas

Important Terms with Notations

Sample Variance

$$S^2 = \frac{\sum (X - \overline{X})^2}{n-1}$$

Sample Standard Deviation

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

Population Variance

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Population Standard Deviation $\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$

Where $\overline{X} = \frac{\sum X}{n}$ (Sample Mean) and

$\mu = \frac{\sum X}{N}$ (Population Mean)

n =Number of observations in the sample(Sample size)

N =Number of observations in the Population (Population Size)

Remarks

1. $S^2 = \frac{\sum (X - \overline{X})^2}{n-1}$ is an unbiased estimator of $\sigma^2 = \frac{\sum (X - \mu)^2}{N}$
2. $\overline{X} = \frac{\sum X}{n}$ is an unbiased estimator of $\mu = \frac{\sum X}{N}$
3. The divisor n-1 is always used while calculating sample variance for ensuring property of being unbiased
4. Standard deviation is always the square root of variance

Example for Standard Deviation

The following data represent the percentage return on investment for 10 mutual funds per annum. Calculate the sample standard deviation.

12, 14, 11, 18, 10.5, 11.3, 12, 14, 11, 9

Solution for the Example

A	B	C	D
1			
2	X	$X - \overline{X}$	$(X - \overline{X})^2$
3	12	-0.28	0.08
4	14	1.72	2.96
5	11	-1.28	1.64
6	18	5.72	32.72
7	10.5	-1.78	3.17
8	11.3	-0.98	0.96
9	12	-0.28	0.08
10	14	1.72	2.96
11	11	-1.28	1.64
12	9	-3.28	10.76
13	Mean =		56.96
14	12.28	Variance=	6.33
15		Standard Deviation=	2.52

Solution for the Example Cont.

From the spreadsheet of Microsoft Excel in the previous slide, it is easy to see

that Mean = $\bar{X} = \frac{\sum X}{n}$ =12.28 (In column A and row14, 12.28 is seen).

Sample Variance = $S^2 = \frac{\sum (X - \bar{X})^2}{n-1}$ 6.33 (In column D and row 14, 6.33 is seen)

Sample Standard Deviation = S = $\sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$ = 2.52

(In column D and row 15, 2.52 is seen)

Coefficient of Variation Relative Dispersion)

Coefficient of Variation (CV) is defined as the ratio of Standard Deviation to Mean.

In symbolic form

CV = $\frac{S}{\overline{X}}$ for the sample data and = $\frac{\sigma}{\mu}$ for the population data.

Coefficient of Variation Example

Consider two Sales Persons working in the same territory. The sales performance of these two in the context of selling PCs are given below. Comment on the results.

Sales Person 1	**Sales Person 2**
Mean Sales (One year average) 50 units	Mean Sales (One year average)75 units
Standard Deviation 5 units	Standard deviation 25 units

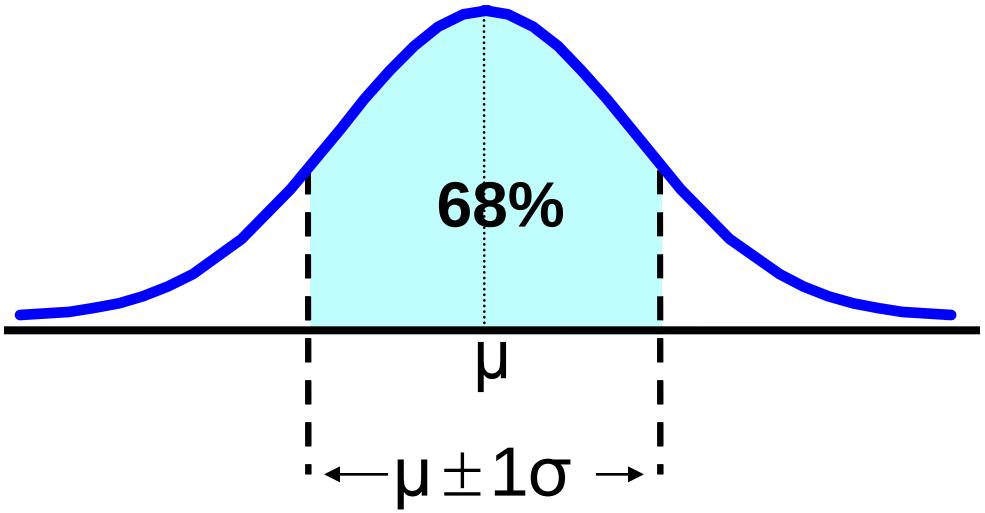
Interpretation for the Example

The CV is 5/50 =0.10 or 10% for the Sales Person1 and 25/75=0.33 or 33% for sales Person2.

The moral of the story is "don't get carried away by absolute number". Look at the scatter. Even though, Sales Person2 has achieved a higher average, his performance is not consistent and seems erratic.

The Empirical Rule

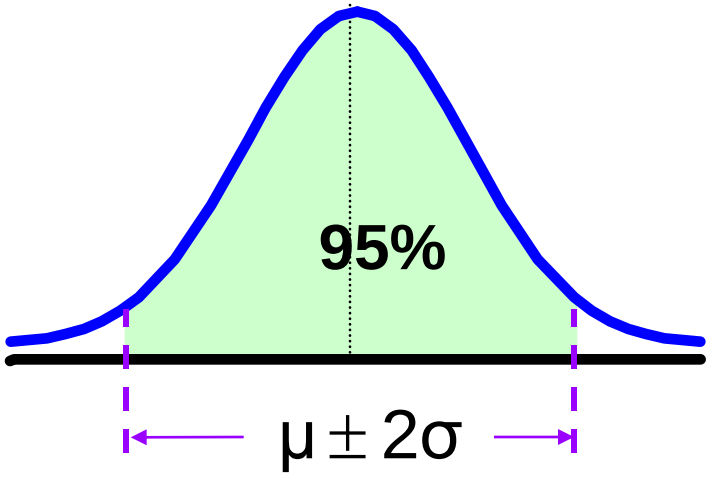
- The empirical rule approximates the variation of data in a bell-shaped distribution
- Approximately 68% of the data in a bell shaped distribution is within 1 standard deviation of the mean or $\mu \pm 1\sigma$

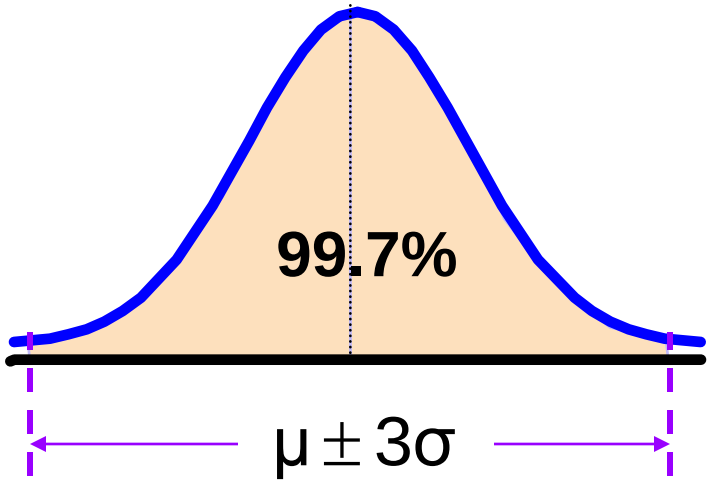


The Empirical Rule

- Approximately 95% of the data in a bell-shaped distribution lies within two standard deviations of the mean, or $\mu \pm 2\sigma$

- Approximately 99.7% of the data in a bell-shaped distribution lies within three standard deviations of the mean, or $\mu \pm 3\sigma$





The Five Number Summary

The five numbers that help describe the center, spread and shape of data are:

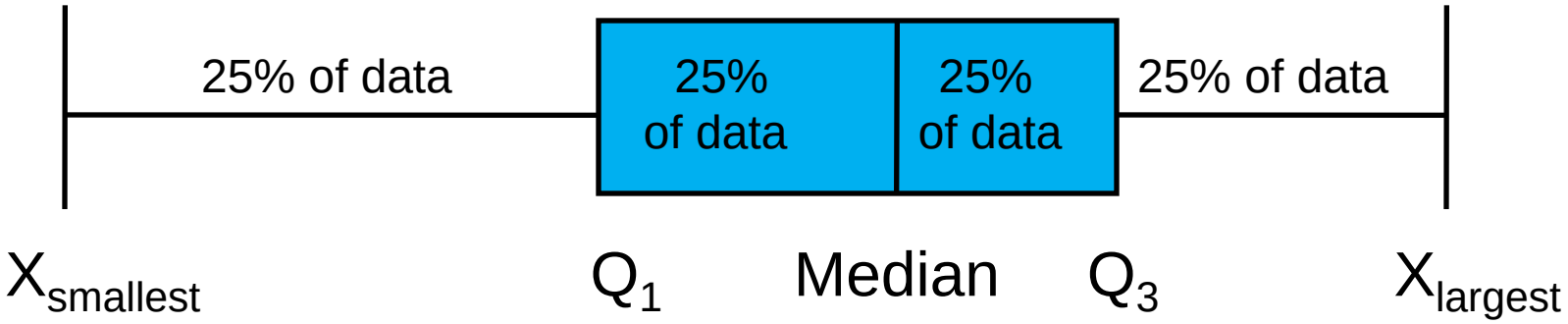
- $X_{smallest}$
- First Quartile (Q_1)
- Median (Q_2)
- Third Quartile (Q_3)
- $X_{largest}$

Five Number Summary and The Boxplot

- The Boxplot: A Graphical display of the data based on the five-number summary:

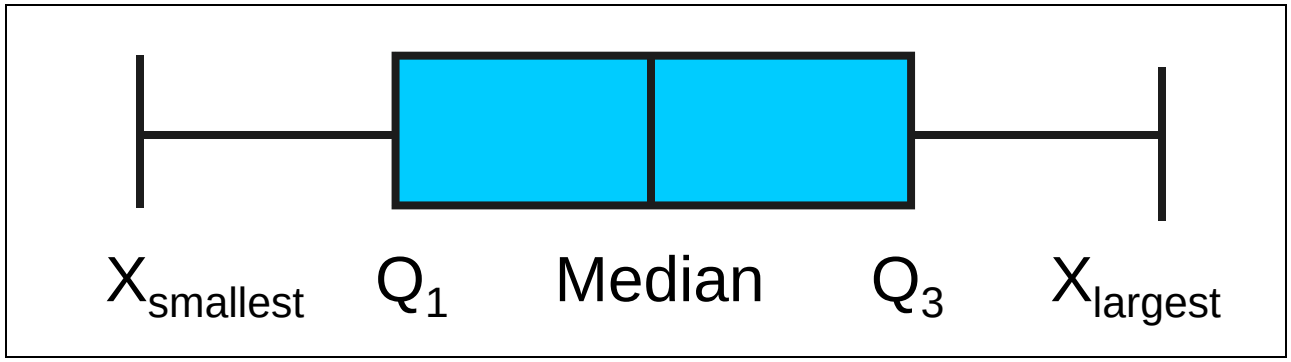
  $X_{smallest}$ -- Q_1 -- Median -- Q_3 -- $X_{largest}$

Example:



Five Number Summary: Shape of Boxplots

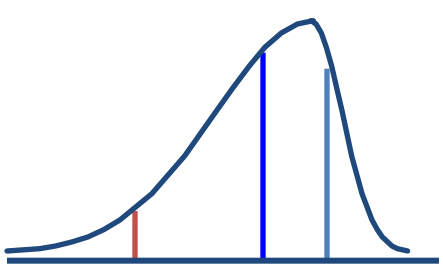
- If data are symmetric around the median then the box and central line are centered between the endpoints



- A Boxplot can be shown in either a vertical or horizontal orientation

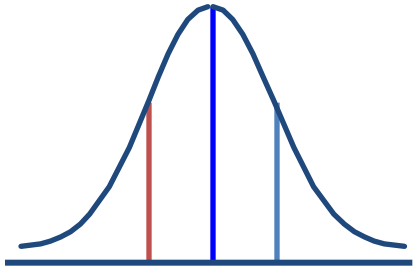
Distribution Shape and The Boxplot

Left-Skewed

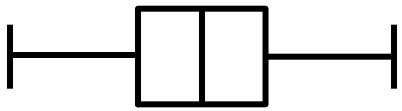


Q_1 Q_2 Q_3

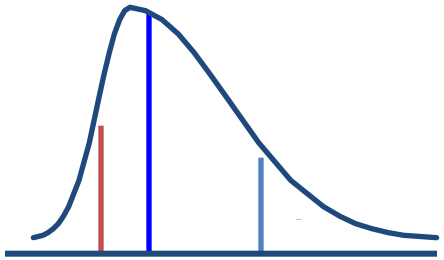
Symmetric



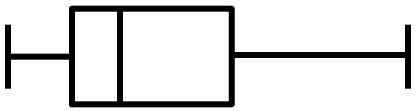
Q_1 Q_2 Q_3



Right-Skewed



Q_1 Q_2 Q_3



Boxplot Example

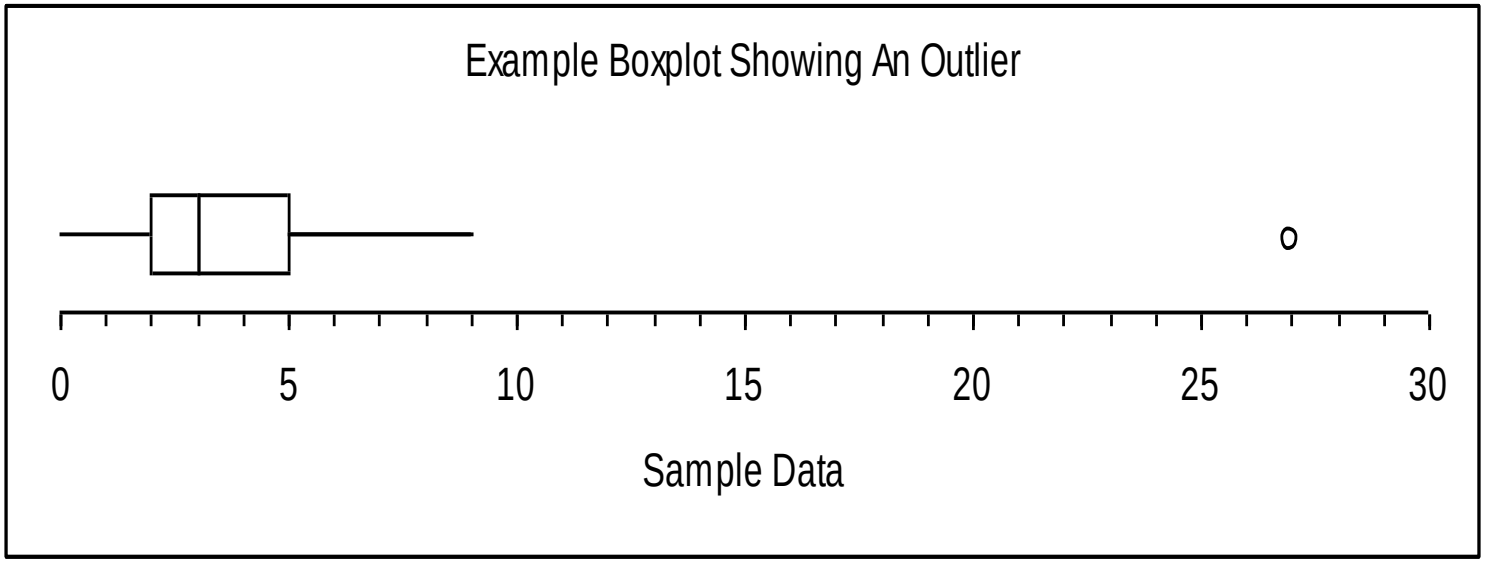
Below is a Boxplot for the following data:

$X_{smallest}$ Q_1 Q_2 Q_3 $X_{largest}$

0 2 2 2 3 3 4 5 5 9 27

Box plot example showing an outlier

- The boxplot below of the same data shows the outlier value of 27 plotted separately
- A value is considered an outlier if it is more than 1.5 times the interquartile range below Q_1 or above Q_3



Descriptive Statistics- CardioGood Fitness

The market research team at AdRight is assigned the task to identify the profile of the typical customer for each treadmill product offered by CardioGood Fitness. The market research team decides to investigate whether there are differences across the product lines with respect to customer characteristics. The team decides to collect data on individuals who purchased a treadmill at a CardioGood Fitness retail store during the prior three months. The data are stored in the CardioGoodFitness.csv file.

Descriptive Statistics- CardioGood Fitness

The team identifies the following customer variables to study: product purchased, TM195, TM498, or TM798; gender; age, in years; education, in years; relationship status, single or partnered; annual household income ($); average number of times the customer plans to use the treadmill each week; average number of miles the customer expects to walk/run each week; and self-rated fitness on an 1-to-5 scale, where 1 is poor shape and 5 is excellent shape.

Perform descriptive analytics to create a customer profile for each CardioGood Fitness treadmill product line.