Statistics & Probability

Chapter 6: DESCRIPTIVE STATISTICS

FPT University

Department of Mathematics

Quy Nhon, 2023

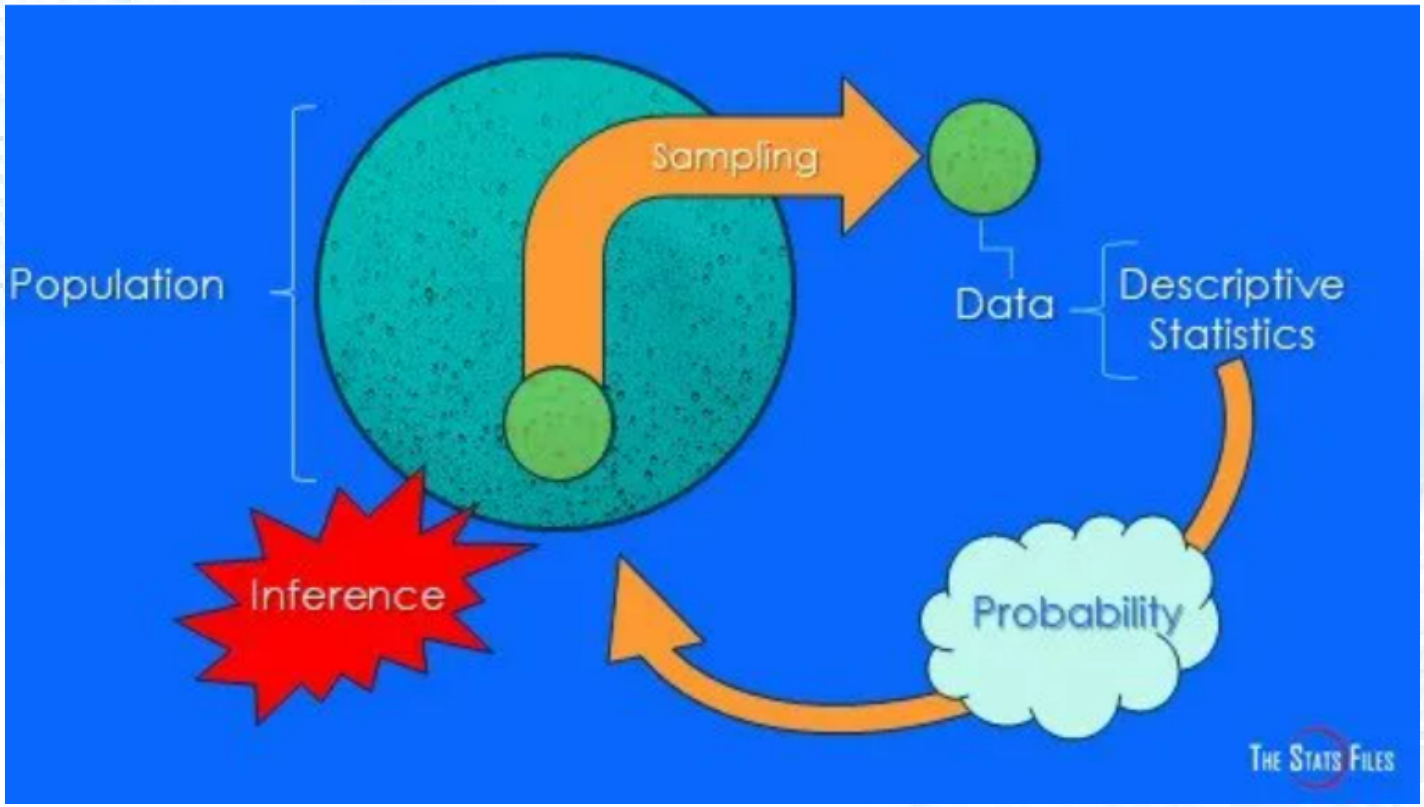
Table of Contents

Table of Contents

What is Statistics?

Statistics is the science of collecting, organizing, analyzing, and interpreting **DATA** in order to make decisions.

- *Descriptive Statistics*: Involves organizing, summarizing, and displaying data. **Example**. Tables, charts,. . .
- *Inferential Statistics*: Involves using sample data to draw conclusions about a population (Chapter 8, 9, 10, 11).



Big picture of Statistics

Sample Mean

Let $x_1, x_2, \ldots, x_n$ be n observations in a sample. Then, the **sample mean** is

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Example (Sample Mean)

Let's consider the eight observations on pull-off force collected from the prototype engine connectors: $12.6,\ 12.9,\ 13.4,\ 12.3,\ 13.6,\ 13.5,\ 12.6$ and 13.1. Then,

$$\overline{x} = \frac{12.6 + 12.9 + 13.4 + 12.3 + 13.6 + 13.5 + 12.6 + 13.1}{8} = 13.0 \text{ (pounds)}.$$

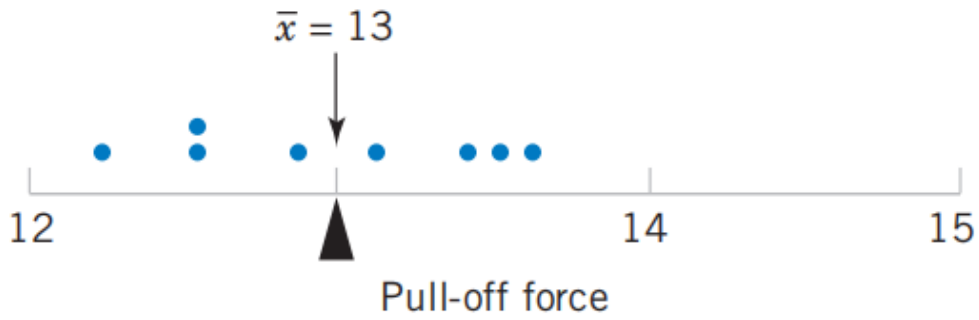


Figure. The sample mean as a balance point for a system of weights.

Population Mean

For a finite population with N equally likely values, $x_1, x_2, \ldots, x_N$, and the probability mass function (PMF) is $f(x_i) = \dfrac{1}{N}$; the **population mean** is

$$\mu = \sum_{i=1}^{N} x_i f(x_i) = \frac{\sum_{i=1}^{N} x_i}{N}.$$

Sample Median

1. The **median** is the middle point in a data set - half of the data points are smaller than the median and half of the data points are larger.
2. To find the median:
   1. Arrange the data points from smallest to largest.
   2. If the number of data points is odd, the median is the middle data point in the list.
   3. If the number of data points is even, the median is the average of the two middle data points in the list.

Example (Sample Median)

Find the median of the following data:

1. $1,\ 4,\ 2,\ 5,\ 0.$

   First, reorder this data, we get $0,\ 1,\ 2,\ 4,\ 5.$ There is an odd number of data points, so the median is the middle data point, i.e., $2.$

2. $10,\ 40,\ 20,\ 50.$

   Put the data in order first, $10,\ 20,\ 40,\ 50.$ There is an even number of data points, so the median is the average of the two middle data points, $\dfrac{20+40}{2} = 30.$

Quiz (Sample Median)

The following data points represent the number of points scored by each player on the W.D. basketball team last game

$$8,\ 5,\ 8,\ 4,\ 8,\ 12,\ 13,\ 5,\ 9.$$

Quiz (Sample Median)

The prices (in dollars) for a sample of round-trip flights from Chicago, Illinois to Cancun, Mexico are listed. Find the median of the flight prices

$$872,\ 432,\ 397,\ 427,\ 388,\ 782,\ 397.$$

Sample Mode

1. The **mode** is the most commonly occurring data point in a dataset.
2. There can be no mode, one mode, or multiple modes in a dataset. For instance,
   + if no entry is repeated the data set has no mode.
   + if two/ three/ four or more entries occur with the same greatest frequency, each entry is a mode and the set is called bimodal/ trimodal/ multimodal.

Example (Sample Mode)

Find the mode of the following data:

1. $0,\ 0,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 2,\ 2,\ 2,\ 3,\ 5.$
   Look for the value that occurs the most, that is 1. Thus, the mode is 1.
2. $0,\ 0,\ 0,\ 1,\ 1,\ 1,\ 1,\ 2,\ 2,\ 2,\ 2,\ 3.$
   Look for the values that occur the most, that are 1 and 2. Hence, the modes are 1 and 2.

Quiz

At a political debate a sample of audience members was asked to name the political party to which they belong. Their responses are shown in the table. What is the mode of the responses?

Political Party	**Frequency**
Democrat	35
Republican	60
Other	25
Did not respond	8

Sample Variance and Sample Standard Deviation

Let $x_1, x_2, \ldots, x_n$ be a sample of n observations.

1. **Sample variance** is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{n\sum_{i=1}^{n} x_i{}^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}.$$

2. **Sample standard deviation** s, is the positive square root of the sample variance.

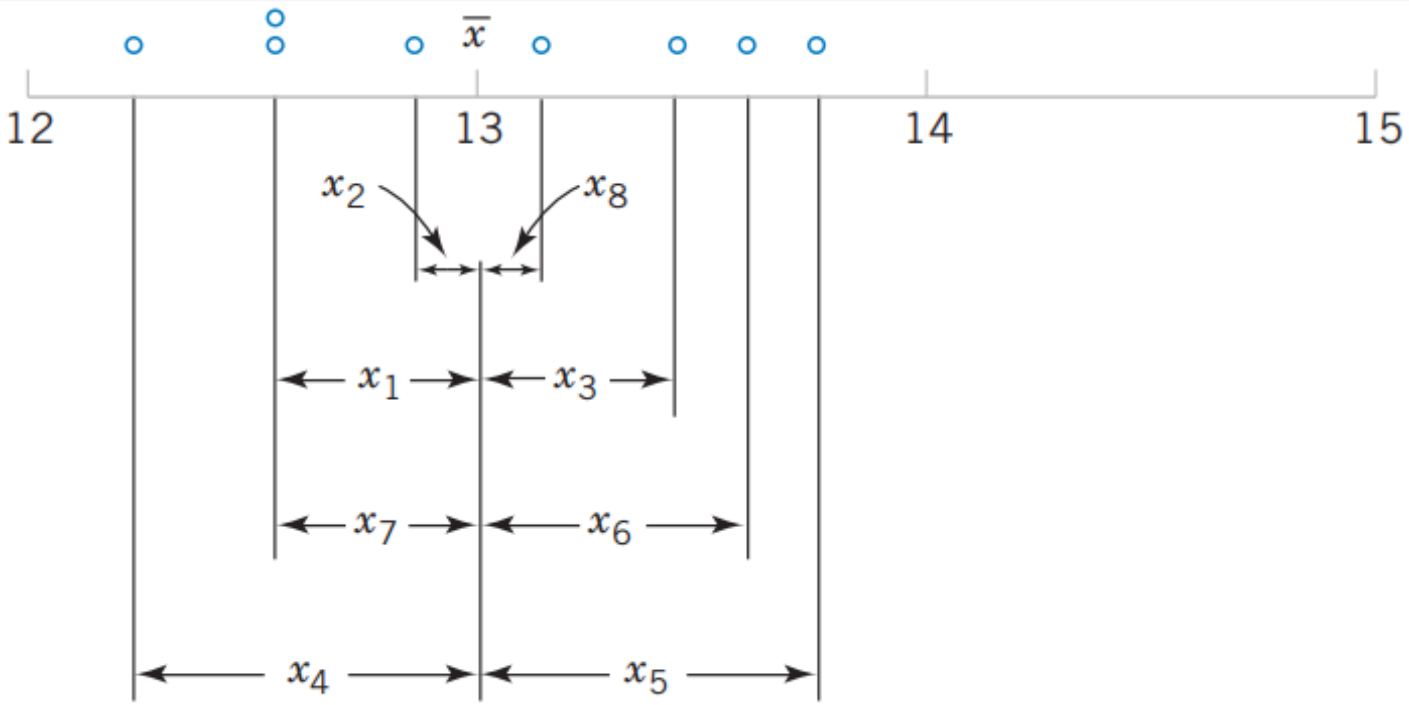


Figure. How the sample variance measures variability through the deviations $x_i - \overline{x}$.

Example (Sample Variance and Sample Standard Deviation)

The following table displays the quantities needed for calculating the sample variance and sample standard deviation for the pull-off force data.

i	x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
1	12.6	-0.4	0.16
2	12.9	-0.1	0.01
3	13.4	0.4	0.16
4	12.3	-0.7	0.49
5	13.6	0.6	0.36
6	13.5	0.5	0.25
7	12.6	-0.4	0.16
8	13.1	0.1	0.01
Sum	104.0	0.0	1.60

Solution. The sample variance is $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \dfrac{1.60}{8-1} = 0.2286$ (pounds)^2.

The sample standard deviation is $s = \sqrt{0.2286} = 0.48$ (pounds).

Recall the sample variance is $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \dfrac{n\sum_{i=1}^{n} x_i{}^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}$.

Quiz

Let's consider the weight of the eight observations collected from the prototype engine connectors: $12,\ 13,\ 9,\ 12,\ 10$ and 12. Find the sample standard deviation.

Population Variance and Population Standard Deviation

Let $x_1, x_2, \ldots, x_N$ be a population of N observations.

1. **Population variance** is
$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}.$$
2. **Population standard deviation**, σ, is the positive square root of the population variance.

Sample Range

Let $x_1, x_2, \ldots, x_n$ be a sample of n observations. The **sample range** is

$$r = \max_{1 \le i \le n}(x_i) - \min_{1 \le i \le n}(x_i).$$

FUQN

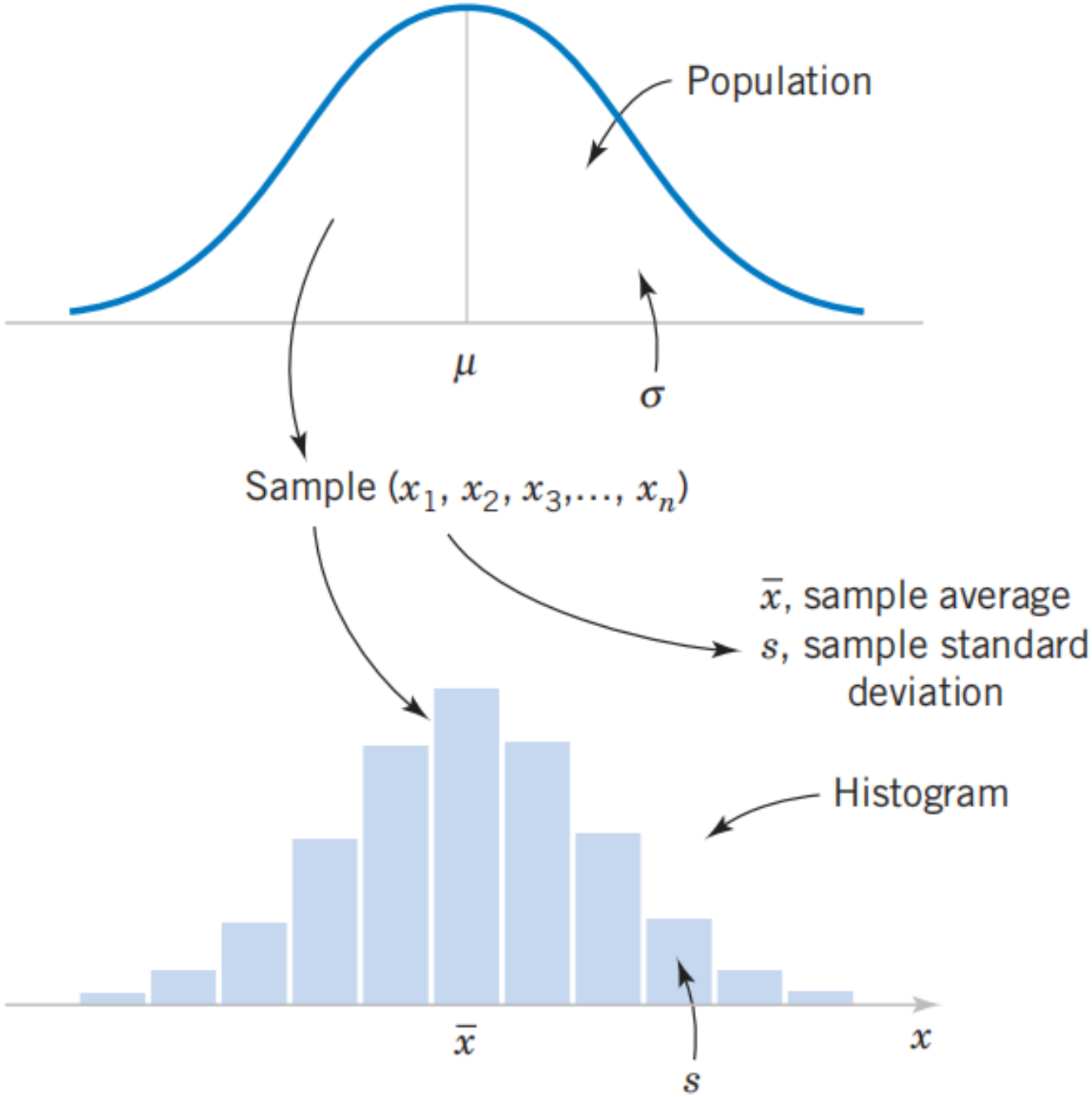


Figure. Relationship between a population and sample.

Table of Contents

Stem-and-Leaf Diagrams

Steps to Construct a Stem-and-Leaf Diagram

1. Classify the data values in terms of the number of digits in each value, such as 2 digit numbers or 3 digit numbers.
2. Fix the key for the stem and leaf plot. For instance, $2 \mid 5 = 25,\ 3 \mid 2 = 3.2$ or $19 \mid 2 = 192$.
3. Consider the first digits as stems and the last digit as leaves.
4. Find the range of the data, that is the lowest and the highest values among the data.
5. Draw a vertical line. Place the stem on the left and the leaf on the right of the vertical line.
6. List the stems in the stem column. Sort them in ascending order.
7. List the leaf values in the column against the stem from lowest to the highest horizontally.

Example (Stem-and-Leaf Diagrams)

Construct a stem-and-leaf plot for the data which is the listening scores of 12 students in a TOEIC test are listed below:

$$55 \;\; 115 \;\; 225 \;\; 240 \;\; 330 \;\; 335 \;\; 385 \;\; 400 \;\; 405 \;\; 405 \;\; 495 \;\; 495.$$

Solution. The stem-and-leaf diagram:

Stem	**Leaf**
5	5
11	5
22	5
24	0
33	0 5
38	5
40	0 5 5
49	5 5

Key: 5 | 5 means 55.

Example (Stem-and-Leaf Diagrams)

The stem-and-leaf plot below shows the quiz scores of students.

a. Find the number of students who scored less than 9 points?

b. Find the number of students who scored a minimum of 9 points?

Stem	**Leaf**
6	6
7	0 5 7 8
8	1 1 3 4 4 6 8 8 9
9	0 2 9
10	0

Key: $7 \mid 5 = 7.5$ points

Solution.

a. There are fourteen scores less than 9 points. They are

$$6.6,\ 7.0,\ 7.5,\ 7.7,\ 7.8,\ 8.1,\ 8.1,\ 8.3,\ 8.4,\ 8.4,\ 8.6,\ 8.8,\ 8.8 \text{ and } 8.9.$$

b. There are four scores which are at least 9 points. They are

$$9.0,\ 9.2,\ 9.9,\ \text{and } 10.0.$$

Quiz

Given a stem-and-leaf diagram as follows:

Stem	Leaf
1	2 4
2	1 5 8
3	2 4 6
5	0 3 4 4
6	2 5 7
8	3 8 9
9	1

Key: $1 \mid 2 = 12$

1. Determine the mode, the mean.
2. Find the range.

Table of Contents

Frequency Distributions and Histograms

Frequency Distribution

A **frequency distribution** is a more compact summary of data than a stem-and-leaf diagram. To construct a frequency distribution:

1. Divide the range of the data into intervals; called class intervals, cells, or bins.
2. The bins should be of equal width.

Note

1. The bins are in the form of Lower Limit – Upper Limit. The lower limit is the smallest data value in the bin and the upper limit is the highest data value in the bin.
   Example: In the bin $4-6$, 4 is the lower limit and 6 is the upper limit.
2. The lower limit is included in the bin but the upper limit cannot be included in the bin.
   Example: The observation 4 will be included in $4-6$ but 6 is not included in the same bin. Rather it will be used in the bin $6-8$. This method is known as an **exclusive method**.
3. Choosing the **number of bins** approximately equal to the square root of the number observations (round up the answer to the next integer) often works well in practice.

Example (Frequency Distribution)

Construct a frequency distribution for the data which is the quiz grades of a group:

$$2.4,\ 4.4,\ 4.6,\ 5.0,\ 5.0,\ 5.8,\ 6.0\ 7.4,\ 8.2,\ 9.0.$$

1. Divide grade ranges into 5 bins:

$$0-2,\ 2-4,\ 4-6,\ 6-8,\ 8-10.$$

2. Count the number of data values in each bins.

Bin	Frequency
$0-2$	0
$2-4$	1
$4-6$	5
$6-8$	2
$8-10$	2

Quiz

The following are age groups of 20 people in a concert.

$$5,\ 65,\ 62,\ 48,\ 5,\ 23,\ 17,\ 40,\ 32,\ 34,\ 35,\ 51,\ 6,\ 18,\ 52,\ 28,\ 39,\ 41,\ 20,\ 69.$$

Construct a frequency distribution.

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Figure. Compressive strength (in psi) of 80 Aluminum-Lithium alloy specimens.

Class	$70 \le x < 90$	$90 \le x < 110$	$110 \le x < 130$	$130 \le x < 150$	$150 \le x < 170$	$170 \le x < 190$	$190 \le x < 210$	$210 \le x < 230$	$230 \le x < 250$
Frequency	2	3	6	14	22	17	10	4	2
Relative frequency	0.0250	0.0375	0.0750	0.1750	0.2750	0.2125	0.1250	0.0500	0.0250
Cumulative relative frequency	0.0250	0.0625	0.1375	0.3125	0.5875	0.8000	0.9250	0.9750	1.0000

Figure. Frequency distribution for the compressive strength data in the above table.

Constructing a Histogram (Equal Bin Widths)

The **histogram** is a visual display of the frequency distribution. To construct a histogram

1. Label the bin (class interval) boundaries on a horizontal scale.
2. Mark and label the vertical scale with the frequencies or the relative frequencies.
3. Above each bin, draw a rectangle where height is equal to the frequency (or relative frequency) corresponding to that bin.

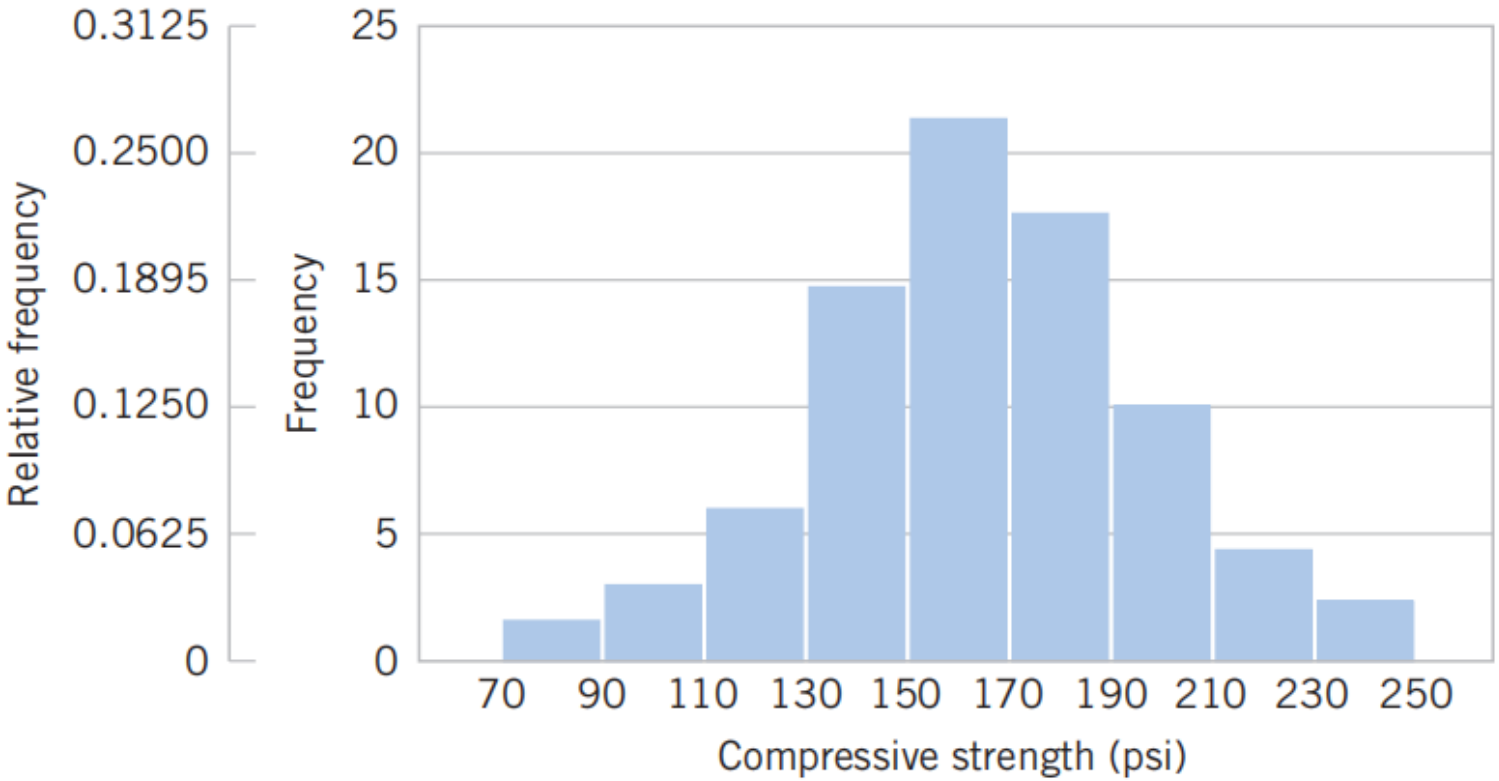


Figure. Histogram of compressive strength for 80 Aluminum-Lithium alloy specimens. FUQN

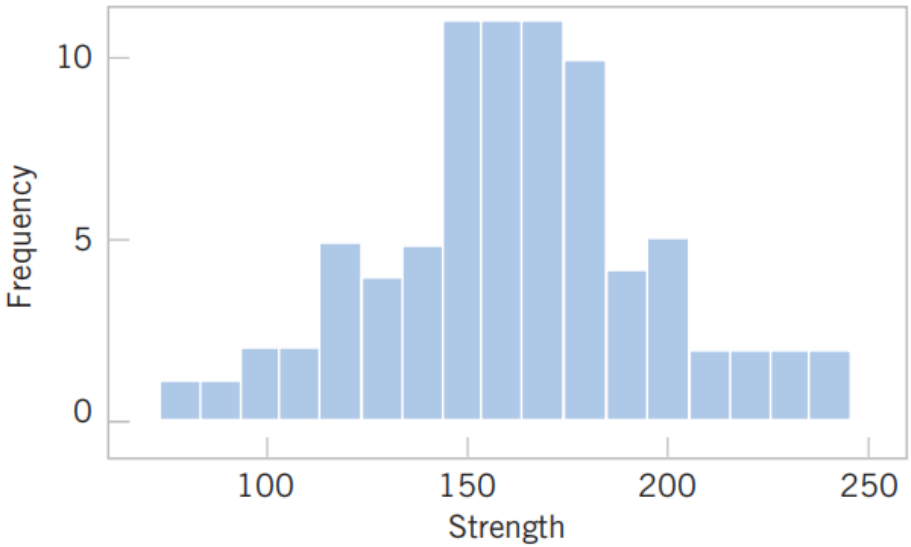


Figure. A histogram of the compressive strength data from Minitab with 17 bins.

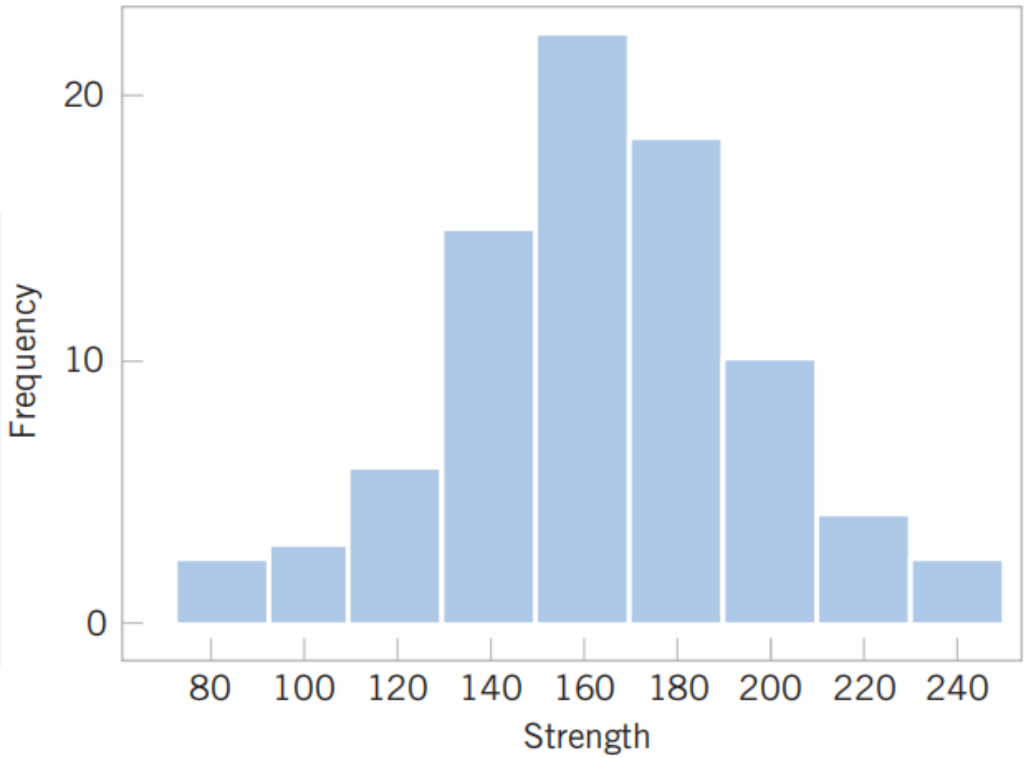


Figure. A histogram of the compressive strength data from Minitab with 9 bins.

1. Histogram are very useful to explore the distribution of data.

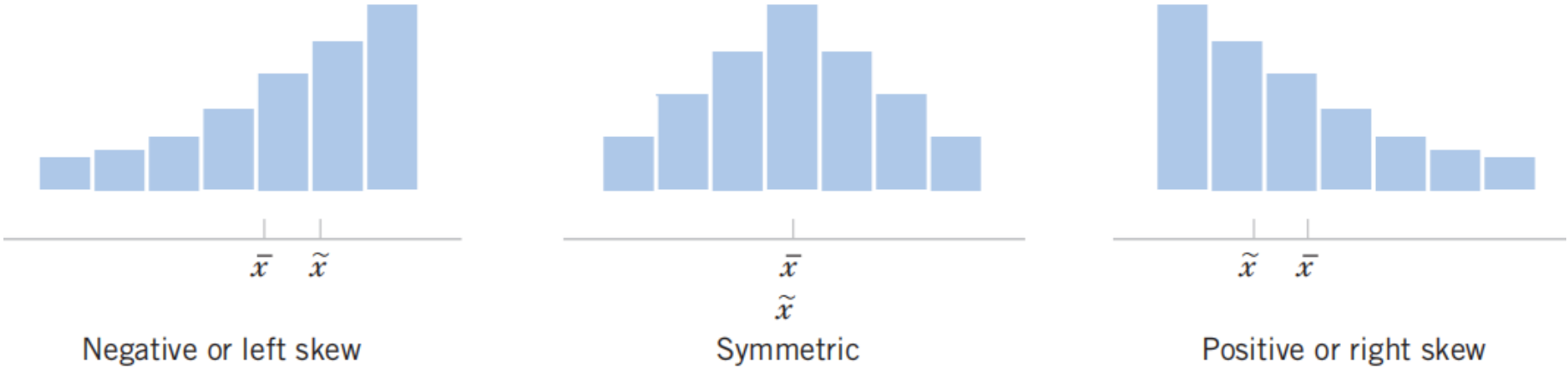


Figure. Histograms for symmetric and skewed distributions.

2. Pareto chart: frequencies are ordered decreasingly.

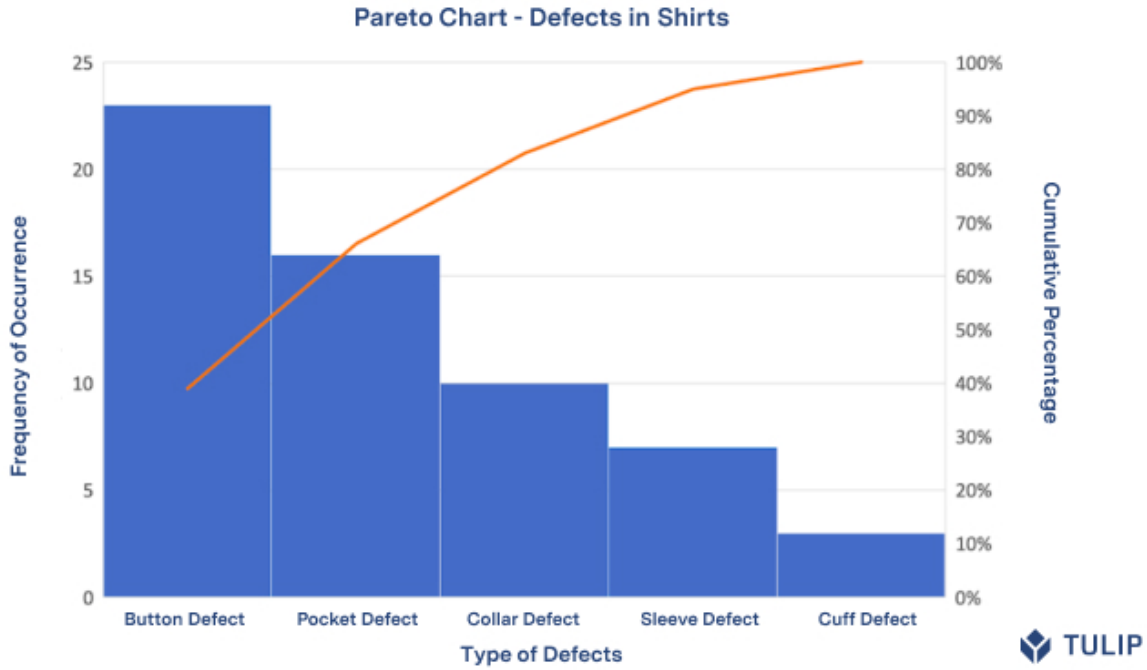
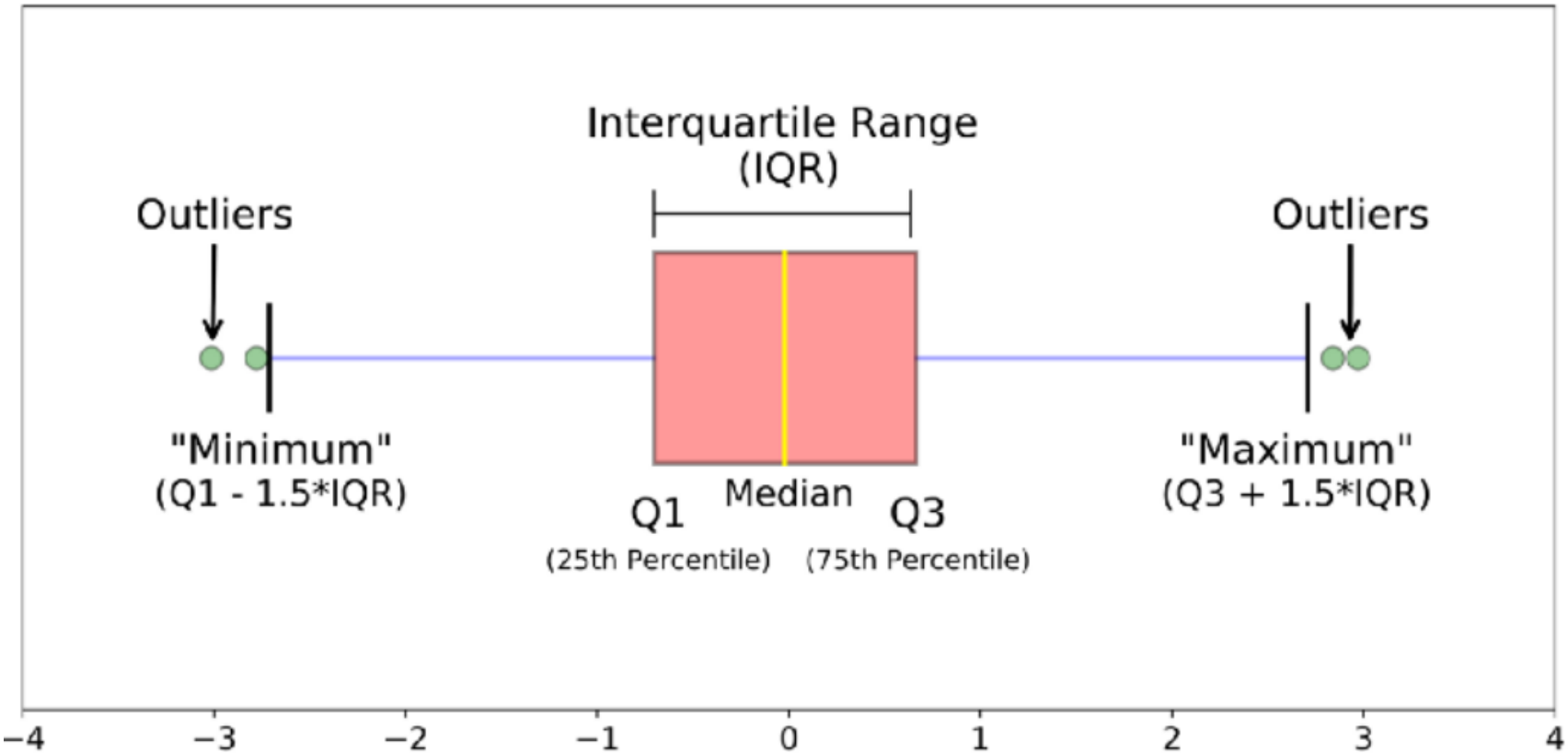


Table of Contents

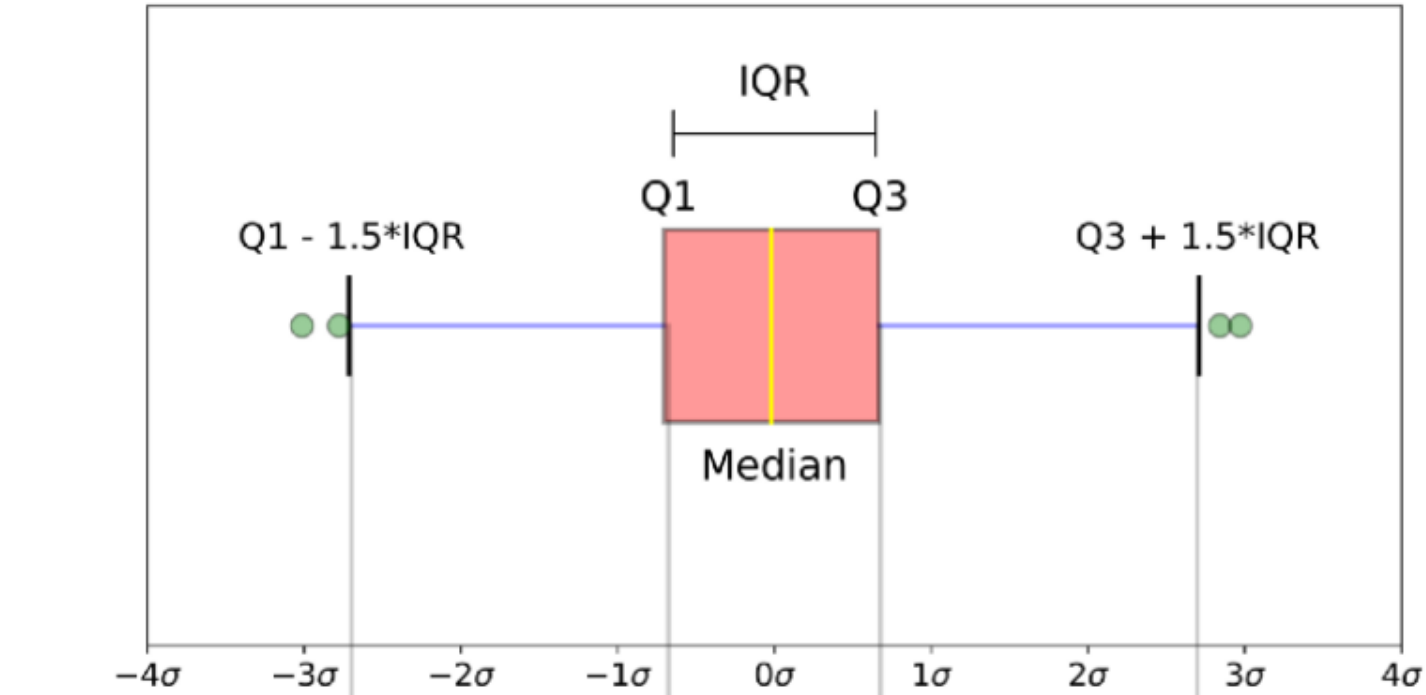
Box Plot (Box and Whisker Plot)

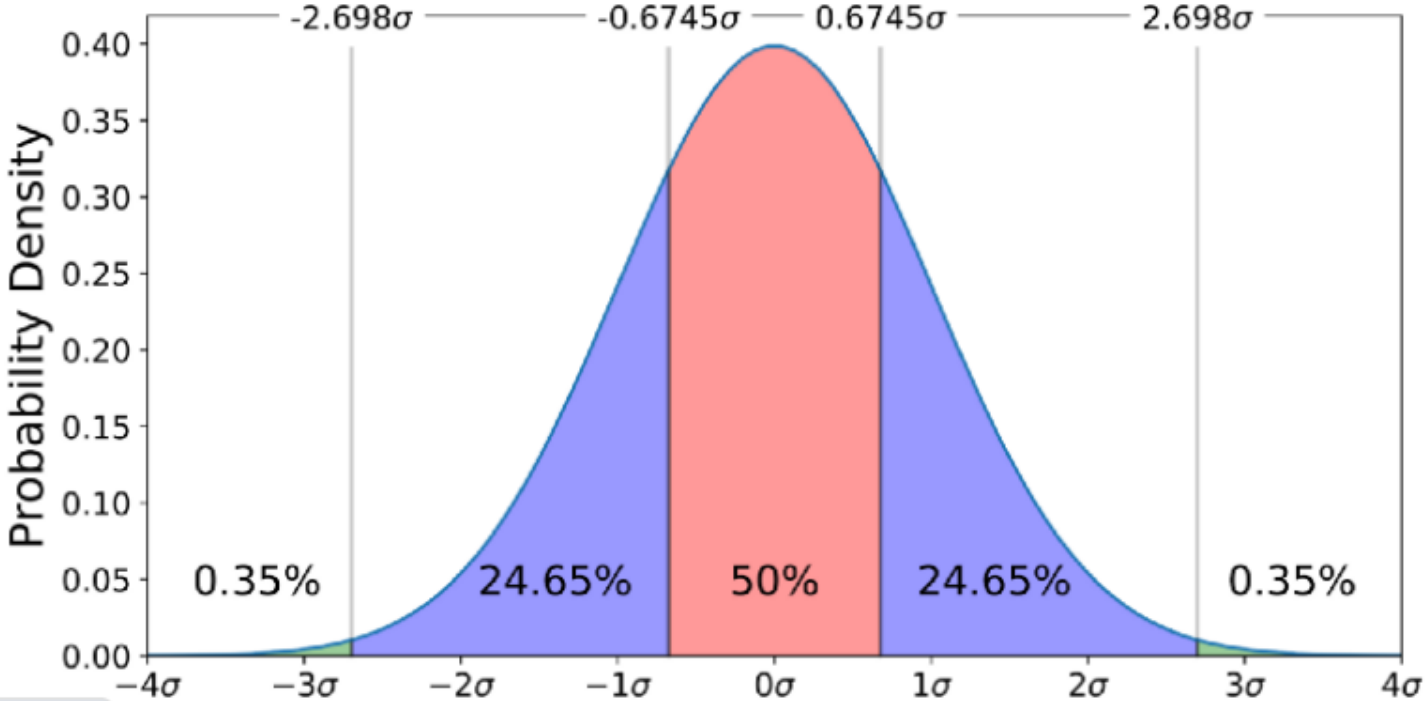
1. A **box plot** or boxplot (also known as box and whisker plot) is a type of chart often used in explanatory data analysis.
2. A box plot describes important features of data: three quartiles, the minimum and maximum values, and unusual observations (outliers).



Definition

1. **Minimum Score**: The lowest score, excluding outliers (shown at the end of the left whisker).
2. **First Quartile**, Q_1 or q_1: is a value that has approximately 25% of the observations below (also known as the lower quartile).
3. **Median**, Q_2 or q_2: has approximately 50% of the observations below its value. The median marks the mid-point of the data and is shown by the line that divides the box into two parts (sometimes known as the second quartile).
4. **Third Quartile**, Q_3 or q_3: has approximately 75% of the observations below its value.
5. **Maximum Score**: The highest score, excluding outliers (shown at the end of the right whisker).
6. **Whiskers**: The upper and lower whiskers represent scores outside the middle 50% (i.e. the lower 25% of scores and the upper 25% of scores).
7. **The Interquartile Range (or IQR)**: IQR $= Q_3 - Q_1$. This is the box plot showing the middle 50% of the observations (i.e., the range between the 25^{th} and 75^{th} percentile).


IQR
Q1
Q3
Q1 - 1.5*IQR
Q3 + 1.5*IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ


-2.698σ
-0.6745σ
0.6745σ
2.698σ
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.35%
24.65%
50%
24.65%
0.35%
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ

Example (Box Plot)

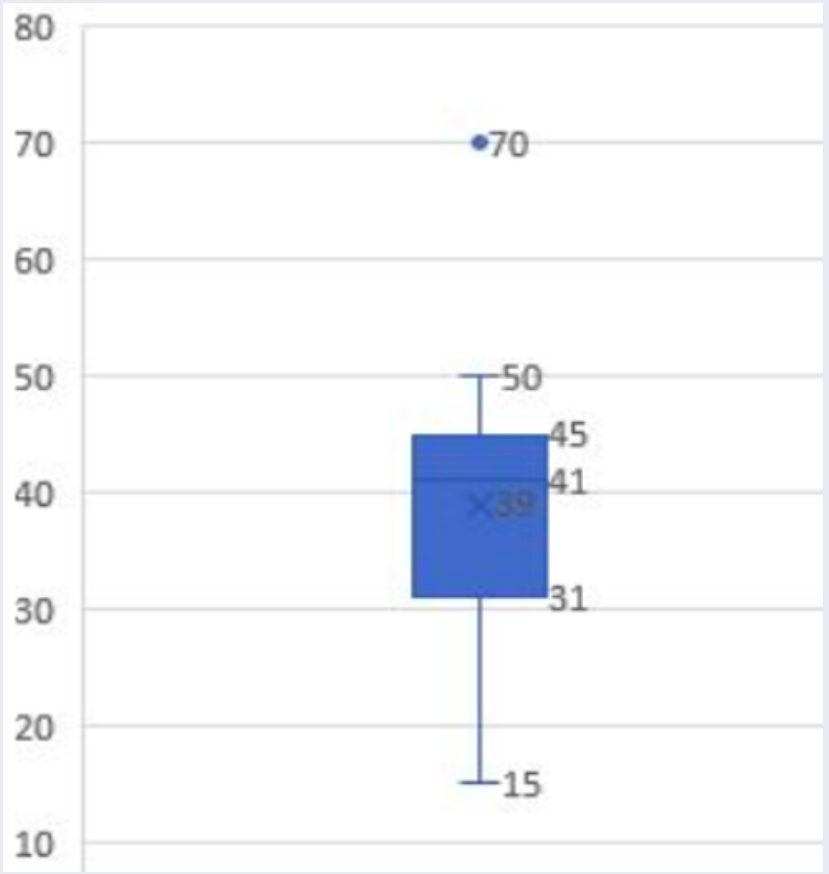
Given a data of ages of 14 random adults from a village:

$$15,\ 20,\ 31,\ 31,\ 32,\ 40,\ 41,\ 41,\ 42,\ 43,\ 45,\ 45,\ 50,\ 70.$$

Draw a box plot for this data.

Solution. We have

- Median: $\dfrac{41+41}{2} = 41.$
- Q_1 is the median of the data points to the left of Median, i.e., $Q_1 = 31.$
- Q_3 is the median of the data points to the right of Median, i.e., $Q_3 = 45.$
- IQR $= Q_3 - Q_1 = 14.$
- $Minimum = Q_1 - 1.5\text{IQR} = 10.$
- $Maximum = Q_3 + 1.5\text{IQR} = 66.$
- $Outliers$ are the points smaller than $Minimum$ and larger than $Maximum$,i.e., $Outlier = 70.$



Quiz

Use the given sample data to find the sample quartiles, the sample mode and the IQR.

$$55,\ 52,\ 52,\ 52,\ 49,\ 74,\ 67,\ 55.$$

Quiz

The "cold start ignition time" of an automobile engine is being investigated by a gasoline manufacturer. The following times (in seconds) were obtained for a test vehicle:

$$1.75,\ 1.92,\ 2.62,\ 2.35,\ 3.09,\ 3.15,\ 2.53,\ 1.91.$$

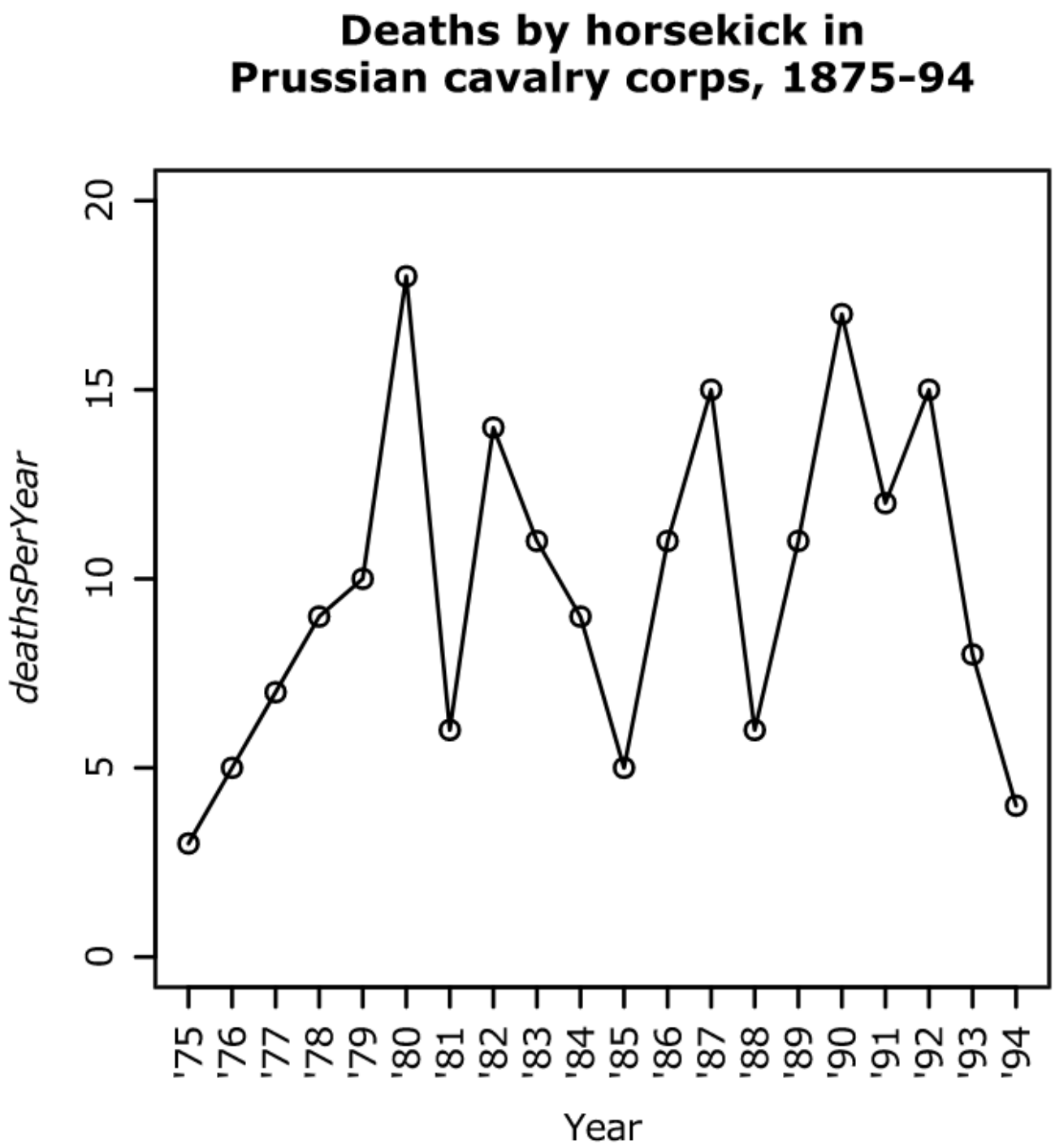
Construct a box plot of the data.

Table of Contents

Time Sequence Plots

Time Sequence Plot

A **time sequence plot** (or **time series plot**) is a graph in which the vertical axis denotes the observed value of the variable (say, x) and the horizontal axis denotes the time (which could be minutes, days, years and etc).


Deaths by horsekick in
Prussian cavalry corps, 1875-94
deathsPerYear
0
5
10
15
20
'75
'76
'77
'78
'79
'80
'81
'82
'83
'84
'85
'86
'87
'88
'89
'90
'91
'92
'93
'94
Year

Thank you!