Descriptive Statistics

Mahbub Latif, PhD

November 2025

Plan

- Experimentation
- Graphical presentation
- Sample Statistics

Introduction

- The first five chapters on probability theory described how the properties of a random variable can be understood using the probability mass function or probability density function of the random variable.

- In most applications, the probability mass function or probability density function of a random variable is not known by an experimenter

- One of the first tasks of the experimenter is to find out as much as possible about the probability distribution of the random variable under consideration

- This is done through experimentation and the collection of a dataset relating to the random variable.

Experimentation

Machine breakdown example

- Consider the machine breakdown example where machines break due to either electrical causes, mechanical causes, or operator misuse
- The probability mass function was known for this example, but in practice, it remains unknown to experimenters
- For this example, to obtain values of probability mass function, the experimenter can observe whether the number of machine breakdowns is actually attributable to each of the three causes
- The experimenter can estimate the probability mass function from the observed data

Amount of milk in a container

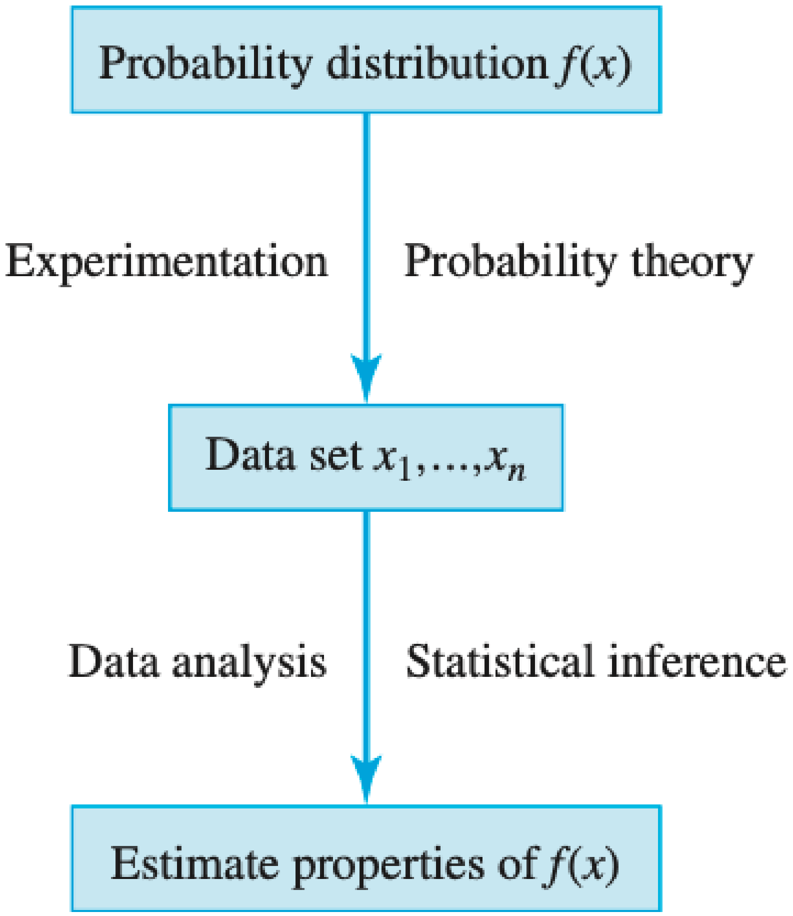
- Data obtained from an experiment can be used to estimate probability density function $f(x)$ too

- Consider the example of the amount of milk in a container, we need to estimate the probability density function $f(x)$

Amount of milk in a container

- To estimate $f(x)$, the experimenter can conduct an experiment because it is not possible to measure all milk containers:
  - Select n milk containers
  - Measure the amount of milk in these selected containers, e.g., $x_1, x_2, \dots, x_n$ is the amount of milk of n milk containers
  - Use the data $x_1, x_2, \dots, x_n$ to estimate probability density function $f(x)$
- Statistical methods that deal with estimating probability mass function or probability density function based on the observed data are known as **statistical inference**

Relationship between probability theory and statistical inference



Probability distribution $f(x)$

Experimentation

Probability theory

Data set $x_1,\ldots,x_n$

Data analysis

Statistical inference

Estimate properties of $f(x)$

Population and Samples

- A *population* consists of all possible observations available from a particular probability distribution

- A probability distribution (density function or mass function) can be considered as a population, characteristics of a population (e.g., mean, variance, etc.) remain unknown to researchers

Population and Samples

- A *sample* is a particular subset of the population that an experimenter measures and uses to investigate the unknown probability distribution.
    - Notation: $x_1, \ldots, x_n$ be a sample from a population
- A *random sample* is one in which the elements of the sample are chosen at random from the population
- A random sample is often used to ensure that the sample is representative of the population.

Types of variables

- A variable is a characteristic that can vary in value among subjects in a sample or population, e.g. last year income, gender, etc.
- Qualitative or categorical variables measure values that differ in quality not in numerical magnitude, e.g., gender, religion, etc.
  - Two types of categorical variables are ordinal (e.g., social status) and nominal (e.g., gender)
- Quantitative variables measure values in numerical magnitude, e.g., age, height, number of family members, etc.
  - Two types of quantitative variables are discrete and continuous

Example (Machine Breakdown)

- The engineer in charge of the maintenance of the machine keeps records on the breakdown causes over a period of a year.

- Altogether there are 46 breakdowns, of which
  - 9 are attributable to electrical causes, 24 are at to mechanical causes, and 13 are to operator misuse.

Example (Machine Breakdown)

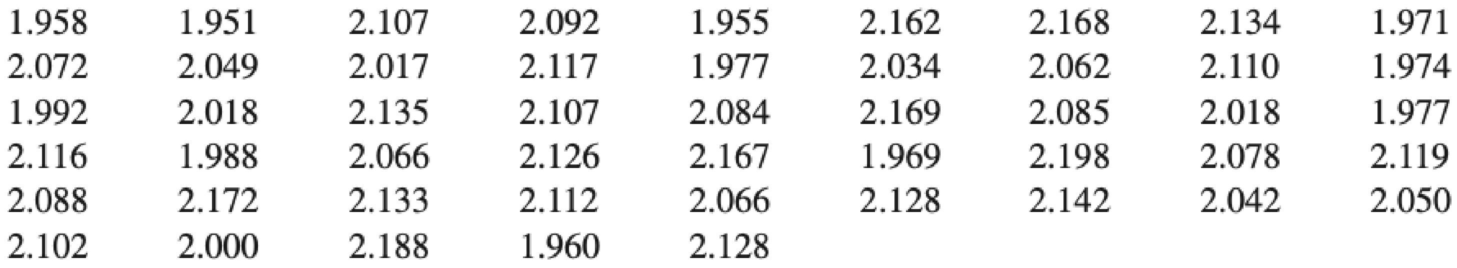
- The data set consists of 46 categorical variables:

$$x_1, x_2, \ldots, x_{46}$$

- Each x's is either "Electrical" or "Mechanical" or "Misuse" category
- In this sample of 46 observations
  - 9 observations are "Electrical"
  - 24 are "Mechanical"
  - 13 are "Misuse" category

Example (Milk container contents)

- A random sample of 50 milk containers is selected, and their milk contents are weighed, which can be used to investigate the unknown underlying probability distribution of the milk container weights.



1.958	1.951	2.107	2.092	1.955	2.162	2.168	2.134	1.971
2.072	2.049	2.017	2.117	1.977	2.034	2.062	2.110	1.974
1.992	2.018	2.135	2.107	2.084	2.169	2.085	2.018	1.977
2.116	1.988	2.066	2.126	2.167	1.969	2.198	2.078	2.119
2.088	2.172	2.133	2.112	2.066	2.128	2.142	2.042	2.050
2.102	2.000	2.188	1.960	2.128				

  - Sample observations $x_1, \ldots, x_{50}$ represent milk contents in litres

- The population in this experiment is the collection of all the milk containers produced and, again, the random selection of the sample should ensure that it is representative.

Data Presentation

Introduction

- Once a data set has been collected, the experimenter's next task is to find an informative way of presenting it.
- In general, a table of numbers is not very informative, whereas a picture or graphical representation of the data set can be quite informative.
- Depending on the type of data (e.g., quantitative or qualitative), different graphical presentations are used
- Quantitative data (discrete and continuous) take values from the real lines, and qualitative data (categorical) represent attributes or characteristics of a unit

Bar chart

- A bar chart is a simple graphical technique for presenting a categorical data set.
- Each category has a bar whose length is proportional to the frequency associated with that category
- A Pareto chart is a bar chart where the categories are arranged in order of decreasing frequency

Example I (Machine breakdown)

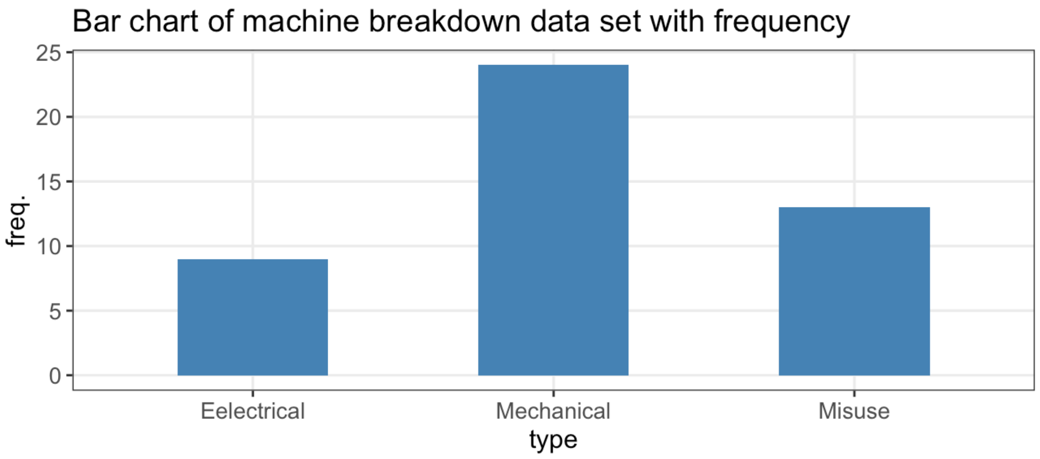
- The engineer in charge of the maintenance of the machine keeps records on the breakdown causes over a period of a year.

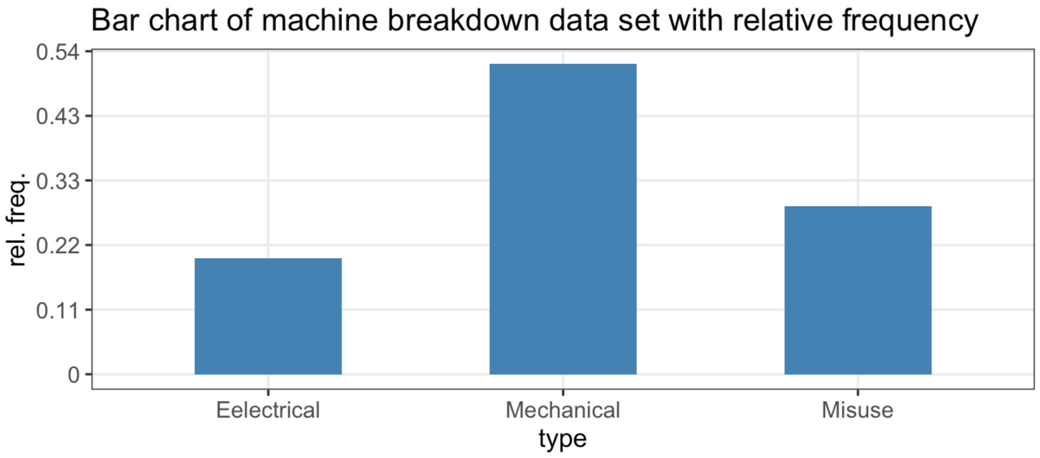
- Altogether there are 46 breakdowns
  - 9 are attributable to electrical causes,
  - 24 are attributable to mechanical causes, and
  - 13 are attributable to operator misuse.

Example I (Machine breakdown)

Frequency distribution

type	frequency	relative frequency
Eelectrical	9	0.196
Mechanical	24	0.522
Misuse	13	0.283


Bar chart of machine breakdown data set with frequency

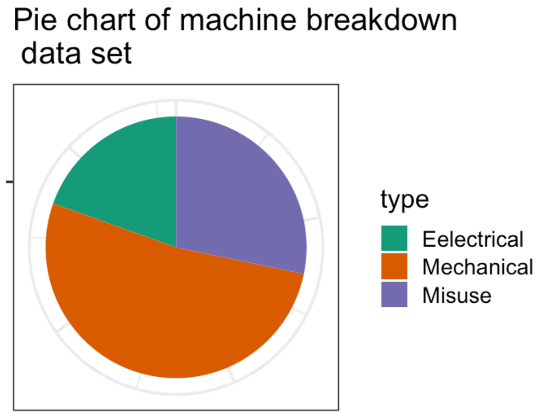

Bar chart of machine breakdown data set with relative frequency

Pie chart

- Pie charts are an alternative way of presenting the frequencies of categorical data in a graphical manner.
- A pie chart emphasizes the proportion of the total data set that is taken up by each of the categories
- If a data set of n observations has r observations in a specific category, then that category receives a "slice" of the pie with an angle of $(r/n) \times 360^\circ$

Pie chart

type	frequency	angle
Eelectrical	9	70.4
Mechanical	24	187.8
Misuse	13	101.7



Histograms

- Histograms look similar to bar charts, but they are used to present quantitative (discrete or continuous) data
- In bar charts, the "x-axis" lists the various categories under consideration, whereas in histograms the "x-axis" is a numerical scale
- A histogram consists of a number of bands (intervals) whose length is proportional to the number of data observations that take a value within that band
- An important consideration in the construction of a histogram is an appropriate choice of the bandwidth (length of the interval), which should be between 6 to 12

Example (Milk container contents)

- The minimum and maximum milk contents are

min	max
1.951	2.198

- To construct ten intervals from 1.95 to 2.20, we need to consider the interval length

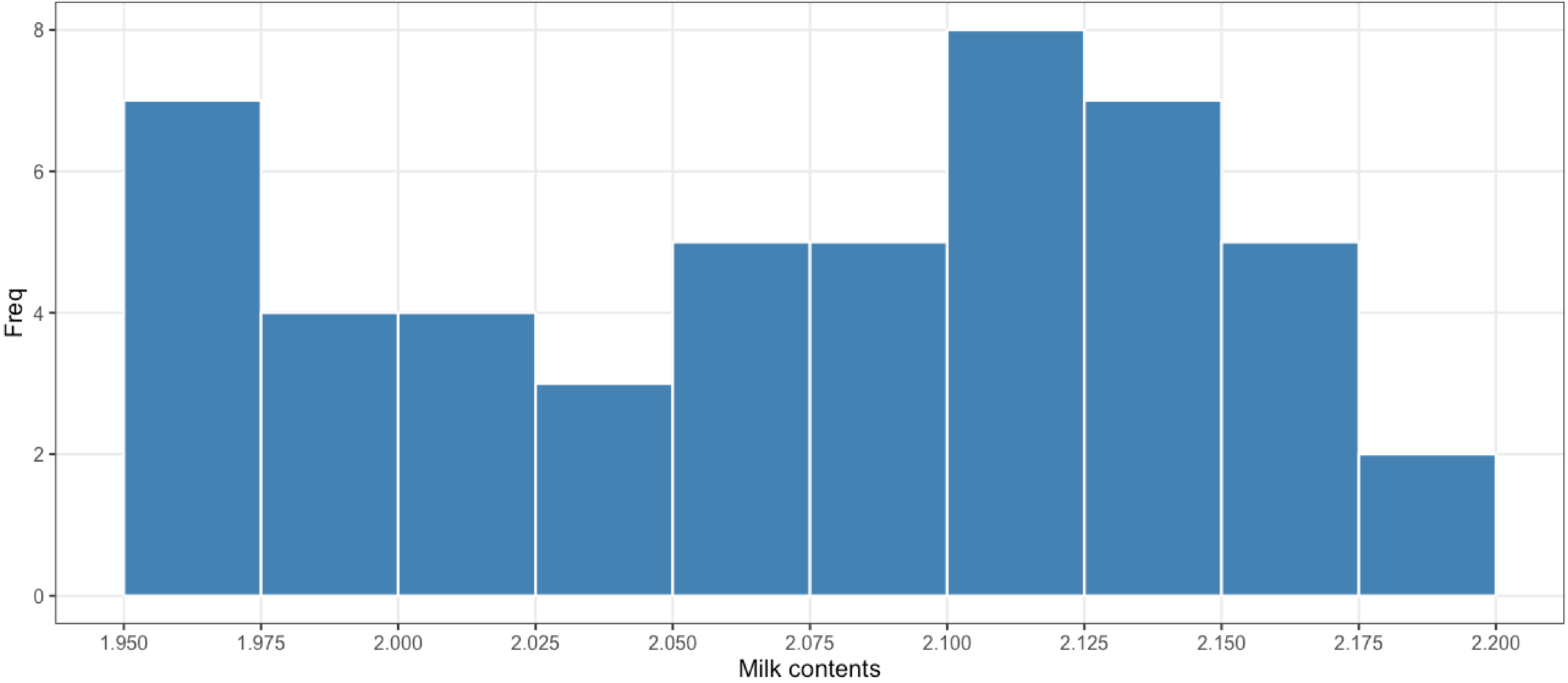
$$L = \frac{2.2 - 1.95}{10} = 0.025$$

Example (Milk container contents)

- Frequency distribution of milk contents, where the interval "1.95-1.975" indicates "[1.95, 1.975)"

Interval	Frequency	Relative frequency
1.950 - 1.975	7	0.14
1.975 - 2.000	4	0.08
2.000 - 2.025	4	0.08
2.025 - 2.050	3	0.06
2.050 - 2.075	5	0.10
2.075 - 2.100	5	0.10
2.100 - 2.125	8	0.16
2.125 - 2.150	7	0.14
2.150 - 2.175	5	0.10
2.175 - 2.200	2	0.04

Example (Milk container contents)



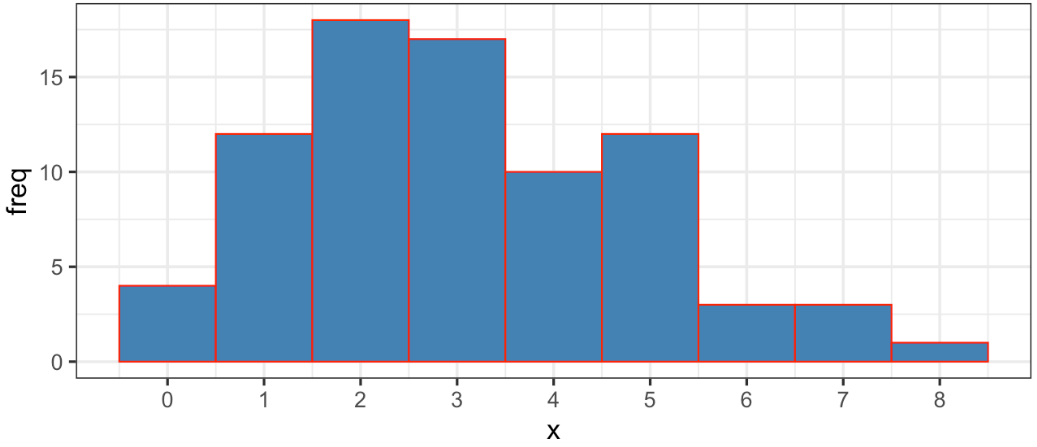
Example II (Defective computer chips)

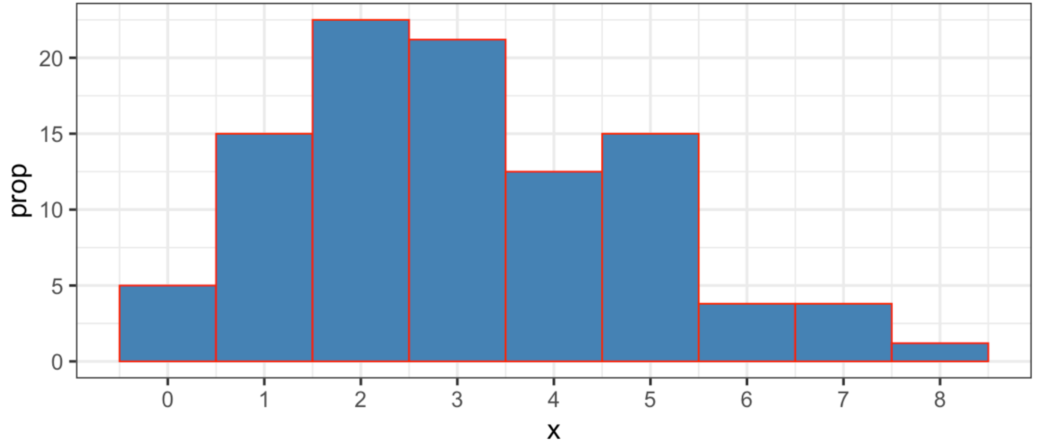
- A company sells computer chips in boxes of 500 chips, the distribution of the number of defective chips in a box is shown in the following table

number of defective chips (x)	frequency (r)
0	4
1	12
2	18
3	17
4	10
5	12
6	3
7	3
8	1

Example II (Defective computer chips)

x	interval	freq	prop
0	(-0.5, 0.5)	4	5.0
1	(0.5, 1.5)	12	15.0
2	(1.5, 2.5)	18	22.5
3	(2.5, 3.5)	17	21.2
4	(3.5, 4.5)	10	12.5
5	(4.5, 5.5)	12	15.0
6	(5.5, 6.5)	3	3.8
7	(6.5, 7.5)	3	3.8
8	(7.5, 8.5)	1	1.2





Skewness of a distribution

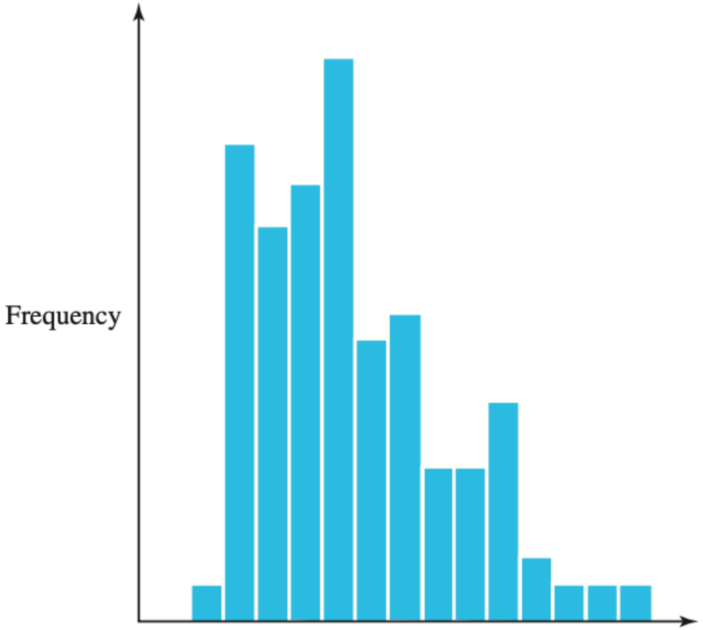


FIGURE 6.18

A histogram with positive skewness

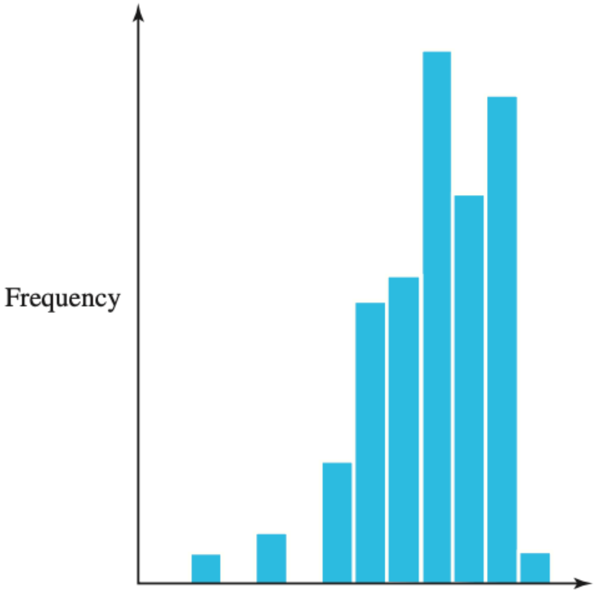


FIGURE 6.19

A histogram with negative skewness

Sample Statistics

Sample Statistics

- Sample statistics provide numerical summary measures of a data set
- Commonly used sample statistics
    - The sample mean, the sample median, and the sample standard deviation
- These statistics provide a numerical summary in the same way that the expectation, median, and standard deviation provide that of a probability distribution

Sample Statistics

- Sample statistics can be classified into two groups:
    - measures of central tendency or location and measures of spread
- Measures of central tendency or location
    - sample mean, sample median, sample mode, etc.
- Measures of spread
    - sample variance, sample standard deviation, inter-quartile range, coefficient of variation, etc.

Sample mean

- For a sample of n observations $x_1, \ldots, x_n$, the sample mean ($\bar{x}$, read "x bar") is defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- For a sample of the following 20 observations, calculate sample mean
  - 0.9, 1.3, 1.8, 2.5, 2.6, 2.8, 3.6, 4.0, 4.1, 4.2, 4.3, 4.3, 4.6, 4.6, 4.6, 4.7, 4.8, 4.9, 4.9, 5.0

Sample mean

- For a sample of the following 20 observations, calculate sample mean
  - 0.9, 1.3, 1.8, 2.5, 2.6, 2.8, 3.6, 4.0, 4.1, 4.2, 4.3, 4.3, 4.6, 4.6, 4.6, 4.7, 4.8, 4.9, 4.9, 5.0
- Sample mean

$$\bar{x} = \frac{\sum x}{n} = \frac{74.5}{20} = 3.725$$

Sample median

- Sample median is the middle most observation of the sample

Sample median

- Steps of obtaining sample median
  - Order the sample observations $\left[\text{e.g. the smallest } x_{(1)} \text{ to the largest } x_{(n)}\right]$

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

  - If n is odd

$$\text{median} = x_{((n+1)/2)}$$

  - If n is even

$$\text{median} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

Sample median

- Compare to sample mean, sample median is less sensitive to extreme values, so for a skewed distribution, sample median is preferable over sample mean as a measure of location

Sample median

- For a sample of the following 20 observations, calculate sample mean
    - 0.9, 1.3, 1.8, 2.5, 2.6, 2.8, 3.6, 4.0, 4.1, 4.2, 4.3, 4.3, 4.6, 4.6, 4.6, 4.7, 4.8, 4.9, 4.9, 5.0
- Here sample size n is even, so sample median is the average of the $(n/2)^{th}$ and $[(n/2)+1]^{th}$ observations of the ordered sample,
    - i.e. average of the 10^{th} and 11^{th} observations

Sample median

- The ordered sample
  - 0.9, 1.3, 1.8, 2.5, 2.6, 2.8, 3.6, 4.0, 4.1, **4.2**, **4.3**, 4.3, 4.6, 4.6, 4.6, 4.7, 4.8, 4.9, 4.9, 5.0
- The sample median

$$\frac{10^{th}\ \text{observation} + 11^{th}\ \text{observation}}{2} = 4.25$$

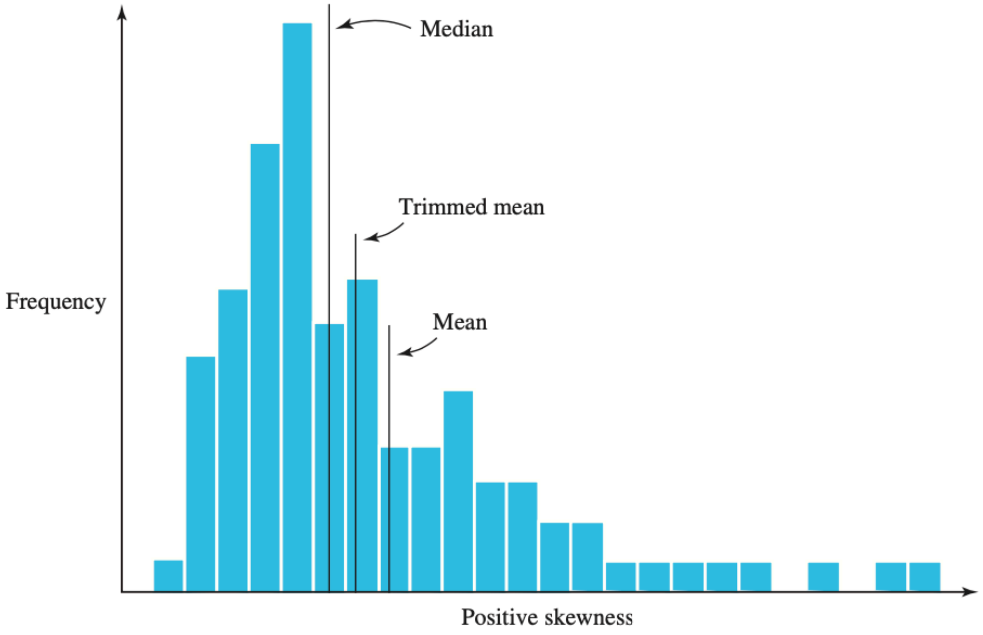
Quantile

- The pth quantile x_p of a sample $x_1, \ldots, x_n$ is calculated as:
  - Ordered sample: $x_{(1)}, \ldots, x_{(n)}$
  - If np is an integer, x_p is the average of $x_{(np)}$ and $x_{(np+1)}$ observations
  - If np is not an integer, x_p is the smallest integer that is greater than np

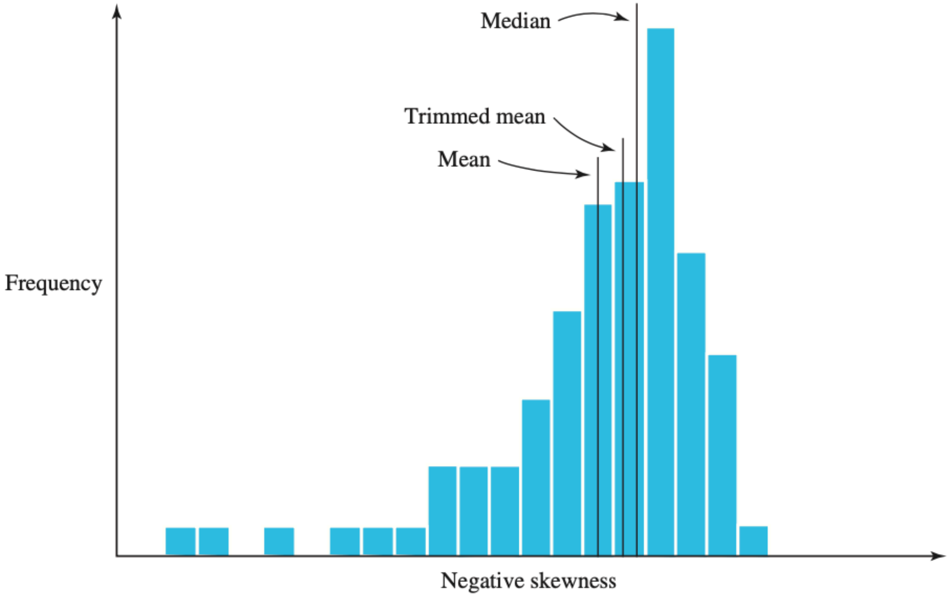
Sample mode

- The sample mode is the most frequent observation of the sample
- For a categorical variable, sample mean and sample median cannot be calculated, mode is the only summary measure one can calculate
- For a quantitative variable, all three measures (mean, median, and model) can be calculated
- The mode of the following sample is 4.6 as it is observed three times in the sample
    - 0.9, 1.3, 1.8, 2.5, 2.6, 2.8, 3.6, 4.0, 4.1, 4.2, 4.3, 4.3, 4.6, 4.6, 4.6, 4.7, 4.8, 4.9, 4.9, 5.0

- Comparison among the measures of location between positively and negatively skewed data



$$Mode < Median < Mean$$



$$Mean < Median < Mode$$

Sample variance

- Sample variance of a sample of n observations $x_1, \ldots, x_n$ is defined as

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

- The sample standard deviation is the positive square-root of sample variance, i.e. $s = +\sqrt{s^2}$

Sample variance

- For the example sample, we get

$$n = 20, \bar{x} = 3.725, \sum_{i=1}^{n} x_i^2 = 308.61$$

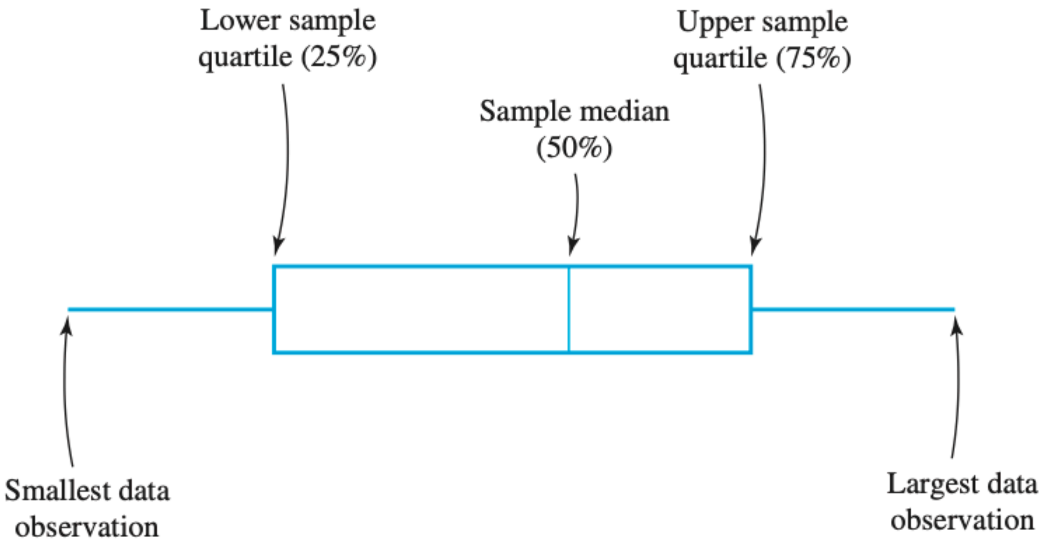
- The sample variance

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{308.61 - 20(3.725^2)}{20-1} = \frac{31.0975}{9} = 1.637$$

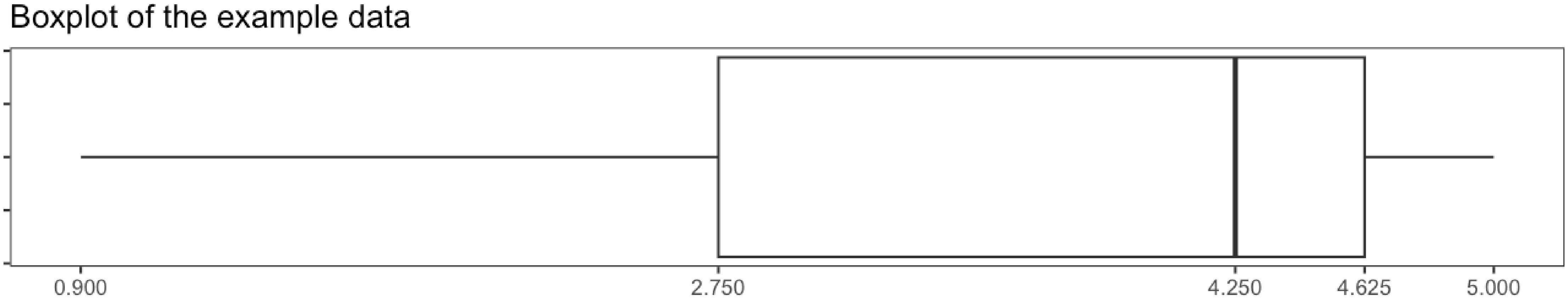
  - The sample standard deviation $s = \sqrt{1.637} = 1.279$

Boxplot

- A boxplot is a schematic presentation of the sample median, the upper and lower sample quartiles, and the largest and smallest data observations.


Lower sample quartile (25%)
Upper sample quartile (75%)
Sample median (50%)
Smallest data observation
Largest data observation

Boxplot



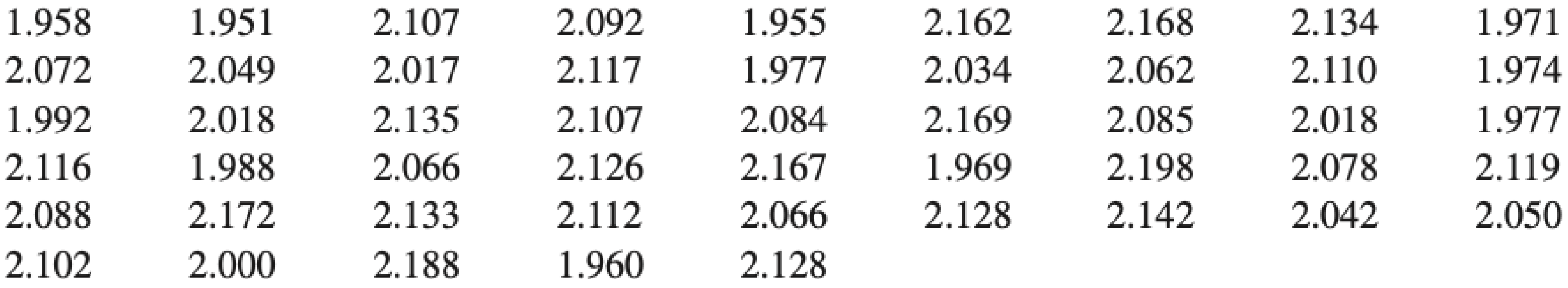
- Minimum: 0.9
- First quartile: 2.75
- Second quartile (median): 4.25
- Third quartile: 4.625
- Maximum: 5

Boxplot

- Boxplot is used to examine whether the distribution is symmetric or skewed
- Distribution is symmetric if median line split the box into two equal parts
- Distribution is negatively (positively) skewed if median line is more closer to the right (left) vertical line of the box

Exercise

- Obtain a boxplot of milk contents data



1.958	1.951	2.107	2.092	1.955	2.162	2.168	2.134	1.971
2.072	2.049	2.017	2.117	1.977	2.034	2.062	2.110	1.974
1.992	2.018	2.135	2.107	2.084	2.169	2.085	2.018	1.977
2.116	1.988	2.066	2.126	2.167	1.969	2.198	2.078	2.119
2.088	2.172	2.133	2.112	2.066	2.128	2.142	2.042	2.050
2.102	2.000	2.188	1.960	2.128				

- Median split the data two equal parts, median of the first part of the data is first quartile and the median of the second part of the data is the third quartile

Coefficient of variation

- Coefficient of variation (CV) measures spread of the data relative to mean and is defined as

$$CV = \frac{s}{\bar{x}}$$

- CV is unit-less and can be used to compare spread of two data sets that are of different units

- Large values of the coefficient of variation imply that the variability is large relative to the sample average, while small values indicate that the variability is small relative to the sample average

Example 42

- A zoologist is interested in the variations in the weights of different kinds of animals.
- A data set of adult male African elephants provided weights with
  - a sample average of $\bar{x}_e = 4550$ kg and a sample standard deviation of $s_e = 150$ kg
- A data set concerning a certain kind of mouse provided weights with
  - a sample average of $\bar{x}_m = 30$ g and a sample standard deviation of $s_m = 1.67$ g.

Example 42

- The variation in the elephant weights is larger than the variation in the mice weights when compared directly because the elephant weights are so much larger.

- However, the coefficient of variation for the elephant and mice weights are

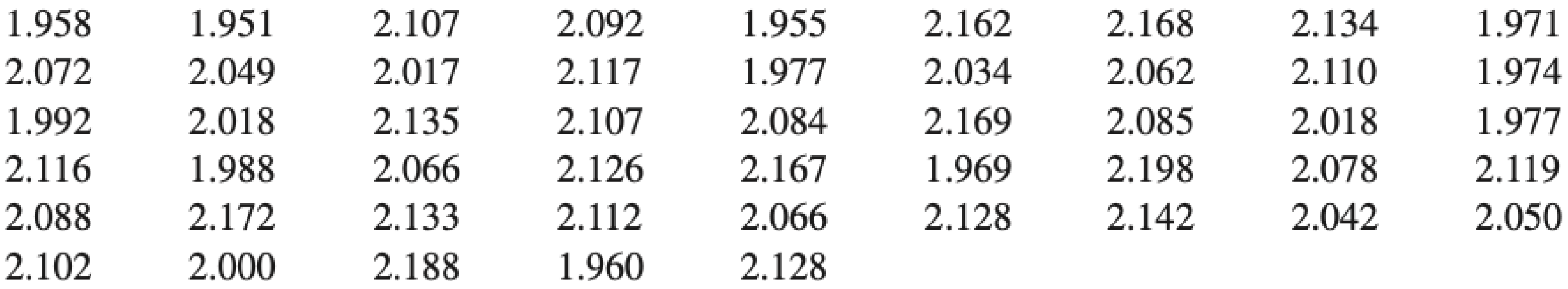
$$\text{Elephant: } CV_e = 150/4550 = 0.033$$

$$\text{Mice: } CV_m = 1.67/30 = 0.056$$

  - It can be seen that the mice have more variability in their weights than the elephants relative to their respective average weights.

Example (Milk container contents)

- A random sample of 50 milk containers is selected and their milk contents are weighed, which can be used to investigate the unknown underlying probability distribution of the milk container weights.



1.958	1.951	2.107	2.092	1.955	2.162	2.168	2.134	1.971
2.072	2.049	2.017	2.117	1.977	2.034	2.062	2.110	1.974
1.992	2.018	2.135	2.107	2.084	2.169	2.085	2.018	1.977
2.116	1.988	2.066	2.126	2.167	1.969	2.198	2.078	2.119
2.088	2.172	2.133	2.112	2.066	2.128	2.142	2.042	2.050
2.102	2.000	2.188	1.960	2.128				

Example (Milk container contents)

- ~~Draw a histogram and identify whether the distribution is symmetric or skewed.~~
- Obtain sample mean and median.
- Obtain sample standard deviation, first and third quartile
- Draw a boxplot.