Module 1
Mathematics for Data Analytics

Jency Thomas

Asst. Professor, CSE

MITS Kochi

- **Module – 1 (Mathematics for Data Analytics)**
- Descriptive statistics - Measures of central tendency and dispersion, Association of two variables - Discrete variables, Ordinal and Continuous variable, Probability calculus - probability distributions, Inductive statistics - Point estimation, Interval estimation, Hypothesis Testing - Basic definitions, t-test

What is Statistics?

- Statistics involves methods to describe, summarize, interpret, and analyze data.
- Helps in drawing meaningful conclusions from data across various fields.
- Helpful for
  - Researchers
  - Government and Organizations
  - Business People

Understanding Population, Sample, and Observations

In statistics, we study data collected from various units. Before diving into analysis, we need precise terminology to describe what we're measuring and where our data comes from. These fundamental concepts form the foundation of statistical thinking and help us communicate clearly about our research.

Core Terminology: Building Your Statistical Vocabulary

Before diving into statistical methods, we need to establish the foundational terms that form the language of statistics. These concepts may seem simple at first, but they provide the critical framework for all statistical thinking.

Observation (ω)

A single unit on which we measure data—could be a person, car, animal, plant, or any entity we're studying. Represented by the Greek symbol ω (omega).

Population (Ω)

The complete collection of all units we're interested in studying. When we write $\omega \in \Omega$, we mean one unit from the entire population—like one person out of all persons of interest.

Sample

A selection of observations $\omega_1, \omega_2, \ldots, \omega_n$ drawn from the population. Always a subset of the population: $\{\omega_1, \omega_2, \ldots, \omega_n\} \subseteq \Omega$. Samples allow us to make inferences about the entire population.

Real-World Examples: Populations and Samples in Action

Social Research in India

If we're interested in the social conditions under which Indian people live, we would define all inhabitants of India as our population Ω. Each individual inhabitant represents an observation ω.

Collecting data from every person would be impossible, so we select a sample—a manageable subset of inhabitants—to represent the whole population.

Africa's Platinum Industry

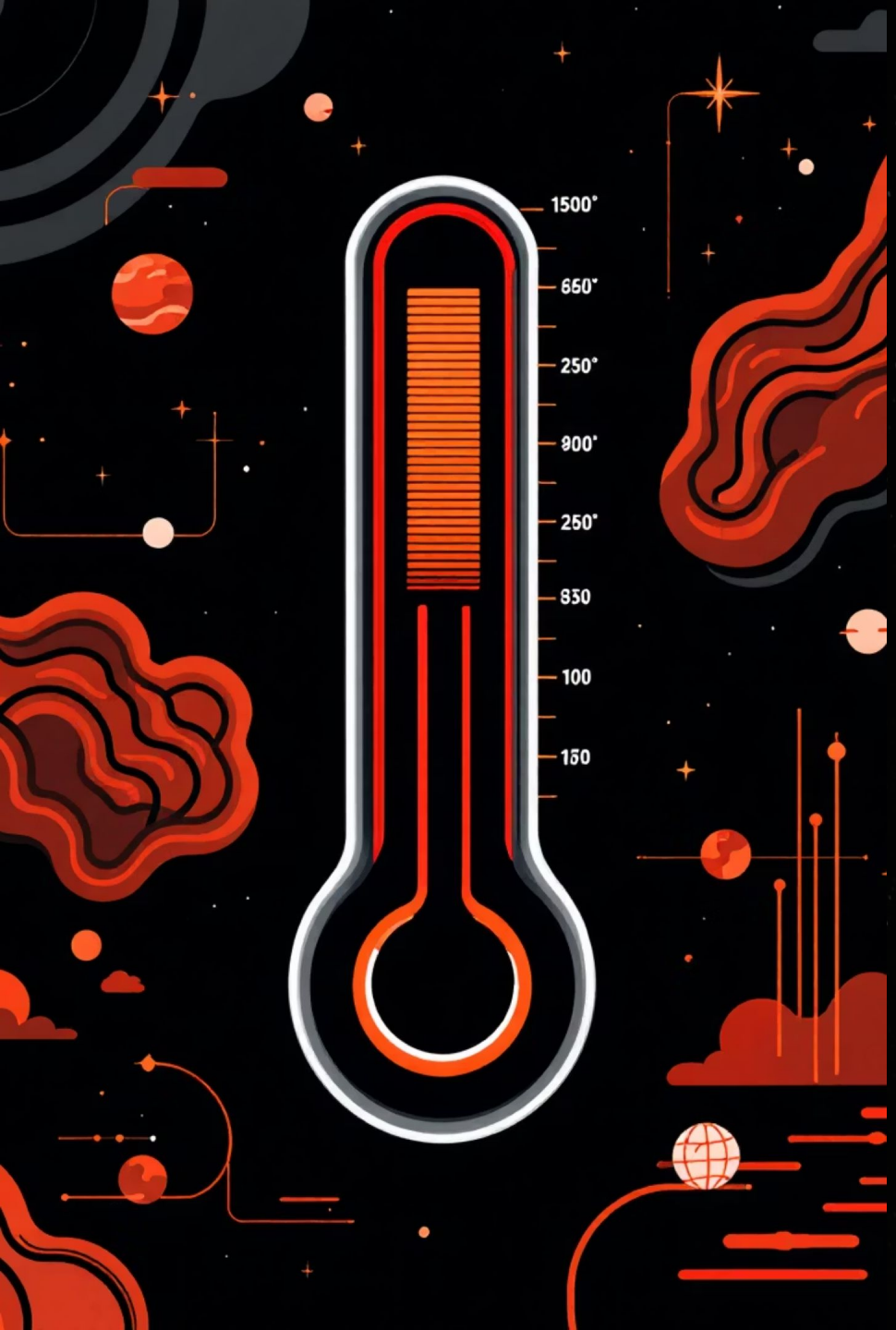
When investigating the economic power of Africa's platinum industry, each platinum-related company is an observation ω, while all such companies collectively form the population Ω.

A sample might consist of a few companies $\omega_1, \omega_2, ..., \omega_n$ selected for detailed economic analysis.

Statistics Course Participants

Consider collecting information about students in a statistics course. All participants constitute the population Ω, and each individual participant represents a unit or observation ω.

We might sample a subset of students to understand trends about the entire class.



When Populations Don't Apply

Sometimes the concept of a population is not applicable or difficult to imagine. Consider measuring temperature in New Delhi every hour. A sample would be the time series of temperatures in a specific window—say, January to March 2016. Here, a population of observational units doesn't clearly exist.

However, if we measure temperatures in several different cities, then all cities form the population Ω, and any subset of cities represents a sample. The key is flexibility in how we define our statistical units based on our research question.

Statistical Variables: Capturing Features of Interest

Once we've specified our population, we need to determine what information we want to collect. A **statistical variable X** captures a particular feature or characteristic of our observations.

01

Define the Population

Identify all units of interest (Ω)

02

Choose Variables

Decide what features to measure (X_1, X_2, ..., X_p)

03

Record Values

Each observation ω takes a specific value x

04

Analyze Data

Use statistical methods to draw conclusions

If observations are human beings, X might describe marital status, gender, age, income level, or any other characteristic. We can study multiple features simultaneously, with each feature collected in a different variable X_1, X_2, ..., X_p.

The Formal Definition of a Variable

Mathematically, a variable is defined as a function that maps each observation to a specific value:

$$X : \Omega \to S \quad \omega \to x$$

This definition states that a variable X takes a value x for each observation $\omega \in \Omega$. The set S contains all possible values the variable can take.

Gender Variable

If X refers to gender, possible values are S = {male, female}. Each person ω is assigned either "male" or "female".

Country of Origin (Cars)

For X = country of origin, values include S = {Italy, South Korea, Germany, France, India, China, Japan, USA, ...}.

Age Variable

A variable X representing age may take any value between 1 and 125 years, with each person ω assigned their specific age x.

Qualitative vs. Quantitative Variables

Qualitative Variables

Definition: Variables whose values cannot be ordered in a logical or natural way.

Examples:

- Color of the eye (blue, brown, green)
- Name of a political party
- Type of transport to work (bus, train, car, bike)

There's no reason to list blue eyes before brown eyes, nor does it make sense to rank buses before trains. The categories are distinct but unordered.

Quantitative Variables

Definition: Variables representing measurable quantities with values that can be ordered logically and naturally.

Examples:

- Size of shoes (6, 7, 8, 9, 10...)
- Price for houses ($200K, $300K, $400K...)
- Number of semesters studied (1, 2, 3, 4...)
- Weight of a person (50 kg, 60 kg, 70 kg...)

These values have a clear numerical ordering and can be compared mathematically.

Important Note: We sometimes assign numbers to qualitative variables for data analysis (e.g., 1 = female, 0 = male), but this doesn't make them quantitative. The assignment is arbitrary—we could just as easily use 2 and 10—so gender remains qualitative despite the numerical coding.

Discrete vs. Continuous Variables

Discrete Variables

Variables that can only take a **finite number of values**. All qualitative variables are discrete (e.g., eye color, geographic region), but quantitative variables can also be discrete.

Examples: Size of shoes (limited sizes available), number of semesters studied (1, 2, 3... but not 2.5), number of children in a family.

Continuous Variables

Variables that can take an **infinite number of values**. These are often described as variables that are "measured rather than counted."

Examples: Time to travel to university, length of an antelope, distance between planets, height of a person.

The crucial point: a person's height might be recorded as 172 cm, but could actually be 172.3 cm, or more precisely 172.342 cm, or really 172.342975328... cm. The infinite decimal possibilities make this continuous.

Understanding Scales of Measurement

Different variables contain different amounts of information. The **scale of a variable** classifies this information level and helps us identify which statistical methods are appropriate to use.

Nominal Scale

Values cannot be ordered. Examples: gender (male–female), application status (pending–not pending). No meaningful way to rank categories.

Ordinal Scale

Values can be ordered, but differences between values lack numerical meaning. Examples: education level (none–primary–secondary–university), satisfaction (unsatisfied–satisfied–very satisfied). We know the order, but can't quantify gaps.

Continuous Scale

Values can be ordered AND differences can be interpreted meaningfully. Example: height (170 cm, 171 cm, 172 cm...). The difference between 170 and 171 equals the difference between 171 and 172.

Subdivisions of the Continuous Scale

The continuous scale can be further divided into three subscales, each offering progressively more mathematical information:

Interval Scale

Only **differences** between values can be interpreted, not ratios. Example: Temperature in °C. The difference between -2°C and 4°C is 6°C, but 4/-2 = -2 doesn't mean -4°C is twice as cold as 2°C.

Ratio Scale

Both **differences and ratios** can be interpreted. Example: Speed. 60 km/h is 40 km/h more than 20 km/h, AND 60 km/h is three times faster than 20 km/h (ratio of 3).

Absolute Scale

Same as ratio scale but measured in **"natural" units** without artificial constructs. Example: Number of semesters studied (1, 2, 3...). No km/h or °C needed—values are inherently countable.

Understanding these scales is crucial for selecting appropriate statistical methods. Throughout this course, we'll see how the scale of measurement determines which analytical techniques we can legitimately apply to our data.

Grouped Data

- If data is available in grouped form, we call the respective variable capturing this information a grouped variable. Sometimes, these variables are also known as categorical variables.
- Any grouped or categorical variable which can only take two values is called a binary variable

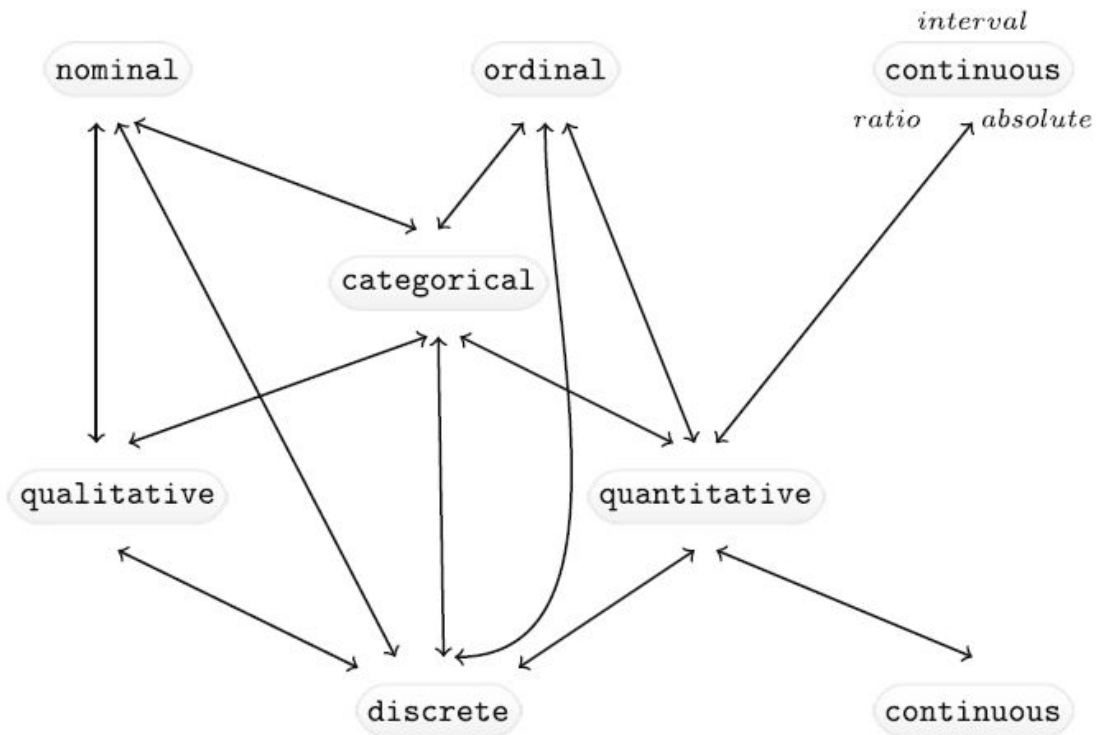


Fig. 1.1 Summary of variable classifications

Exercises

1.6 Exercises

Exercise 1.1 Describe both the population and the observations for the following research questions:

(a) Evaluation of the satisfaction of employees from an airline.
(b) Description of the marks of students from an assignment.
(c) Comparison of two drugs which deal with high blood pressure.

- *Solution to Exercise 1.1*
- (a) The population consists of all employees of the airline. This may include administration staff, pilots, stewards, cleaning personnel, and others. Each single employee relates to an observation in the survey.
- (b) The population comprises all students who take part in the examination. Each student represents an observation.
- (c) All people suffering high blood pressure in the study area (city, province, country,. . .), are the population of interest. Each of these persons is an observation.

Exercise 1.2 A national park conducts a study on the behaviour of their leopards. A few of the park's leopards are registered and receive a GPS device which allows measuring the position of the leopard. Use this example to describe the following concepts: population, sample, observation, value, and variable.

- The *population* in this study refers to all leopards in the park.
- Only a few of the leopards are equipped with the GPS devices. This is the *sample* on which the study is conducted in.
- Each leopard refers to an *observation*.
- The measurements are taken for each leopard in the sample. The GPS coordinates allow to determine the position during the entire day.
- Important *variables* to capture would therefore be $X1$ = “latitude”, $X2$ = “longitude”, and $X3$ = “time”.
- Each variable would take on certain *values* for each observation; for example, the first leopard may have been observed at latitude 32◦ at a certain time point, and thus $x11 = 32◦$.

Exercise 1.3 Which of the following variables are qualitative, and which are quantitative? Specify which of the quantitative variables are discrete and which are continuous:

Time to travel to work, shoe size, preferred political party, price for a canteen meal, eye colour, gender, wavelength of light, customer satisfaction on a scale from 1 to 10, delivery time for a parcel, blood type, number of goals in a hockey match, height of a child, subject line of an email.

- Qualitative: Preferred political party, eye colour, gender, blood type, subject line of an email.
- Quantitative and discrete: Shoe size, customer satisfaction on a scale from 1 to 10, number of goals in a hockey match.
- Quantitative and continuous: Time to travel to work, price of a canteen meal, wavelength of light, delivery time of a parcel, height of a child.

Exercise 1.4 Identify the scale of the following variables:

(a) Political party voted for in an election
(b) The difficulty of different levels in a computer game
(c) Production time of a car
(d) Age of turtles
(e) Calender year
(f) Price of a chocolate bar
(g) Identification number of a student
(h) Final ranking at a beauty contest
(i) Intelligence quotient.

Item	Scale
(a)	Nominal
(b)	Ordinal
(c)	Ratio
(d)	Ratio
(e)	Interval
(f)	Ratio
(g)	Nominal
(h)	Ordinal
(i)	Interval

- **(a) Political party voted for in an election**

An: Nominal
Different parties are just **names/categories** with **no natural order**.

- **(b) The difficulty of different levels in a computer game**

An: Ordinal
Levels have a **ranking** (Easy < Medium < Hard), but the **difference between levels is not measurable**.

- **(c) Production time of a car**

An: Ratio
Time has a **true zero** and allows meaningful comparisons (e.g., 4 hours is twice 2 hours).

- **(d) Age of turtles**

An: Ratio
Age has a **true zero** (birth) and supports meaningful ratios (10 years is twice 5 years).

- **(e) Calendar year**

An: Interval
Difference between years is meaningful, but **zero is arbitrary** (there is no true “year 0”).

- **(f) Price of a chocolate bar**

An: Ratio
Price has a **true zero** (₹0 means no cost), and ratios make sense (₹40 is twice ₹20).

- **(g) Identification number of a student**

An: Nominal
ID number is just a **label** — it has **no quantity or order**.

- **(h) Final ranking at a beauty contest**

An: Ordinal
Ranks show **order**, but the **difference between ranks** (1st, 2nd, 3rd) is **not measurable**.

- **(i) Intelligence quotient (IQ)**

An:Interval
IQ has meaningful **differences**, but **zero is arbitrary** and ratios (e.g., 120 is not "twice as intelligent" as 60) do **not** make sense.

Solution to Exercise 1.4

(a) The choice of a political party is measured on a nominal scale. The names of the parties do not have a natural order.
(b) Typically the level of a computer game is measured on an ordinal scale: for example, level 10 may be more difficult than level 5, but this does not imply that level 10 is twice as difficult as level 5, or that the difference in difficulty between levels 2 and 3 is the same as the difference between levels 10 and 11.
(c) The production time of a car is measured on a continuous scale (ratio scale). In practice, it may be measured in days from the start of the production.
(d) This variable is measured on a continuous scale (ratio scale). Typically, the age is captured in years starting from the day of birth.
(e) Calender year is a continuous variable which is measured on an interval scale. Note that the year which we define as "zero" is arbitrary, and it varies from culture to culture. Because the year zero is arbitrary, and we also have dates before this year, the calender year is measured on an interval scale.
(f) The scale is continuous (ratio scale).
(g) The scale of ID numbers is nominal. The ID number may indeed consist of numbers; however, "112233" does not refer to something half as much/good as "224466". The number is descriptive.
(h) The final rank is measured on an ordinal scale. The ranks can be clearly ordered, and the participants can be ranked by using their final results. However the first winner may not have "double" the beauty of the second winner, it is merely a ranking.
(i) The intelligence quotient is a variable on a continuous scale. It is constructed in such a way that differences are interpretative—i.e. being 10 points above or 10 points below the average score of 100 points means the same deviation from the average. However, ratios cannot be interpreted, so the intelligence quotient is measured on an interval scale.

Measures of Central Tendency

- A data set may contain many variables and observations.
- However, we are not always interested in each of the measured values but rather in a summary which interprets the data.
- Statistical functions can be used to summarize the data in a meaningful yet concise way.

Example 3.0.1 Suppose someone from Munich (Germany) plans a holiday in Bangkok (Thailand) during the month of December and would like to get information about the weather when preparing for the trip. Suppose last year's maximum temperatures during the day (in degrees Celsius) for December 1–31 are as follows:

22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26,

25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.

1. Arithmetic mean

- The **arithmetic mean** is one of the most intuitive measures of central tendency.
- Suppose a variable of size n consists of the values x1, x2, . . . , xn.
- The arithmetic mean of this data is defined as

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

- To calculate the arithmetic mean for grouped data, we need the following frequency table:

Class intervals a_j	$a_1 = e_0 - e_1$	$a_2 = e_1 - e_2$	...	$a_k = e_{k-1} - e_k$
Absolute freq. n_j	n_1	n_2	...	n_k
Relative freq. f_j	f_1	f_2	...	f_k

Note that $a_1, a_2, \ldots, a_k$ are the k class intervals and each interval $a_j\,(j = 1, 2, \ldots, k)$ contains n_j observations with $\sum_{j=1}^{k} n_j = n$. The relative frequency of the jth class is $f_j = n_j/n$ and $\sum_{j=1}^{k} f_j = 1$. The mid-value of the jth class interval is defined as $m_j = (e_{j-1} + e_j)/2$, which is the mean of the lower and upper limits of the interval. The **weighted arithmetic mean** for grouped data is defined as

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j m_j = \sum_{j=1}^{k} f_j m_j.$$

Example 3.0.1 Suppose someone from Munich (Germany) plans a holiday in Bangkok (Thailand) during the month of December and would like to get information about the weather when preparing for the trip. Suppose last year's maximum temperatures during the day (in degrees Celsius) for December 1–31 are as follows:

22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26,
25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.

- The arithmetic mean is therefore

$$\bar{x} = \frac{22 + 24 + 21 + \cdots + 28 + 29}{31} = 26.48\,^{\circ}\mathrm{C}.$$

Example 3.0.1 Suppose someone from Munich (Germany) plans a holiday in Bangkok (Thailand) during the month of December and would like to get information about the weather when preparing for the trip. Suppose last year's maximum temperatures during the day (in degrees Celsius) for December 1–31 are as follows:

22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26,
25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.

- Let us assume the data in the above is summarized in categories as follows:

Class intervals	< 20	$(20 - 25]$	$(25, 30]$	$(30, 35]$	> 35
Absolute frequencies	$n_1 = 0$	$n_2 = 12$	$n_3 = 18$	$n_4 = 1$	$n_5 = 0$
Relative frequencies	$f_1 = 0$	$f_2 = \frac{12}{31}$	$f_3 = \frac{18}{31}$	$f_4 = \frac{1}{31}$	$f_5 = 0$

- We can calculate the (weighted) arithmetic mean as

$$\bar{x} = \sum_{j=1}^{k} f_j m_j = 0 + \frac{12}{31} \cdot 22.5 + \frac{18}{31} \cdot 27.5 + \frac{1}{31} 32.5 + 0 \approx 25.7.$$

- The weighted mean is meant to estimate the arithmetic mean in those situations where only grouped data is available. It is therefore typically used to obtain an approximation of the true mean.

Properties of the Arithmetic Mean.

(i) The sum of the deviations of each variable around the arithmetic mean is zero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0. \tag{3.3}$$

(ii) If the data is linearly transformed as $y_i = a + bx_i$, where a and b are known constants, it holds that

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = \frac{1}{n}\sum_{i=1}^{n} a + \frac{b}{n}\sum_{i=1}^{n} x_i = a + b\bar{x}. \tag{3.4}$$

- Recall previous Examples where we considered the temperatures in December in Bangkok. We measured them in degrees Celsius, but someone from the USA might prefer to know them in degrees Fahrenheit. With a linear transformation, we can create a new temperature variable as

$$\text{Temperature in °F} = 32 + 1.8\,\text{Temperature in °C}.$$

Using $\bar{y} = a + b\bar{x}$, we get $\bar{y} = 32 + 1.8 \cdot 26.48 \approx 79.7\,°\text{F}$.

- **Increasing all salaries by ₹1000**
- Let average salary = ₹30,000
  New salary = old salary + 1000
- New average:
- $\bar{y} = 1000 + 30{,}000 = ₹31{,}000$

- **Example 3: Scaling marks**
- If marks are doubled:
- $y = 2x$
- Mean also doubles:
- $\bar{y} = 2\bar{x}$

2. Median and Quantiles

- The median is the value which divides the observations into two equal parts such that at least 50% of the values are greater than or equal to the median and at least 50% of the values are less than or equal to the median.
- The median is denoted by $\tilde{x}_{0.5}$; then, in terms of the empirical cumulative distribution function, the condition $F(\tilde{x}_{0.5}) = 0.5$ is satisfied.
- Consider the n observations $x_1, x_2, \ldots, x_n$ which can be ordered as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.
- The calculation of the median depends on whether the number of observations n is odd or even.
- When n is odd, then $\tilde{x}_{0.5}$ is the middle ordered value.
- When n is even, then $\tilde{x}_{0.5}$ is the arithmetic mean of the two middle ordered values:

$$\tilde{x}_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{if } n \text{ is even.} \end{cases}$$

- Consider the previous example where we evaluated the temperature in Bangkok in December.The ordered values x(i), i = 1, 2, . . . , 31, are as follows:

°C	21	22	22	23	24	24	25	25	25	25	25	25	26	26	26	26
(i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
°C	27	27	27	28	28	28	29	29	29	29	29	30	30	30	31	
(i)	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	

- We have n = 31

  therefore ˜x0.5 = x((n+1)/2) = x((31+1)/2) = x(16) = 26.

  - Therefore, at least 50% of the 31 observations are greater than or equal to 26 and at least 50%are less than or equal to 26.

If we deal with grouped data, we can calculate the median under the assumption that the values within each class are equally distributed. Let $K_1, K_2, \ldots, K_k$ be k classes with observations of size $n_1, n_2, \ldots, n_k$, respectively. First, we need to determine which class is the median class, i.e. the class that includes the median. We define the median class as the class K_m for which

$$\sum_{j=1}^{m-1} f_j < 0.5 \quad \text{and} \quad \sum_{j=1}^{m} f_j \geq 0.5 \tag{3.6}$$

hold. Then, we can determine the median as

$$\tilde{x}_{0.5} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m \tag{3.7}$$

where e_{m-1} denotes the lower limit of the interval K_m and d_m is the width of the interval K_m.

Example 3.1.4 Recall Example 3.1.1 where we looked at the grouped temperatur data:

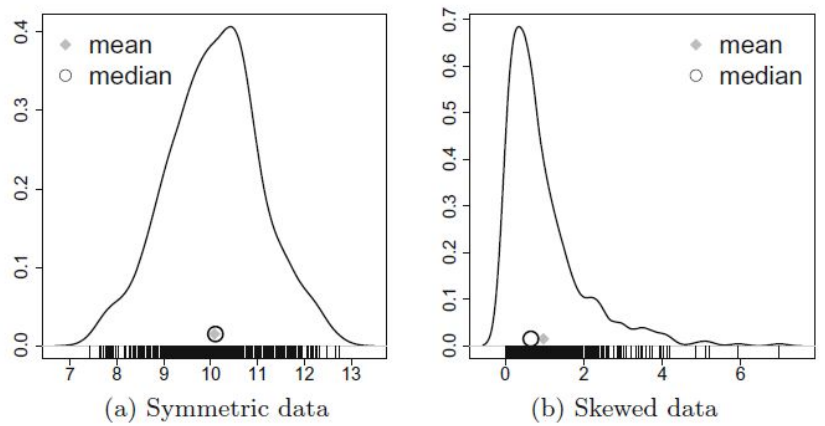
Class intervals	<20	(20–25]	(25, 30]	(30, 35]	>35
n_j	$n_1 = 0$	$n_2 = 12$	$n_3 = 18$	$n_4 = 1$	$n_5 = 0$
f_j	$f_1 = 0$	$f_2 = \frac{12}{31}$	$f_3 = \frac{18}{31}$	$f_4 = \frac{1}{31}$	$f_5 = 0$
$\sum_j f_j$	0	$\frac{12}{31}$	$\frac{30}{31}$	1	1

For the third class ($m = 3$), we have

$$\sum_{j=1}^{m-1} f_j = \frac{12}{31} < 0.5 \quad \text{and} \quad \sum_{j=1}^{m} f_j = \frac{30}{31} \geq 0.5.$$

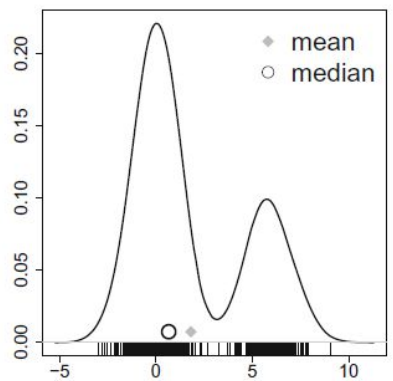
We can therefore calculate the median as

$$\tilde{x}_{0.5} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m = 25 + \frac{0.5 - \frac{12}{31}}{\frac{18}{31}} \cdot 5 \approx 25.97.$$

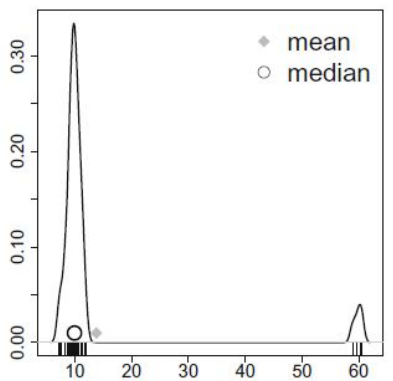


(a) Symmetric data

(b) Skewed data



(c) Bimodal data



(d) Data with outliers

Arithmetic mean and median for different data

3. Quantiles

- Quantiles are a generalization of the idea of the median.
- The median is the value which splits the data into two equal parts.
- Similarly, a quantile partitions the data into other proportions.
- For example, a 25 %-quantile splits the data into two parts such that at least 25% of the values are less than or equal to the quantile and at least 75% of the values are greater than or equal to the quantile.
- In general, let α be a number between zero and one.
- The ($\alpha \times 100$)%-quantile, denoted as $\tilde{x}_\alpha$, is defined as the value which divides the data in proportions of ($\alpha \times 100$)% and $(1 - \alpha) \times 100$% such that at least $\alpha \times 100$% of the values are less than or equal to the quantile and at least $(1 - \alpha) \times 100$% of the values are greater than or equal to the quantile.

- In terms of the empirical cumulative distribution function, we can write
- $F(\tilde{x}_\alpha) = \alpha$.
- It follows immediately that for n observations, at least nα values are less than or equal to $\tilde{x}_\alpha$ and at least n(1 − α) observations are greater than or equal to $\tilde{x}_\alpha$
- The median is the 50 %-quantile ˜x0.5.
- If α takes the values 0.1, 0.2, . . . , 0.9, the quantiles are called deciles.
- If α takes the values 0.2, 0.4, 0.6, and 0.8, the quantiles are known as quintiles and they divide the data into five equal parts.
- If α takes the values 0.25, 0.5, and 0.75, the quantiles are called quartiles.

Name	α values	Meaning
Median	0.5	Middle of data
Quartiles	0.25, 0.5, 0.75	4 equal parts
Quintiles	0.2, 0.4, 0.6, 0.8	5 equal parts
Deciles	0.1 to 0.9	10 equal parts
Percentiles	0.01 to 0.99	100 equal parts

Consider n ordered observations $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The $\alpha \cdot 100\,\%$-quantile $\tilde{x}_\alpha$ is calculated as

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{if } n\alpha \text{ is not an integer number,} \\ & \text{choose } k \text{ as the smallest integer } > n\alpha, \\ \frac{1}{2}(x_{(n\alpha)} + x_{(n\alpha+1)}) & \text{if } n\alpha \text{ is an integer.} \end{cases} \tag{3.8}$$

- Consider the previous example. The ordered values x(i), i = 1, 2, . . . , 31 are as follows:

°C	21	22	22	23	24	24	25	25	25	25	25	25	26	26	26	26
(i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
°C	27	27	27	28	28	28	29	29	29	29	29	30	30	30	31	
(i)	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	

To determine the quartiles, i.e. the 25, 50, and 75 % quantiles, we calculate $n\alpha$ as $31 \cdot 0.25 = 7.75, 31 \cdot 0.5 = 15.5$, and $31 \cdot 0.75 = 23.25$. Using (3.8), it follows that

$$\tilde{x}_{0.25} = x_{(8)} = 25, \quad \tilde{x}_{0.5} = x_{(16)} = 26,$$
$$\tilde{x}_{0.75} = x_{(24)} = 29.$$

3. Quantile–Quantile Plots (QQ-Plots)

- If we plot the quantiles of two variables against each other, we obtain a Quantile–Quantile plot (QQ-plot).
- This provides a simple summary of whether the distributions of the two variables are similar with respect to their location or not.

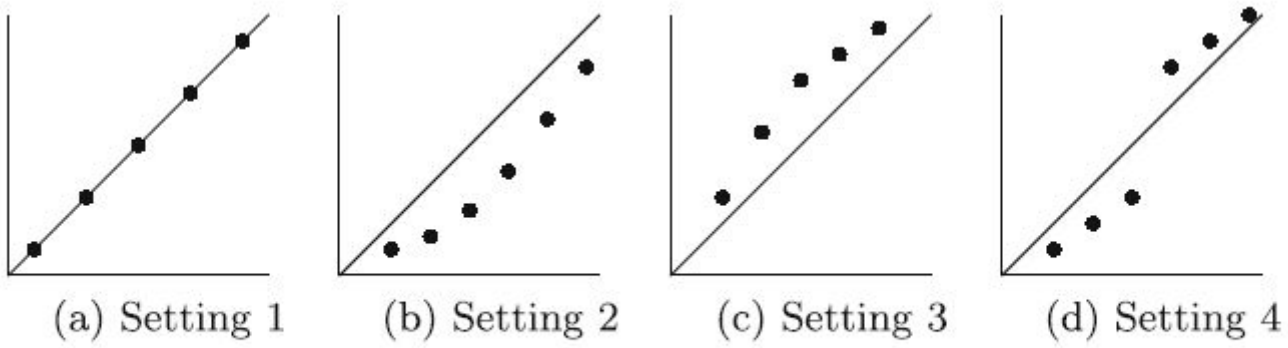


Fig. 3.3 Different patterns for a QQ-plot

- (a) If all the pairs of quantiles lie (nearly) on a straight line at an angle of 45% from the x-axis, then the two samples have similar distributions (Fig. 3.3a).
- (b) If the y-quantiles are lower than the x-quantiles, then the y-values have a tendency to be lower than the x-values (Fig. 3.3b).
- (c) If the x-quantiles are lower than the y-quantiles, then the x-values have a tendency to be lower than the y-values (Fig. 3.3c).
- (d) If the QQ-plot is like Fig. 3.3d, it indicates that there is a break point up to which the y-quantiles are lower than the x-quantiles and after that point, the y-quantiles are higher than the x-quantiles.

- Eg:X = temperatures in City A
- Y = temperatures in City B
- Each variable has many observations, not just one value.
- Compute the same quantiles (same α values) for both datasets and compare them.
- So for α = 0.1, 0.2, 0.3, …, 0.9:
  - compute the α-quantile of X
  - compute the α-quantile of Y
- Each α gives one pair:
  - (quantile *of X*, *quantile of Y*)
- These pairs are plotted in a QQ-plot.

- Let X (Math marks): $50, 55, 60, 65, 70$

  Y (Physics marks): $40, 45, 50, 55, 60$
- Step 2: Compute the same quantiles
- Let's take $\alpha = 0.2, 0.4, 0.6, 0.8$

(\alpha)	Quantile of X (Math)	Quantile of Y (Physics)
0.2	52.5	42.5
0.4	57.5	47.5
0.6	62.5	52.5
0.8	67.5	57.5

- **Form pairs**
- (52.5,42.5),(57.5,47.5),(62.5,52.5),(67.5,57.5)
- A QQ-plot compares **the same percentiles of two datasets** to see how their distributions differ in location and shape.
- If we plot this we get graph similar to the second diagram, All Y-quantiles are lower than X-quantiles ie, Y-values tend to be lower than X-values.

4.Mode

- The mode $\bar{x}_M$ of n observations x1, x2, . . . , x_n is the value which occurs the most compared with all other values, i.e. the value which has maximum absolute frequency.
- It may happen that two or more values occur with the same frequency in which case the mode is not uniquely defined.
- A formal definition of the mode is

$$\bar{x}_M = a_j \Leftrightarrow n_j = \max\{n_1, n_2, \ldots, n_k\}.$$

- The mode is typically applied to any type of variable for which the number of different values is not too large.
- If continuous data is summarized in groups, then the mode can be used as well.

5. Geometric Mean

- Consider n observations x1, x2, . . . , xn which are all positive and collected on a quantitative variable.
- The geometric mean $\bar{x}_G$ of this data is defined as

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} .$$

- The geometric mean plays an important role in fields where we are interested in products of observations, such as when we look at percentage changes in quantities.

- Suppose we have a starting value at some baseline time point 0 (zero), which may be denoted as B_0.
- At time t , this value may have changed and we therefore denote it as B_t , t = 1, 2, . . . , T.
- The ratio of B_t andB_{t-1}, is called the the growth factor.

$$x_t = \frac{B_t}{B_{t-1}},$$

- The growth rate rt is defined as

$$r_t = ((x_t - 1) \cdot 100)\,\%$$

Time t	Inventory B_t	Growth factor x_t	Growth rate r_t
0	B_0	–	–
1	B_1	$x_1 = B_1/B_0$	$((x_1 - 1) \cdot 100)\,\%$
2	B_2	$x_2 = B_2/B_1$	$((x_2 - 1) \cdot 100)\,\%$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	B_T	$x_T = B_T/B_{T-1}$	$((x_T - 1) \cdot 100)\,\%$

We can calculate B_t $(t = 1, 2, \ldots, T)$ by using the growth factors:

$$B_t = B_0 \cdot x_1 \cdot x_2 \cdot \ldots \cdot x_t.$$

The average growth factor from B_0 to B_T is the geometric mean or geometric average of the growth factors:

$$\begin{aligned} \bar{x}_G &= \sqrt[T]{x_1 \cdot x_2 \cdot \ldots \cdot x_T} \\ &= \sqrt[T]{\frac{B_0 \cdot x_1 \cdot x_2 \cdot \ldots \cdot x_T}{B_0}} \\ &= \sqrt[T]{\frac{B_T}{B_0}}. \end{aligned} \tag{3.11}$$

Therefore, B_t at time t can be calculated as $B_t = B_0 \cdot \bar{x}_G^t$.

Example 3.1.8 Suppose someone wants to deposit money, say €1000, in a bank. The bank advisor proposes a 5-year savings plan with the following plan for interest rates: 1 % in the first year, 1.5 % in the second year, 2.5 % in the third year, and 3 % in the last 2 years. Now he would like to calculate the average growth factor and average growth rate for the invested money. The concept of the geometric mean can be used as follows:

Example 3.1.8 Suppose someone wants to deposit money, say €1000, in a bank. The bank advisor proposes a 5-year savings plan with the following plan for interest rates: 1 % in the first year, 1.5 % in the second year, 2.5 % in the third year, and 3 % in the last 2 years. Now he would like to calculate the average growth factor and average growth rate for the invested money. The concept of the geometric mean can be used as follows:

Year	Euro	Growth factor	Growth rate (%)
0	1000	–	–
1	1010	1.01	1.0
2	1025.15	1.015	1.5
3	1050.78	1.025	2.5
4	1082.30	1.03	3.0
5	1114.77	1.03	3.0

The geometric mean is calculated as

$$\bar{x}_G = (1.01 \cdot 1.015 \cdot 1.025 \cdot 1.03 \cdot 1.03)^{\frac{1}{5}} = 1.021968$$

which means that he will have on average about 2.2 % growth per year. The savings after 5 years can be calculated as

$$€\,1000 \cdot 1.021968^5 = €\,1114.77.$$

It is easy to compare two different saving plans with different growth strategies using the geometric mean.

6. Harmonic Mean

- The harmonic mean is typically used whenever different x_i contribute to the mean with a different weight w_i , i.e. when we implicitly assume that the weight of each x_i is not one.
- It can be calculated as

$$\bar{x}_H = \frac{w_1 + w_2 + \cdots + w_k}{\frac{w_1}{x_1} + \frac{w_2}{x_2} + \cdots + \frac{w_k}{x_k}} = \frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{k} \frac{w_i}{x_i}}.$$

For example, when calculating the average speed, each weight relates to the relative distance travelled, n_i/n, with speed x_i. Using $w_i = n_i/n$ and $\sum_i w_i = \sum_i n_i/n = 1$, the harmonic mean can be written as

$$\bar{x}_H = \frac{1}{\sum_{i=1}^{k} \frac{w_i}{x_i}}. \tag{3.13}$$

Example 3.1.9 Suppose an investor bought shares worth €1000 for two consecutive months. The price for a share was €50 in the first month and €200 in the second month. What is the average purchase price? The number of shares purchased in the first month is $1000/50 = 20$. The number of shares purchased in the second month is $1000/200 = 5$. The total number of shares purchased is thus $20 + 5 = 25$, and the total investment is €2000. It is evident that the average purchase price is $2000/25 = €80$. This is in fact the harmonic mean calculated as

$$\bar{x}_H = \frac{1}{\frac{0.5}{50} + \frac{0.5}{200}} = 80$$

because the weight of each purchase is $n_i/n = 1000/2000 = 0.5$. If the investment was €1200 in the first month and €800 in the second month, then we could use the harmonic mean with weights $1200/2000 = 0.6$ and $800/2000 = 0.4$, respectively, to obtain the results.

- An investor invests €1000 in each of two months.
  Share price: Month 1: €50, Month 2: €200
- We are not averaging prices directly.

  We are finding: Average $price = \frac{Total\ money\ spent}{Total\ shares\ bought}$

- Month 1

$$\text{Shares} = \frac{1000}{50} = 20$$

- Month 2

$$\text{Shares} = \frac{1000}{200} = 5$$

- Total shares:

$$20 + 5 = 25$$

- Total money invested:

$$1000 + 1000 = 2000$$

Average price=2000/25=80

Weighted harmonic mean:

$$\bar{x}_H = \frac{1}{\frac{w_1}{x_1} + \frac{w_2}{x_2}}$$

Prices:

- $x_1 = 50$
- $x_2 = 200$

Each month has equal investment:

$$w_1 = w_2 = \frac{1000}{2000} = 0.5$$

- Therefore

$$\bar{x}_H = \frac{1}{\frac{0.5}{50} + \frac{0.5}{200}}$$

$$= \frac{1}{0.01 + 0.0025} = \frac{1}{0.0125} = 80$$

(b) A hiking enthusiast has a new app for his smartphone which summarizes his hikes by using a GPS device. Let us look at the distance hiked (in km) and maximum altitude (in m) for the last 10 hikes: **(6)**

Distance	12.5	29.9	14.8	18.7	7.6	16.2	16.5	27.4	12.1	17.5
Altitude	342	1245	502	555	398	670	796	912	238	466

Calculate the arithmetic mean and median for both distance and altitude.

Measures of Dispersion

- Measures of central tendency, give us an idea about the location where most of the data is concentrated.
- However, two different data sets may have the same value for the measure of central tendency, say the same arithmetic means, but they may have different concentrations around the mean.
- In this case, the location measures may not be adequate enough to describe the distribution of the data.
- The concentration or dispersion of observations around any particular value is another property which characterizes the data and its distribution.
- We now introduce statistical methods which describe the **variability** or **dispersion** of data.

1.Range and Interquartile Range

Consider a variable X with n observations $x_1, x_2, \ldots, x_n$. Order these n observations as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The range is a measure of dispersion defined as the difference between the maximum and minimum value of the data as

$$R = x_{(n)} - x_{(1)}. \tag{3.14}$$

The **interquartile range** is defined as the difference between the 75th and 25th quartiles as

$$d_Q = \tilde{x}_{0.75} - \tilde{x}_{0.25}. \tag{3.15}$$

It covers the centre of the distribution and contains 50 % of the observations.

- Recall our previous examples where we looked at the temperature in Bangkok during December.
- The ordered values x(i), i = 1, . . . , 31, are as follows:

°C	21	22	22	23	24	24	25	25	25	25	25	25	26	26	26	26
(i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
°C	27	27	27	28	28	28	29	29	29	29	29	30	30	30	31	
(i)	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	

- We obtained the quantiles in Example 3.1.5 as $\tilde{x}_{0.25} = 25$ and $\tilde{x}_{0.75} = 29$. The interquartile range is therefore $d_Q = 29 - 25 = 4$, which means that 50% of the data is centred between 25 and 29 ◦C. The range is $R = 31 - 21 = 10$ ◦C, meaning that the temperature is varying at most by 10 ◦C.

2. Absolute Deviation,Variance, and Standard Deviation

Consider the deviations of n observations around a certain value "A" and combine them together, for instance, via the arithmetic mean of all the deviations:

$$D = \frac{1}{n}\sum_{i=1}^{n}(x_i - A). \tag{3.16}$$

This measure has the drawback that the deviations $(x_i - A)$, $i = 1, 2, \ldots, n$, can be either positive or negative and, consequently, their sum can potentially be very small or even zero. Using D as a measure of variability is therefore not a good idea since D may be small even for a large variability in the data.

Using absolute values of the deviations solves this problem, and we introduce the following measure of dispersion:

$$D(A) = \frac{1}{n}\sum_{i=1}^{n}|x_i - A|. \tag{3.17}$$

It can be shown that the absolute deviation attains its minimum when A corresponds to the median of the data:

$$D(\tilde{x}_{0.5}) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \tilde{x}_{0.5}|. \tag{3.18}$$

We call $D(\tilde{x}_{0.5})$ the **absolute median deviation**. When $A = \bar{x}$, we speak of the **absolute mean deviation** given by

$$D(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|. \tag{3.19}$$

Another solution to avoid the positive and negative signs of deviation in (3.16) is to consider the squares of deviations $x_i - A$, rather than using the absolute value. This provides another measure of dispersion as

$$s^2(A) = \frac{1}{n}\sum_{i=1}^{n}(x_i - A)^2 \tag{3.20}$$

which is known as the **mean squared error** (MSE) with respect to A. The MSE is another important measure in statistics, see Chap. 9, Eq. (9.4), for details. It can be shown that $s^2(A)$ attains its minimum value when $A = \bar{x}$. This is the (sample) **variance**

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{3.21}$$

After expanding $\tilde{s}^2$, we can write (3.21) as

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2. \tag{3.22}$$

The positive square root of the variance is called the (sample) **standard deviation**, defined as

$$\tilde{s} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{3.23}$$

- The standard deviation has the same unit of measurement as the data whereas the unit of the variance is the square of the units of the observations.
- The variance is a measure which we use to obtain measures of association between variables and to draw conclusions from a sample about a population of interest.
- The standard deviation is typically preferred for a descriptive summary of the dispersion of data.
- The standard deviation measures how much the observations vary or how they are dispersed around the arithmetic mean.
- A low value of the standard deviation indicates that the values are highly concentrated around the mean.
- A high value of the standard deviation indicates lower concentration of the observations around the mean, and some of the observed values may even be far away from the mean.
- If there are extreme values or outliers in the data, then the arithmetic mean is more sensitive to outliers than the median. In such a case, the absolute median deviation may be preferred over the standard deviation.

- Example 3.2.1 Suppose three students Christine, Andreas, and Sandro arrive at different times in the class to attend their lectures. Let us look at their arrival time in the class after or before the starting time of lecture, i.e. let us look how early or late they were (in minutes).

Week	1	2	3	4	5	6	7	8	9	10
Christine	0	0	0	0	0	0	0	0	0	0
Andreas	−10	+10	−10	+10	−10	+10	−10	+10	−10	+10
Sandro	3	5	6	2	4	6	8	4	5	7

$$\tilde{s}_C^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{1}{10}((0-0)^2 + \cdots + (0-0)^2) = 0$$

$$\tilde{s}_A^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{1}{10}((-10-0)^2 + \cdots + (10-0)^2) \approx 111.1$$

$$\tilde{s}_S^2 = \frac{1}{10}\sum_{i=1}^{10}(x_i - \bar{x})^2 = \frac{1}{10}((3-5)^2 + \cdots + (7-5)^2) \approx 3.3$$

$$D(\tilde{x}_{0.5,C}) = \frac{1}{10}\sum_{i=1}^{n}|x_i - \tilde{x}_{0.5}| = |0-0| + \cdots + |0-0| = 0$$

$$D(\tilde{x}_{0.5,A}) = \frac{1}{10}\sum_{i=1}^{n}|x_i - \tilde{x}_{0.5}| = |-10-0| + \cdots + |10-0| = 10$$

$$D(\tilde{x}_{0.5,S}) = \frac{1}{10}\sum_{i=1}^{n}|x_i - \tilde{x}_{0.5}| = |3-5| + \cdots + |7-5| = 1.4.$$

Variance for Grouped Data. The variance for grouped data can be calculated using

$$s_b^2 = \frac{1}{n}\sum_{j=1}^{k} n_j (a_j - \bar{x})^2 = \frac{1}{n}\left(\sum_{j=1}^{k} n_j a_j^2 - n\bar{x}^2\right) = \frac{1}{n}\sum_{j=1}^{k} n_j a_j^2 - \bar{x}^2, \tag{3.24}$$

where a_j is the middle value of the jth interval. However, when the data is artificially grouped and the knowledge about the original ungrouped data is available, we can also use the arithmetic mean of the jth class:

$$s_b^2 = \frac{1}{n}\sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2. \tag{3.25}$$

If the data within each class is known, we can use the Theorem of Variance Decomposition (see p. 136 for the theoretical background) to determine the variance. This allows us to represent the total variance as the sum of the **variance between the different classes** and the **variance within the different classes** as

$$\tilde{s}^2 = \underbrace{\frac{1}{n}\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2}_{\text{between}} + \underbrace{\frac{1}{n}\sum_{j=1}^{k} n_j \tilde{s}_j^2}_{\text{within}}. \tag{3.27}$$

In (3.27), $\tilde{s}_j^2$ is the variance of the jth class:

$$\tilde{s}_j^2 = \frac{1}{n_j}\sum_{x_i \in K_j} (x_i - \bar{x}_j)^2. \tag{3.28}$$

Linear Transformations. Let us consider a linear transformation $y_i = a + bx_i$ $(b \neq 0)$ of the original data x_i, $(i = 1, 2, \ldots, n)$. We get the arithmetic mean of the transformed data as $\bar{y} = a + b\bar{x}$ and for the variance:

$$\begin{aligned} \tilde{s}_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{b^2}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= b^2\tilde{s}_x^2. \end{aligned} \tag{3.29}$$

Standardization. A variable is called standardized if its mean is zero and its variance is 1. Standardization can be achieved by using the following transformation:

$$y_i = \frac{x_i - \bar{x}}{\tilde{s}_x} = -\frac{\bar{x}}{\tilde{s}_x} + \frac{1}{\tilde{s}_x}x_i = a + bx_i. \tag{3.30}$$

3.Coefficient of Variation

- The coefficient of variation v is a measure of dispersion which uses both the standard deviation and mean and thus allows a fair comparison.
- It is properly defined only when all the values of a variable are measured on a ratio scale and are positive such that $\bar{x} > 0$ holds.
- It is defined as

$$v = \frac{s}{\bar{x}}.$$

- The coefficient of variation is a unit-free measure of dispersion.
- It is often used when the measurements of two variables are different but can be put into relation by using a linear transformation $y_i = bx_i$.
- It is possible to show that if all values x_i of a variable X are transformed into a variable Y with values $y_i = b \cdot x_i$, $b > 0$, then v does not change.

Example 3.2.8 If we want to compare the variability of hotel prices in two selected cities in Germany and England, we could calculate the mean prices, together with their standard deviation. Suppose a sample of prices of say 100 hotels in two selected cities in Germany and England is available and suppose we obtain the mean and standard deviations of the two cities as $x_1 = €130$, $x_2 = £230$, $s_1 = €99$, and $s_2 = £212$. Then, $v_1 = 99/130 \approx 0.72$ and $v_2 = 212/230 = 0.92$. This indicates higher variability in hotel prices in England. However, if the data distribution is skewed or bimodal, then it may be wise not to choose the arithmetic mean as a measure of central tendency and likewise the coefficient of variation.

Association of two variables

- When two variables are not independent, then they are associated. Their association can be weak or strong.
- Measures of association describe the degree of association between two variables and can have a direction as well.
- If variables are defined on a nominal scale, then nothing can be said about the direction of association, only about the strength.

Summarizing the Distribution of Two Discrete Variables

- When both variables are discrete, then it is possible to list all combinations of values of the two variables and to count how often these combinations occur in the data.
- When we have two discrete variables, instead of listing all raw data again and again, we summarize the data using a contingency table.
- A contingency table simply shows how often each combination of categories occurs.

Contingency Tables for Discrete Data

- Suppose we have data on two discrete variables. This data can be described in a two-dimensional **contingency table**.
- Example 4.1.1: An airline conducts a customer satisfaction survey. The survey includes questions about travel class and satisfaction levels with respect to different categories such as seat comfort, in-flight service, meals, safety, and other indicators. Consider the information on X, denoting the travel class (Economy = “E”, Business = “B”, First = “F”), and “Y”, denoting the overall satisfaction with the flight on a scale from 1 to 4 as 1 (poor), 2 (fair), 3 (good), and 4 (very good).A possible response from 12 customers may look as follows:

	Passenger number											
i	1	2	3	4	5	6	7	8	9	10	11	12
Travel class	E	E	E	B	E	B	F	E	E	B	E	B
Satisfaction	2	4	1	3	1	2	4	3	2	4	3	3

- We can calculate the absolute frequencies for each of the combination of observed values. For example, there are 2 passengers (passenger numbers 3 and 5) who were flying in economy class and rated the flight quality as poor, there were no passengers from both business class and first class who rated the flight quality as poor; there were 2 passengers who were flying in economy class and rated the quality as fair (2), and so on

Table 4.1 Contingency table for travel class and satisfaction

		Overall rating of flight quality				Total (row)
		Poor	Fair	Good	Very good	
Travel class	Economy	2	2	2	1	7
	business	0	1	2	1	4
	first	0	0	0	1	1
	Total (column)	2	3	4	3	12

- Now we extend this example and discuss a general framework to summarize the absolute frequencies of two discrete variables in contingency tables. We use the following notations: Let x1, x2, . . . , xk be the k classes of a variable X and let y1, y2, . . . , yl be the l classes of another variable Y . We assume that both X and Y are discrete variables. It is possible to summarize the absolute frequencies ni j related to (xi , y j), i = 1, 2, . . . , k, j = 1, 2, . . . , l, in a k × l **contingency table** as shown below.

Table 4.2 $k \times l$ contingency table

		Y					
		y_1		y_j		y_l	Total (rows)
X	x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	n_{1+}
	x_2	n_{21}	$\cdots$	n_{2j}	$\cdots$	n_{2l}	n_{2+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}	n_{i+}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	n_{k+}
	Total (columns)	n_{+1}	$\cdots$	n_{+j}	$\cdots$	n_{+l}	n

We denote the sum of the ith row as $n_{i+} = \sum_{j=1}^{l} n_{ij}$ and the sum over the jth column as $n_{+j} = \sum_{i=1}^{k} n_{ij}$. The total number of observations is therefore

$$n = \sum_{i=1}^{k} n_{i+} = \sum_{j=1}^{l} n_{+j} = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} \,. \tag{4.1}$$

We denote the sum of the ith row as $n_{i+} = \sum_{j=1}^{l} n_{ij}$ and the sum over the jth column as $n_{+j} = \sum_{i=1}^{k} n_{ij}$. The total number of observations is therefore

$$n = \sum_{i=1}^{k} n_{i+} = \sum_{j=1}^{l} n_{+j} = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} \,. \tag{4.1}$$

Remark 4.1.1 Note that it is also possible to use the relative frequencies $f_{ij} = n_{ij}/n$ instead of the absolute frequencies n_{ij} in Table 4.2, see Example 4.1.2.

- Let us first consider a 2 × 2 contingency table which is a special case of a k × l contingency table,

		Y		
		y_1	y_2	Total (row)
X	x_1	a	b	$a+b$
	x_2	c	d	$c+d$
	Total (column)	$a+c$	$b+d$	n

The variables X and Y are independent if

$$\frac{a}{a+c} = \frac{b}{b+d} = \frac{a+b}{n}$$

or equivalently if

$$a = \frac{(a+b)(a+c)}{n}.$$

Example 4.2.1 Suppose a vaccination against flu (influenza) is given to 200 persons. Some of the persons may get affected by flu despite the vaccination. The data is summarized in Table 4.6. Using the notations of Table 4.5, we have $a = 90, b = 10, c = 40, d = 60$, and thus, $(a + b)(a + c)/n = 100 \cdot 130/200 = 65$ which is less than $a = 90$. Hence, being affected by flu is not independent of the vaccination, i.e. whether one is vaccinated or not has an influence on getting affected by flu. In the vaccinated group, only 10 of 100 persons are affected by flu while in the group not vaccinated 60 of 100 persons are affected. Another interpretation is that if independence holds, then we would expect 65 persons to be not affected by flu in the vaccinated group but we observe 90 persons. This shows that vaccination has a protective effect.

		Persons		
		Not affected	Affected	Total (row)
Vaccination	Vaccinated	90	10	100
	Not vaccinated	40	60	100
	Total (column)	130	70	200

Pearson's x^2 Statistic

- used for measuring the association between variables in a contingency table.
- The x^2 statistic or x^2coefficient for a k × l contingency table is given as

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}.$$

A simpler formula for 2×2 contingency tables is

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

- The idea behind the x^2 coefficient is that when the relationship between two variables is stronger, then the deviations between observed and expected frequencies are expected to be higher (because the expected frequencies are calculated assuming independence) and this indicates a stronger relationship between the two variables.

- If observed and expected frequencies are identical or similar, then this is an indication that the association between the two variables is weak and the variables may even be independent.
- The x^2statistic for a k × l contingency table sums up all the differences between the observed and expected frequencies, squares them, and scales them with respect to the expected frequencies.
- The squaring of the difference makes the statistic independent of the positive and negative signs of the difference between observed and expected frequencies.
- The range of values for x^2 is

$$0 \leq \chi^2 \leq n(\min(k, l) - 1).$$

- A value of x^2 close to zero indicates a weak association and a value of x^2 close to n(min(k, l) − 1) indicates a strong association between the two variables.
- Note that the range of x^2 depends on n, k and l, i.e. the sample size and the dimension of the contingency table.
- The x^2 statistic is a symmetric measure in the sense that its value does not depend on which variable is defined as X and which as Y .

- Find x^2 value for this data.

Table 4.3 Contingency table for travel class and satisfaction

		Overall rating of flight quality				
		Poor	Fair	Good	Very good	Total (rows)
Travel class	Economy	10	33	15	4	62
	Business	0	3	20	2	25
	First	0	0	5	8	13
	Total (columns)	10	36	40	14	100

Example 4.2.2 Consider Examples 4.1.2 and 4.1.4. Using the values from Table 4.4, we can calculate the χ^2 statistic as

$$\chi^2 = \frac{(10-6.2)^2}{6.2} + \frac{(33-22.32)^2}{22.32} + \cdots + \frac{(8-1.82)^2}{1.82} = 57.95064$$

- *The maximum possible value for the x^2 statistic is 100(min(4, 3) − 1) = 200. Thus, $x^2 \approx 57$ indicates a moderate association between "travel class" and "overall rating of flight quality" of the passengers.*

Cramer's **V** Statistic

- A problem with Pearson's x^2 coefficient is that the range of its maximum value depends on the sample size and the size of the contingency table.
- These values may vary in different situations.
- To overcome this problem, the coefficient can be standardized to lie between 0 and 1 so that it is independent of the sample size as well as the dimension of the contingency table.
- Since n(min(k, l) − 1) was the maximal value of the x^2 statistic, dividing x^2 by this maximal value automatically leads to a scaled version with maximal value 1.

- Cramer's V statistic which for a k × l contingency table is given by

$$V = \sqrt{\frac{\chi^2}{n(\min(k,l)-1)}}$$

- The closer the value of V gets to 1, the stronger the association between the two variables.

Example 4.2.3 Consider Example 4.2.2. The obtained χ^2 statistic is 57.95064. To obtain Cramer's V, we just need to calculate

$$V = \sqrt{\frac{\chi^2}{n(\min(k, l) - 1)}} = \sqrt{\frac{57.95064}{100(3 - 1)}} \approx 0.54. \qquad (4.10)$$

- This indicates a moderate association between "travel class" and "overall rating of flight quality" because 0.54 lies in the middle of 0 and 1.

Contingency Coefficient **C**

- Another option to standardize x^2 is given by a corrected version of Pearson's contingency coefficient:

$$C_{\text{corr}} = \frac{C}{C_{\max}} = \sqrt{\frac{\min(k,l)}{\min(k,l)-1}}\sqrt{\frac{\chi^2}{\chi^2+n}},$$

with

$$C = \sqrt{\frac{\chi^2}{\chi^2+n}} \quad \text{and} \quad C_{\max} = \sqrt{\frac{\min(k,l)-1}{\min(k,l)}}.$$

- It always lies between 0 and 1. The closer the value of C is to 1, the stronger the association.

Example 4.2.4 We know from Example 4.2.2 that the χ^2 statistic for travel class and satisfaction level is 57.95064. To calculate C_{corr}, we need the following calculations:

$$C = \sqrt{\frac{57.95064}{57.95064 + 100}} = 0.606, \quad C_{\max} = \sqrt{\frac{\min(4, 3) - 1}{\min(4, 3)}} = \sqrt{\frac{2}{3}} = 0.816,$$

$$C_{\text{corr}} = \frac{C}{C_{\max}} = \frac{0.606}{0.816} \approx 0.74 .$$

- There is a moderate to strong association between “travel class” and “overall rating of flight quality” of the passengers.

Relative Risks and Odds Ratios

- Relative risk and odds ratio are used specifically when one variable is a binary exposure and the other is a binary outcome, and the aim is to compare risks between groups.
- An exposure variable answers to the questions like
  - Was the person exposed to something?
- Examples:
  - Smoking: Yes / No
  - Drug given: Yes / No
  - Radiation exposure: Yes / No
  - Treatment received: Yes / No

Exposure usually represents a cause or risk factor.

- What is an *outcome variable*?
- An outcome variable usually answers "*What happened as a result?*"
- Examples:
  - Disease: Yes / No
  - Recovery: Yes / No
  - Death: Yes / No
  - Test positive: Yes / No
- Outcome represents the effect or result.

Relative Risks and Odds Ratios

- Suppose we have two variables X and Y with their conditional distributions $f_{i|j}^{X|Y}$ and $f_{j|i}^{Y|X}$.
- In the context of a 2 × 2 contingency table,

$$f_{1|1}^{X|Y} = n_{11}/n_{+1},$$

$$f_{1|2}^{X|Y} = n_{12}/n_{+2},.$$

$$f_{2|2}^{X|Y} = n_{22}/n_{+2}$$

$$f_{2|1}^{X|Y} = n_{21}/n_{+1}.$$

- The relative risks are defined as the ratio of two conditional distributions, for example

$$\frac{f_{1|1}^{X|Y}}{f_{1|2}^{X|Y}} = \frac{n_{11}/n_{+1}}{n_{12}/n_{+2}} = \frac{a/(a+c)}{b/(b+d)} \quad \text{and} \quad \frac{f_{2|1}^{X|Y}}{f_{2|2}^{X|Y}} = \frac{n_{21}/n_{+1}}{n_{22}/n_{+2}} = \frac{c/(a+c)}{d/(b+d)}.$$

- The odds ratio is defined as the ratio of these relative risks

$$\frac{f_{1|1}^{X|Y}/f_{1|2}^{X|Y}}{f_{2|1}^{X|Y}/f_{2|2}^{X|Y}} = \frac{f_{1|1}^{X|Y} f_{2|2}^{X|Y}}{f_{2|1}^{X|Y} f_{1|2}^{X|Y}} = \frac{a\,d}{b\,c}$$

- The relative risks compare proportions, while the odds ratio compares odds.

Example 4.2.5 A classical example refers to the possible association of smoking with a particular disease. Consider the following data on 240 individuals:

		Smoking		Total (row)
		Yes	No	
Disease	Yes	34	66	100
	No	22	118	140
	Total (column)	56	184	240

We calculate the following relative risks:

$$\frac{f_{1|1}^{X|Y}}{f_{1|2}^{X|Y}} = \frac{34/56}{66/184} \approx 1.69 \quad \text{and} \quad \frac{f_{2|1}^{X|Y}}{f_{2|2}^{X|Y}} = \frac{22/56}{118/184} \approx 0.61\,. \tag{4.15}$$

- Thus, the proportion of individuals with the disease is 1.69 times higher among smokers when compared with non-smokers. Similarly, the proportion of healthy individuals is 0.61 times smaller among smokers when compared with non-smokers.

- The relative risks are calculated to compare the proportion of sick or healthy patients between smokers and non-smokers. Using these two relative risks, the odds ratio is obtained as

$$OR = \frac{34 \times 118}{66 \times 22} = 2.76.$$

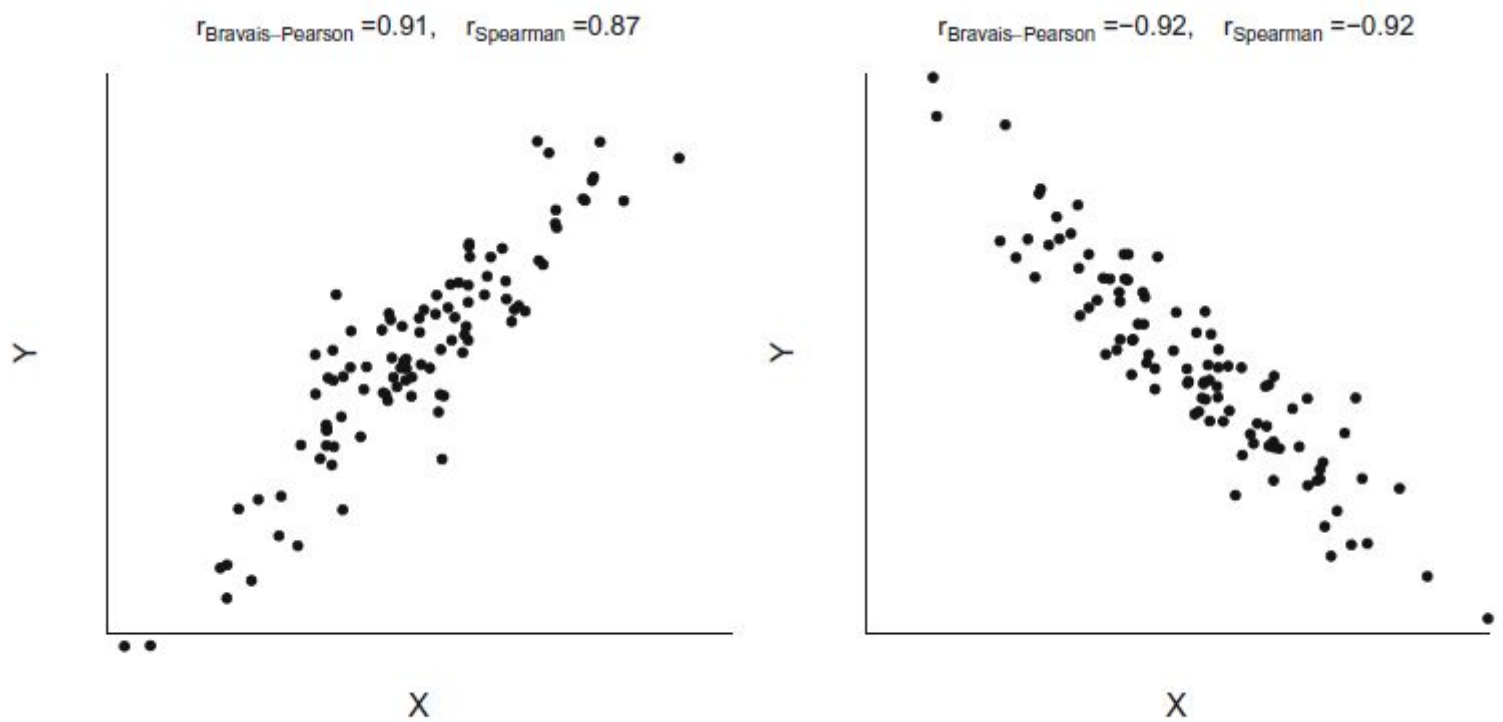
- We can interpret this outcome as follows:
- (i) the chances of smoking are 2.76 times higher for individuals with the disease compared with healthy individuals .

- ii) the chances of having the particular disease is 2.76 times higher for smokers compared with non-smokers.
- If we interchange either one of the “Yes” and “No” columns or the “Yes” and “No” rows, we obtain OR = 1/2.76 ≈ 0.36, giving us further interpretations:
- (iii) the chances of smoking are 0.36 times lower for individuals without disease compared with individuals with the disease, and
- (iv) the chance of having the particular disease is 0.36 times lower for non-smokers compared with smokers.

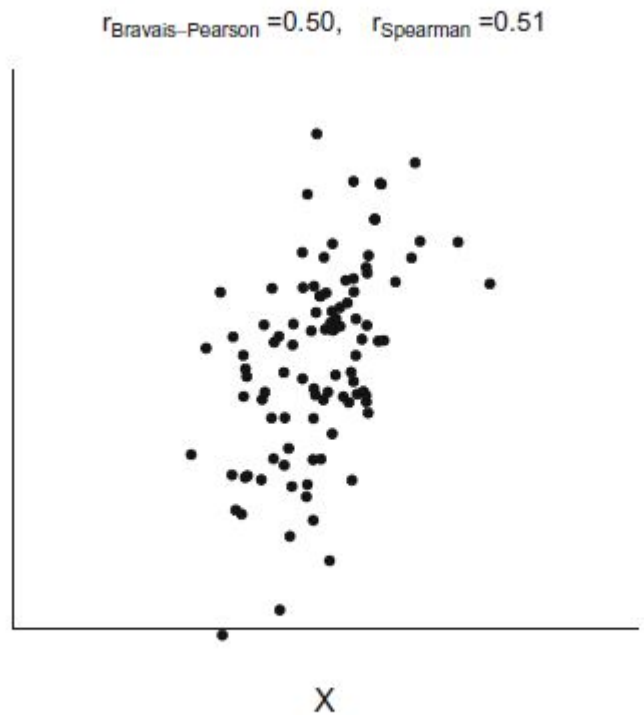
Association Between Ordinal and Continuous Variables

Graphical Representation of Two Continuous Variables

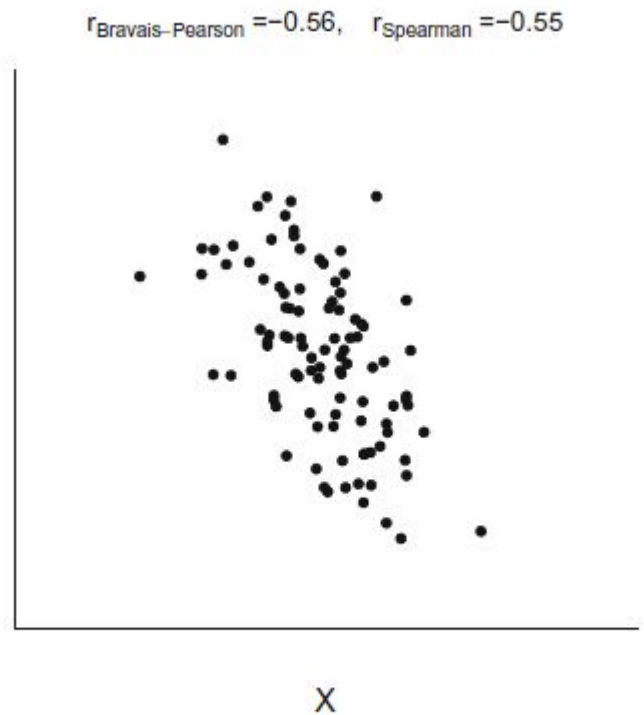
- A simple way to graphically summarize the association between two continuous variables is to plot the paired observations of the two variables in a two-dimensional coordinate system.
- If n paired observations for two continuous variables X and Y are available as (xi , yi), i = 1, 2, . . . , n, then all such observations can be plotted in a single graph.
- This graph is called a **scatter plot**



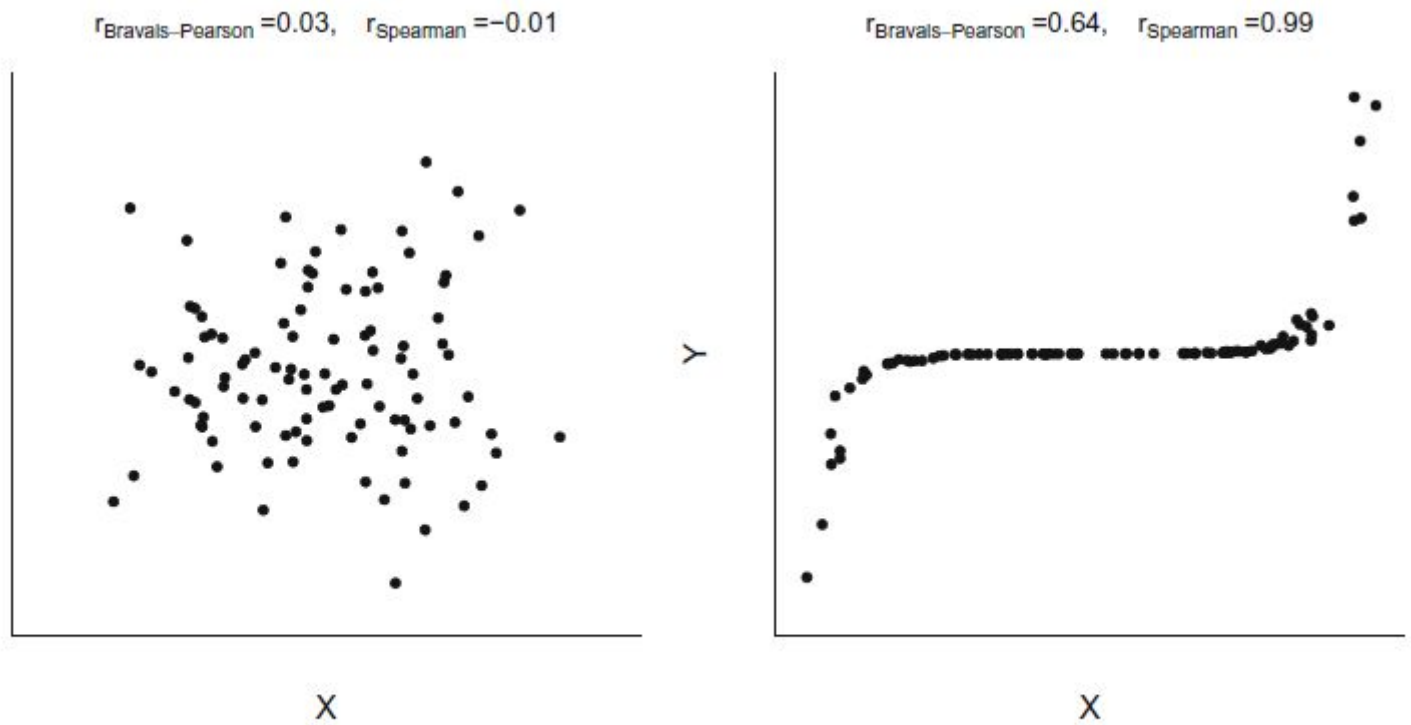
(a) Strong positive linear relationship (b) Strong negative linear relationship



(c) Moderate positive relationship



(d) Moderate negative relationship



(a) No clear relationship

(b) Nonlinear relationship

Correlation Coefficient

- Suppose two variables X and Y are measured on a continuous scale and are linearly related like Y = a + b X where a and b are constant values.
- The **correlation coefficient** r (X, Y) = r measures the degree of linear relationship between X and Y using

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

with

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = n\tilde{s}_X^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = n\tilde{s}_Y^2,$$

and

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\,.$$

- In a decathlon competition, a group of athletes are competing with each other in 10 different track and field events. Suppose we are interested in how the results of the 100-m race relate to the results of the long jump competition. The correlation coefficient for the 100-m race (X, in seconds) and the long jump event (Y , in metres) for 5 athletes participating in the 2004 Olympic Games.

i	x_i	y_i
Roman Sebrle	10.85	7.84
Bryan Clay	10.44	7.96
Dmitriy Karpov	10.50	7.81
Dean Macey	10.89	7.47
Chiel Warners	10.62	7.74

$$\bar{x} = \frac{1}{5}(10.85 + 10.44 + 10.50 + 10.89 + 10.62) = 10.66$$

$$\bar{y} = \frac{1}{5}(7.84 + 7.96 + 7.81 + 7.47 + 7.74) = 7.764$$

$$S_{xx} = (10.85 - 10.66)^2 + (10.44 - 10.66)^2 + \cdots + (10.62 - 10.66)^2 = 0.1646$$

$$S_{yy} = (7.84 - 7.764)^2 + (7.96 - 7.764)^2 + \cdots + (7.74 - 7.764)^2 = 0.13332$$

$$S_{xy} = (10.85 - 10.66)(7.84 - 7.764) + \cdots + (10.62 - 10.66)(7.74 - 7.764)$$
$$= -0.1027$$

The correlation coefficient therefore is

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{-0.1027}{\sqrt{0.1646 \times 0.13332}} \approx -0.69 .$$

- Since −0.69 is negative, we can say that (i) there is a negative correlation between the 100-m race and the long jump event, i.e., shorter running times result in longer long jump results, and (ii) this association is moderate to strong.

Spearman's Rank Correlation Coefficient

- The correlation coefficient is independent of the units of measurement of X and Y .
- The correlation coefficient is symmetric, i.e. r(X, Y) = r(Y, X).
- The limits of r are $-1 \leq r \leq 1$.
- If all the points in a scatter plot lie exactly on a straight line, then the linear relationship between X and Y is perfect and $|r| = 1$
- If the relationship between X and Y is (i) perfectly linear and increasing, then r = +1 and (ii) perfectly linear and decreasing, then r = −1.

- The signs of r thus determine the direction of the association. If r is close to zero, then it indicates that the variables are independent or the relationship is not linear.
- Note that if the relationship between X and Y is nonlinear, then the degree of linear relationship may be low and r is then close to zero even if the variables are clearly not independent.

- To measure the degree of agreement, or, in general, the degree of association, one can use **Spearman's rank correlation coefficient.**
- As the name says, this correlation coefficient uses only the ranks of the values and not the values themselves.
- Thus, this measure is suitable for both ordinal and continuous variables. We introduce the following notations:
- Let R(xi) denote the rank of the ith observation on X, i.e. the rank xi among the ordered values of X.
- Similarly, R(yi) denotes the rank of the ith observation of y.
- The difference between the two rank values is di = R(xi) − R(yi).

- Spearman's rank correlation coefficient is defined as

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}.$$

- The values of R lie between −1 and +1 and measure the degree of correlation between the ranks of X and Y .
- Note that it does not matter whether we choose an ascending or descending order of the ranks, the value of R remains the same.
- When all the observations are assigned exactly the same ranks, then R = 1 and when all the observations are assigned exactly the opposite ranks, then R = −1.

- Calculate Spearman's rank correlation coefficient for the first five observations of the decathlon data. Again we list the results of the 100-m race (X) and the results of the long jump competition (Y). In addition, we assign ranks to both X and Y . For example, the shortest time receives rank 1, whereas the longest time receives rank 5. Similarly, the shortest long jump result receives rank 5.

i	x_i	$R(x_i)$	y_i	$R(y_i)$	d_i	d_i^2
Roman Sebrle	10.85	4	7.84	4	0	0
Bryan Clay	10.44	1	7.96	5	−4	16
Dmitriy Karpov	10.50	2	7.81	3	−1	1
Dean Macey	10.89	5	7.47	1	−4	16
Chiel Warners	10.62	3	7.74	2	−1	1
Total						34

- Spearman's rank correlation coefficient can be calculated as

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 34}{5 \cdot 24} = -0.7.$$

- We therefore have a moderate to strong negative association between the 100-m race and the long jump event.
- We now know that for the 5 athletes above longer running times relate to shorter jumping distances which in turn means that a good performance in one discipline implies a good performance in the other discipline.

- **Correlation (r)** measures:
  - Strength and direction of a *linear* relationship between two variables.
  - Ie,
  - Do X and Y increase/decrease together in a straight-line manner?
  - How strong is that straight-line trend?

If the relationship is:

- perfectly linear increasing → r = +1
- perfectly linear decreasing → r = −1
- nonlinear (curve-shaped) → r ≈ 0 (even if clearly related!)

- Spearman's rank correlation (ρ or R) measures:
  - Strength and direction of a *monotonic* relationship between two variables.
- A monotonic relationship is a relationship where:
  - As X increases, Y consistently increases OR
  - As X increases, Y consistently decreases

- It ignores actual values
- Uses only the order (ranks)
- ie it tells us When X increases in rank, does Y also increase (or decrease) in rank?

- **ρ = +1**
  - Whenever X increases in rank, Y also increases in rank exactly.
  - Example:
  - Tallest student → highest rank in weight
  - Second tallest → second highest weight
- **ρ = –1**
  - Higher rank in X always corresponds to lower rank in Y.
  - Example:
  - Higher class rank number → lower marks
    (Rank 1 = highest marks)

- **ρ ≈ 0.7**
  - There is a strong tendency that higher X values correspond to higher Y values, but not perfectly.

- **ρ ≈ 0**
  - Knowing the rank of X gives no useful information about the rank of Y.