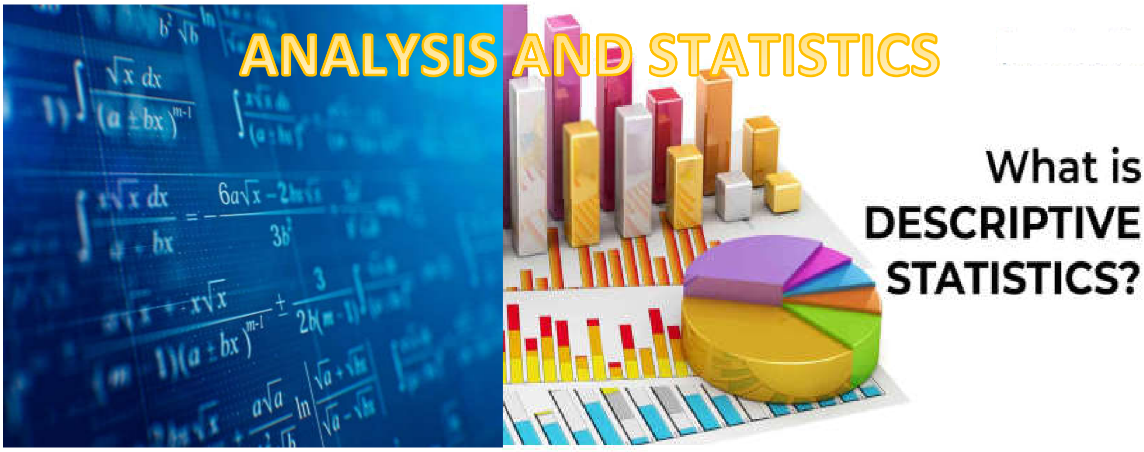
MOHAMED KHIDER UNIVERSITY, BISKRA

FACULTY OF EXACTE SCIENCES

DEPARTMENT OF MATHEMATICS

COURSE OF ANALYSIS AND STATISTICS



Presented by

Dr. LAIADI ABDELKADER

November 2025

Contents

List of Figures

Introduction

This handout is part of the official program of the Mathematics and Statistics module intended mainly for students in the first year of a biology degree, but may possibly be useful for students in other mathematics modules (within the framework of the L.M.D. system), and anyone wishing to know continued functions, derivatives, the integral of a function and also descriptive statistics.

The required mathematical level is that of the first year of a Mathematics, economics, physics or S.T. degree. The content of this subject is the basis of any introduction to mathematical analysis. It allows the student to acquire the maximum of mathematical techniques necessary for most of the subjects studied throughout their course, namely: probability and statistics, numerical analysis, mathematical programming and...etc.

Descriptive statistics is a basic area of study in statistics. It demonstrates the relationship between variables in a given sample, and it is often used to clean up and summarize scattered data, which is crucial for making inferential statistical comparisons and conducting research [2]. Descriptive statistics are often used in statistical modeling. A statistical model expresses the previous knowledge of the probability experiment that produced the observed data [3]. The model assumes that the observation X is created by one of the model's probability distributions[3]. There are multiple measurements in descriptive statistics. There are measures of central tendency, frequency, position, and dispersion, and they also contain categories of variables such as ratio, intervals, and nominal and ordinal variables [2]. Many concepts within descriptive statistics can also be explained using basic calculus, and those descriptive statistics can be put into analyzing real-life data such as housing prices and the stock market.

This handout contains three main chapters, where are presented

1) Some definitions and theorems on limits, continuity and derivatives of functions.

2) Some methods for calculating the integral of a function.

3) Introduction to descriptive statistics.

Each chapter ends with some corrected exercises to check the acquisition of the essential concepts that have been introduced.

I could not end this foreword without a great, more personal tribute to my colleague professors who have seriously examined this handout. I would also like to thank our readers in particular. Finally, errors may be found, please report them to the author.

Chapter 1

CONTINUOUS AND DERIVATIVE FUNCTIONS

1.1 Domain of a function

Definition 1. The **domain** of a function f consists of all values of x for which the value $f(x)$ is defined. It denoted by $dom\ f(x)$ or D_f. We can write

$$D_f = \{x \in R \backslash f(x) \text{ is defined}\}$$

Definition 2. The **range** of a function is the set of values that the function assumes. This set is the values that the function shoots out after we plug an x value in. They are the y values.

All the problem is asking you is to find what values of x can be plugged into the function. This is useful to know since some functions have limits on what is permissible as an input. For instance, consider the function:

$$f(x) = \frac{1}{x}$$

We know that we can never divide by 0 so here our domain can not include the value $x = 0$. However, all other values of x would be OK. We can plug in any other number into our function and we would get an output. If we ever put a number into a function and we can't get an output then we know that there is some sort of domain issue. The most common occurrences of this happen with:

- dividing by 0
- negative square roots
- negative logs

Examples. 1) The domain of $f(x) = \frac{e^x+1}{x^2-2}$ consists of all real numbers except for $x = \sqrt{2}$ and $x = -\sqrt{2}$, since for those numbers division by zero occurs.

One can write the answer as either

$$\begin{aligned} D_f &= \{x \in R \backslash x \neq \pm\sqrt{2}\} \\ &= R - \{\sqrt{2}, -\sqrt{2}\} \\ &= \Big]-\infty, -\sqrt{2}\Big[\cup \Big]-\sqrt{2}, \sqrt{2}\Big[\cup \Big]\sqrt{2}, +\infty\Big[\end{aligned}$$

2) The domain of $f(x) = \sqrt{x^2 - 5x + 6}$ consists of all numbers x for which $x^2 - 5x + 6 \geq 0$ (in order for the square root to assume a real value). Since $x^2 - 5x + 6 = (x-2)(x-3)$ (we use the formula for roots of a quadratic equation, $x = \frac{5 \pm \sqrt{\Delta}}{2}, \Delta = b^2 - 4ac = 1$), we would like to find x for which $(x-2)(x-3) \geq 0$.

x	$-\infty$		2		3		$+\infty$
$x^2 - 5x + 6$		$+$	0	$-$	0	$+$	

Thus, the domain of f is

$$D_f =]-\infty, 2] \cup [3, +\infty[$$

3) The domain of $f(x) = \ln(1 - x^2)$ consists of all numbers x for which $1 - x^2 > 0$. Since

$1 - x^2 = (1 - x)(1 + x)$,

x	$-\infty$		-1		1		$+\infty$
$1 - x^2$		$-$	0	$+$	0	$-$	

Hence, the domain of f is

$$D_f =]-1, 1[$$

1.2 limit of functions

Definition 1 Suppose L denotes a finite number. We say the limite of $f(x)$ as x aproaches a is L when $f(x)$ approaches to L as x aproaches a, we write

$$\lim_{x \longrightarrow a} f(x) = L$$

Two-sides limits the right hand limit

$$\lim_{x \longrightarrow a^+} f(x) = L \text{ or } \lim_{x \overset{>}{\longrightarrow} a} f(x) = L$$

the left hand limit

$$\lim_{x \longrightarrow a^-} f(x) = L \text{ or } \lim_{x \overset{<}{\longrightarrow} a} f(x) = L$$

Proposition: The limit $f(x)$ exists if and only if

$$\lim_{x \longrightarrow a^-} f(x) = \lim_{x \longrightarrow a^+} f(x) = L$$

1.2.1 Properties of limits

Suppose that f, g two functions defined on $I \subset R$.

1) If $f(x) \leq g(x)$ for all $x \in I$ and $\lim_{x \longrightarrow a} f(x)$, $\lim_{x \longrightarrow a} g(x)$ exist, then

$$\lim_{x \longrightarrow a} f(x) \leq \lim_{x \longrightarrow a} g(x), a \in I$$

2) Suppose that f, g and h three functions defined on $I \subset R$.

If $f(x) \leq h(x) \leq g(x)$ for all $x \in I$ and $\lim_{x \longrightarrow a} f(x) = \lim_{x \longrightarrow a} g(x) = L$, then the limit of $h(x)$ as $x \longrightarrow a$ exists and

$$\lim_{x \longrightarrow a} h(x) = L$$

1.2.2 Algebraic properties

Limits of functions respect algebraic operations.

Suppose that $a \in R$ and $\lim_{x \longrightarrow a} f(x)$, $\lim_{x \longrightarrow a} g(x)$ exist. If $\lim_{x \longrightarrow a} f(x) = L$, $\lim_{x \longrightarrow a} g(x) = M$ then

1) $\lim_{x \longrightarrow a} (f(x) \pm g(x)) = \lim_{x \longrightarrow a} f(x) \pm \lim_{x \longrightarrow a} g(x) = L \pm M$

2) $\lim_{x \longrightarrow a} (f(x) \times g(x)) = \lim_{x \longrightarrow a} f(x) \times \lim_{x \longrightarrow a} g(x) = L \times M$

3) $\lim_{x \longrightarrow a} \left(\frac{f(x)}{g(x)} \right) = \frac{\lim_{x \longrightarrow a} f(x)}{\lim_{x \longrightarrow a} g(x)} = \frac{L}{M} \, (M \neq 0)$.

Interminate forms: The cases of the interminate form (the limit is not directly exist)

$$\frac{0}{0}, \frac{\infty}{\infty}, \infty - \infty, 0 \times \infty$$

1.2.3 Hôpital's rule

L'Hôpital's rule states that for functions f and g which are defined on an open interval I and differentiable on $I \backslash \{c\}$ for a (possibly infinite) accumulation point c of I, if

$$\lim_{x \longrightarrow c} \frac{f(x)}{g(x)} = \frac{0}{0} \text{ or } \lim_{x \longrightarrow c} \frac{f(x)}{g(x)} = \frac{\infty}{\infty}$$

and $g'(x) \neq 0$ for all x in $I \backslash \{c\}$ and $\lim_{x \longrightarrow c} \frac{f'(x)}{g'(x)}$ exists, then

$$\lim_{x \longrightarrow c} \frac{f(x)}{g(x)} = \lim_{x \longrightarrow c} \frac{f'(x)}{g'(x)}$$

We note that one can not use Hôpital's rule "in reverse" to deduce that $\frac{f'}{g'}$ has a limit if $\frac{f}{g}$ has a limit.

Example 1 Let $f(x) = x + \sin x$ and $g(x) = x$. Then $\lim\limits_{x \longrightarrow +\infty} \frac{f(x)}{g(x)} = \frac{+\infty}{+\infty}$ and

$$\lim_{x \longrightarrow +\infty} \frac{f(x)}{g(x)} = \lim_{x \longrightarrow +\infty} \left(1 + \frac{\sin x}{x}\right) = 1$$

because

$$1 - \frac{1}{x} \leq 1 + \frac{\sin x}{x} \leq \frac{1}{x} + 1$$

but the limit

$$\lim_{x \longrightarrow +\infty} \frac{f'(x)}{g'(x)} = \lim_{x \longrightarrow +\infty} (1 + \cos x)$$

does not exist.

Examples. Calculate the following limits

$$1) \lim_{x \longrightarrow 0} \frac{\sqrt{x^2+4}-2}{x^2}, 2) \lim_{x \longrightarrow +\infty} \frac{2x^3+x^2+3}{x^2+1}, 3) \lim_{x \longrightarrow 0} \frac{\cos 2x - 1}{x^2},$$
$$4) \lim_{x \longrightarrow 0} \frac{\ln(x+1) - x}{x^3}, 5) \lim_{x \longrightarrow +\infty} \frac{\sqrt{x^2+x+1}}{x+1+\sqrt{x^2+x+1}}$$

Solution. Because $\lim\limits_{x \longrightarrow 0} \frac{\sqrt{x^2+4}-2}{x^2} = \frac{0}{0}$ is the interminate form $\frac{0}{0}$.However, by rationalization of the numerator (conjugate method) we obtain

$$\begin{aligned} \lim_{x \longrightarrow 0} \frac{\sqrt{x^2+4}-2}{x^2} &= \lim_{x \longrightarrow 0} \frac{\sqrt{x^2+4}-2}{x^2} \times \frac{\sqrt{x^2+4}+2}{\sqrt{x^2+4}+2} \\ &= \lim_{x \longrightarrow 0} \frac{1}{\sqrt{x^2+4}+2} \\ &= \frac{1}{4} \end{aligned}$$

2) $\lim\limits_{x \longrightarrow +\infty} \frac{2x^3+x^2+3}{x^2+1} = \lim\limits_{x \longrightarrow +\infty} \frac{2x^3}{x^2} = \lim\limits_{x \longrightarrow +\infty} 2x = +\infty.$

3) $\lim\limits_{x \longrightarrow 0} \frac{\cos 2x - 1}{x^2} = \frac{0}{0}$. It is the interminate form $\frac{0}{0}$. By using the Hôpital's rule twice then

$$\lim_{x \longrightarrow 0} \frac{\cos 2x - 1}{x^2} = \lim_{x \longrightarrow 0} \frac{-2\sin 2x}{2x} = \lim_{x \longrightarrow 0} -2\cos 2x = -2$$

4) We have by Hôpital's rule

$$\lim_{x \longrightarrow 0} \frac{\ln(x+1) - x}{x^3} = \lim_{x \longrightarrow 0} \frac{\frac{1}{x+1} - 1}{3x^2} = \lim_{x \longrightarrow 0} \frac{-1}{3(x+1)x^2} = -\infty$$

5) We have

$$\lim_{x \longrightarrow +\infty} \frac{\sqrt{x^2+x+1}}{x+1+\sqrt{x^2+x+1}} = \lim_{x \longrightarrow +\infty} \frac{\sqrt{x^2}}{x+1+\sqrt{x^2}} = \lim_{x \longrightarrow +\infty} \frac{x}{x+1+x} = \frac{1}{2}$$

1.3 Continuous Functions

Any function $y = f(x)$ whose graph can be sketched over its domain in one continuous motion without lifting the pencil is an example of a continuous function. In this lecture we investigate more precisely what it means for a function to be continuous.

Definition 1 We say the function f is continuous at a number a if

$$\lim_{x \longrightarrow a} f(x) = f(a)$$

(i.e. we can make the value of $f(x)$ as close as we like to $f(a)$ by taking x sufficiently close to a). Note that this definition implies that the function f has the following three properties if f is continuous at a:

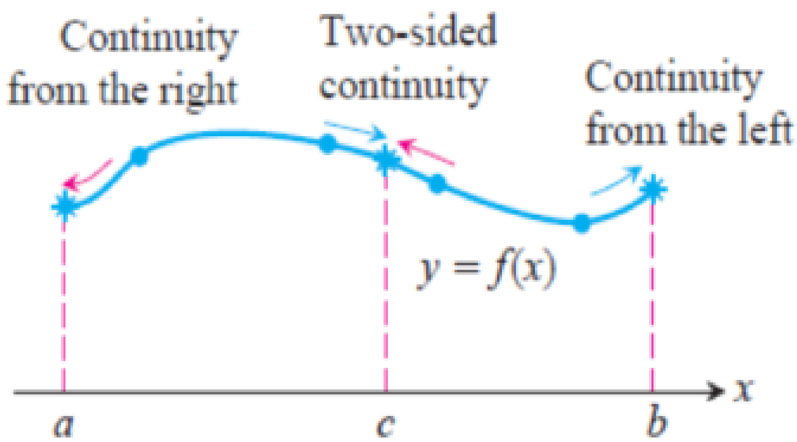
1. $f(a)$ is defined (a is in the domain of f).
2. $\lim_{x \longrightarrow a} f(x)$ exists.
3. $\lim_{x \longrightarrow a} f(x) = f(a)$.

Definition 2 1) A function f is continuous from the right at a number a if $\lim_{x \longrightarrow a^+} f(x) = f(a) \left(\lim_{x \overset{>}{\longrightarrow} a} f(x) = f(a) \right)$.

2) A function f is continuous from the left at a number a if $\lim_{x \longrightarrow a^-} f(x) = f(a) \left(\lim_{x \overset{<}{\longrightarrow} a} f(x) = f(a) \right)$.

Remark A function f is continuous at a number a if

$$\lim_{x \longrightarrow a^+} f(x) = \lim_{x \longrightarrow a^-} f(x) = f(a)$$



Example 1 Is the function

$$f(x) = \begin{cases} x+1, x \leq 1 \\ 2, 1 < x \leq 2 \\ \frac{1}{x-3}, x > 2 \end{cases}$$

continuous at $x = 1$ and 2 ?

Solution: We could answer these questions by examining the graph of $f(x)$, but lets try them without the graph.

At $a = 1, f(1) = 2$ and the left and right limits are equal,

$$\lim_{x \overset{<}{\longrightarrow} 1} (x+1) = 2 = f(1)$$

$$\lim_{x \overset{>}{\longrightarrow} 1} 2 = f(1)$$

so f is continuous at 1.

At $a = 2, f(2) = 2$, but the left and right limits are not equal.

$$\lim_{x \overset{<}{\longrightarrow} 2} 2 = 2 = f(2)$$

$$\lim_{x \overset{>}{\longrightarrow} 2} \frac{1}{x-3} = -1 \neq f(2)$$

so f is not continuous at 2. f is continuous from the left at 2, but not from the right.

Definition 3 (Continuous function on an interval) A function f is continuous on an interval if it is continuous at every number in the interval.
(If f is defined only on one side of an endpoint of the interval, we understand continuous at the endpoint to mean continuous from the right or continuous from the left at the endpoint as appropriate.)

1.3.1 Properties of Continuous Functions

If the functions f and g are continuous at $x = a$, then the following functions $f + g$, $f.g$, $k.f$ and $\frac{f}{g}\,(g \neq 0)$ are continuous at $x = a$ where k is a constant.

Theorem

(a) Every polynomial

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_0$$

is continuous because

$$\lim_{x \longrightarrow a} P(x) = P(a)$$

(b) If $P(x)$ and $Q(x)$ are polynomials then the rational function $\frac{P(x)}{Q(x)}$ is continuous wherever it is defined provided that $Q(x) \neq 0$

(c) All trigonometric functions wherever they are defined.

Example 2

Show that the following functions are continuous everywhere on their respective domains

$$f(x) = \frac{x}{1 + x^4}, g(x) = \left| \frac{x + 1}{x^2 - 2} \right|$$

1) The numerator is a rational power of the identity function; the denominator is an everywhere-positive polynomial. Therefore, the quotient is continuous.

2) The quotient is continuous for all $x \neq \pm\sqrt{2}$ and the function is the composition of this quotient with the continuous absolute value function.

Example 3 Let

$$f(x) = \begin{cases} \frac{\sin x}{x}, x < 0 \\ x^2 + 1, x \geq 0 \end{cases}$$

1) Determine the domain of f ?

2) Is the function f continuous on R ?

Solution: 1) $D_f = R =]-\infty, 0[\cup [0, +\infty[$

2) First, we have

$$\lim_{\substack{< \\ x \longrightarrow 0}} \frac{\sin x}{x} = 1 = f(0)$$

$$\lim_{\substack{> \\ x \longrightarrow 0}} x^2 + 1 = 1 = f(0)$$

hence f is continuous at $x = 0$ and also we have

$\frac{\sin x}{x}$ is continuous on $]-\infty, 0[$

$x^2 + 1$ is continuous on $[0, +\infty[$

implies that f is contiuous on R.

1.3.2 Composition of Continuous Functions

If g is continuous at a and f is continuous at $g(a)$, then

$$\lim_{x \longrightarrow a} f \circ g(x) = \lim_{x \longrightarrow a} f(g(x)) = f(\lim_{x \longrightarrow a} (g(x))) = f(g(a))$$

so $f \circ g(x) = f(g(x))$ is continuous at a.

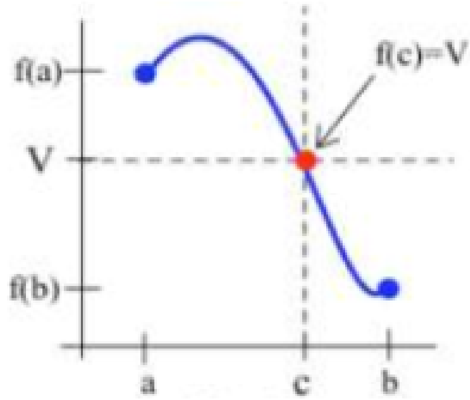
1.3.3 Intermediate Value Theorem for Continuous Functions

If f is continuous on the interval $[a, b]$ and V is any value between $f(a)$ and $f(b)$, then there is a number c between a and b so that $f(c) = V$.

(that is, f actually takes each intermediate value between $f(a)$ and $f(b)$.)

If the graph of f connects the points $(a, f(a))$ and $(b, f(b))$ and V is any number between

$f(a)$ and $f(b)$, then the graph of f must cross the horizontal line $y = V$ somewhere between $x = a$ and $x = b$ (Figure below). Since f is continuous, its graph cannot "hop" over the line $y = V$



Example 4 If we consider the function $f(x) = x^2 - 1$, on the interval $[0;2]$, we see that $f(0) = -1 < 0$ and $f(2) = 3 > 0$. Therefore the intermediate value theorem says that there must be some number c between 0 and 2 $(0 < c < 2)$ with $f(c) = 0$. The graph of $f(x)$ crosses the x axis at the point where $x = c$. What is the value of c in this case?

$$f(x) = x^2 - 1 = 0$$

This is true if $x = \pm 1$, therefore for $c = 1 \in [0;2]$, we have $f(c) = 0$.

Example 5 Check whether there is a solution to the equation $2x^3 - x^2 - 3 = 0$ between the interval $[0,2]$.

Solution:

Let us find the values of the given function at the $x = 0$ and $x = 2$.

$$f(x) = 2x^3 - x^2 - 3$$

Substitute $x = 0$ in the given function

$f(0) = 2(0)^3 - (0)^2 - 3 = -3$

Substitute $x = 2$ in the given function

$f(2) = 2(2)^3 - (2)^2 - 3 = 9$

Therefore, we conclude that at $x = 0$, the curve is below zero; while at $x = 2$, it is above zero.

Since the given equation is a polynomial, its graph will be continuous (see Figure 1.1).

Thus, applying the intermediate value theorem, we can say that the graph must cross at some point between $[0, 2]$. Hence, there exists a solution to the equation $2x^3 - x^2 - 3 = 0$ between the interval $[0, 2]$ $(f(c) = 0)$.

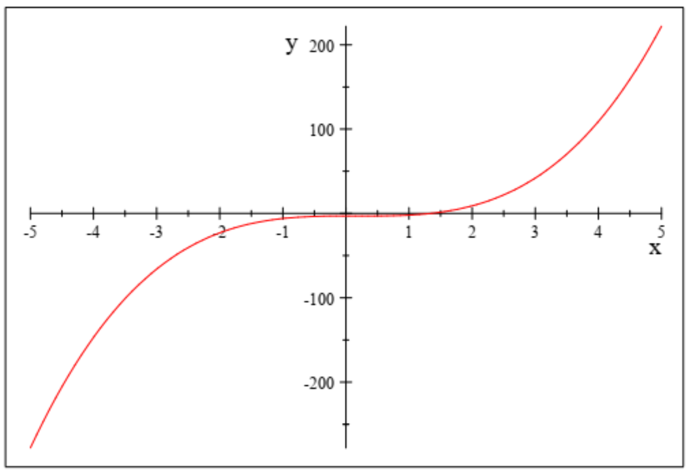


Figure 1.1: Plot of the function $2x^3 - x^2 - 3$

1.3.4 Continuous Extension to a Point

Generally, a function (such as a rational function) may have a limit even at a point where it is not defined. If $f(c)$ is not defined, but

$$\lim_{x \longrightarrow c} f(x) = L$$

exists, we can define a new function $\widetilde{f}(x)$ by the rule

$$\widetilde{f}(x) = \begin{cases} f(x), x \in D_f \\ L, x = c \end{cases}$$

The function $\widetilde{f}(x)$ is continuous at $x = c$. It is called the continuous extension of f to $x = c$. For rational functions f, continuous extensions are usually found by cancelling common factors.

Example 6

Show that $f(x)$ has a continuous extension to $x = 0$ and find that extension

$$f(x) = \frac{\sin x}{x}$$

Solution.

The domain of this function is $D_f = R^*$

Although $f(0)$ is not defined, we have (see figure 1.2)

$$\lim_{x \longrightarrow 0} f(x) = \lim_{x \longrightarrow 0} \frac{\sin x}{x} = 1$$

Hence f has a continuous extension to $x = 0$ and the extension is

$$\widetilde{f}(x) = \begin{cases} \frac{\sin x}{x}, x \in R^* \\ 1, x = 0 \end{cases}$$

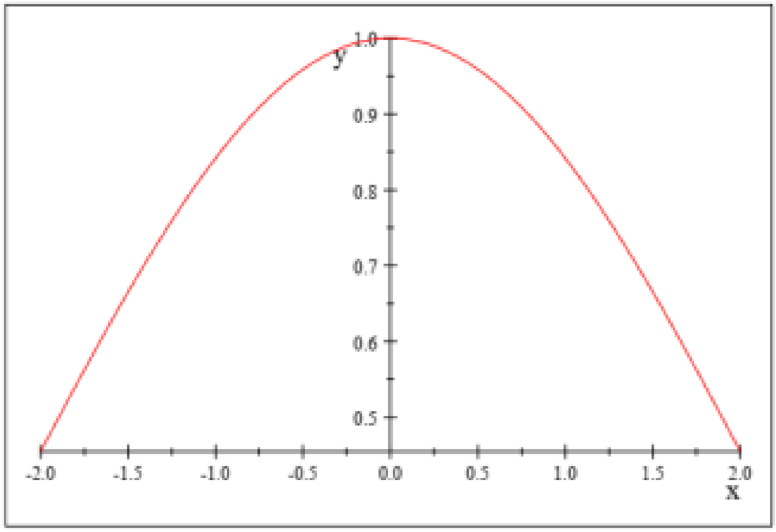


Figure 1.2: Plot of the function $\frac{\sin x}{x}$

1.4 Derivative of functions

1.4.1 Derivable functions

Definition 1 A function f is derivable at an interior point x_0 of its domain if the difference quotient

$$\frac{f(x) - f(x_0)}{x - x_0}, x \neq x_0$$

approaches a limit as x approaches x_0, in which case the limit is called the derivative of f at x_0, and is denoted by $f'(x)$; thus,

$$\lim_{x \longrightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0) \tag{1}$$

It is sometimes convenient to let $x = x_0 + h$ and write (1) as

$$\lim_{h \longrightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h} = f'(x_0)$$

For convenience we define the zeroth derivative of f to be f itself; thus

$$f^{(0)} = f$$

We assume that you are familiar with the other standard notations for derivatives; for example,

$$f^{(2)} = f'', f^{(3)} = f'''$$

and so on, and

$$\frac{d^n f}{dx^n} = f^{(n)}$$

Example 1. The function $f : R \longrightarrow R$ defined by $f(x) = x^2$ is derivable on R with derivative $f'(x) = 2x$ since

$$\lim_{x \longrightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{x \longrightarrow x_0} \frac{x^2 - x_0^2}{x - x_0} = \lim_{x \longrightarrow x_0} \frac{(x - x_0)(x + x_0)}{x - x_0} = 2x_0$$

Example 2. If n is a positive integer and

$$f(x) = x^n$$

then

$$\frac{f(x) - f(x_0)}{x - x_0} = \frac{x^n - x_0^n}{x - x_0} = \frac{x - x_0}{x - x_0} \sum_{k=0}^{n-1} x^{n-k-1} x_0^k$$

so

$$f'(x_0) = \lim_{x \longrightarrow x_0} \sum_{k=0}^{n-1} x^{n-k-1} x_0^k = n x_0^{n-1}$$

Since this holds for every x_0, we drop the subscript and write

$$f'(x) = \frac{d}{dx}(x^n) = n x^{n-1}$$

Interpretations of the Derivative

If $f(x)$ is the position of a particle at time $x \neq x_0$, the difference quotient

$$\frac{f(x) - f(x_0)}{x - x_0}$$

is the average velocity of the particle between times x_0 and x. As x approaches x_0, the average applies to shorter and shorter intervals. Therefore, it makes sense to regard the limit (1), if it exists, as the particle's instantaneous velocity at time x_0. This interpretation may be useful even if x is not time, so we often regard $f'(x)$ as the instantaneous rate of change of $f(x)$ at x_0, regardless of the specific nature of the variable x. The derivative also has a geometric interpretation. The equation of the line through two points $(x_0, f(x_0))$ and $(x_1, f(x_1))$ on the curve $y = f(x)$ (Figure 1.3) is

$$y = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0)$$

Varying x_1 generates lines through $(x_0, f(x_0))$ that rotate into the line

$$y = f(x_0) + f'(x_0)(x - x_0)$$

as x_1 approaches x_0. This is the tangent to the curve $y = f(x)$ at the point $(x_0, f(x_0))$. Figure 1.4 below depicts the situation for various values of x_1.

Here is a less intuitive definition of the tangent line: If the function

$$T(x) = f(x_0) + m(x - x_0)$$

approximates f so well near x_0 that

$$\lim_{x \longrightarrow x_0} \frac{f(x) - T(x)}{x - x_0} = 0$$

we say that the line $y = T(x)$ is tangent to the curve $y = f(x)$ at $(x_0, f(x_0))$.

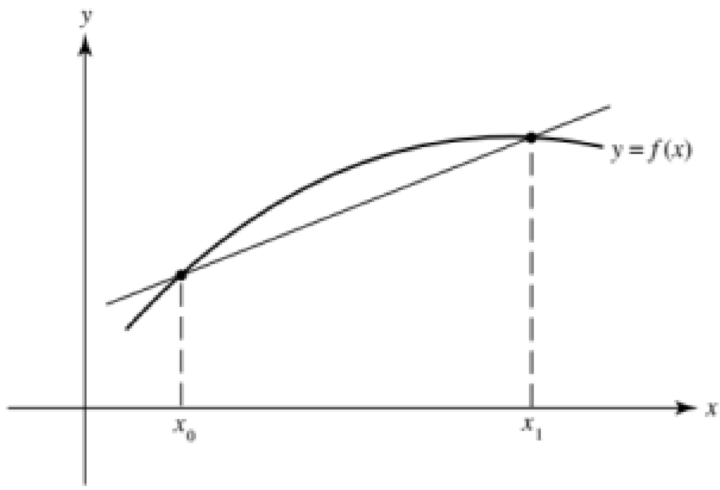


Figure 1.3: Plot $y = f\,(x)$

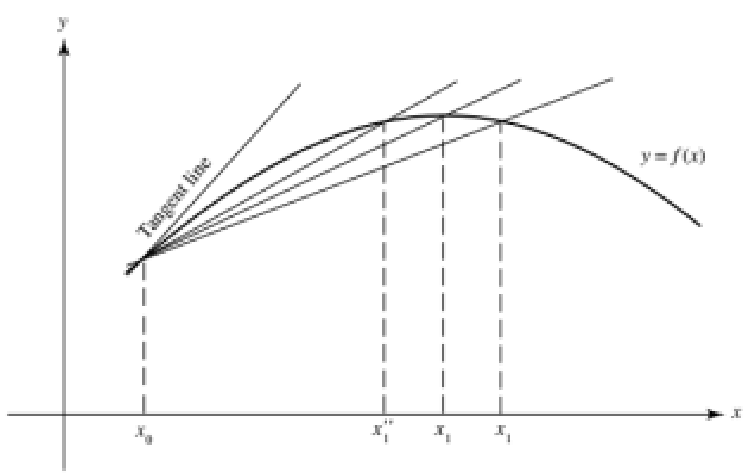


Figure 1.4: Tengent of the curve $y = f\,(x)$

Left and right derivatives. For the most part, we will use derivatives that are defined only at the interior points of the domain of a function. Sometimes, however, it is convenient to use one-sided left or right derivatives that are defined at the endpoint of an interval.

1) A function f is derivable at x_0 from the right if

$$\lim_{x \longrightarrow x_0^+} \frac{f(x) - f(x_0)}{x - x_0} \text{ exists}$$

2) A function f is derivable at x_0 from the left if

$$\lim_{x \longrightarrow x_0^-} \frac{f(x) - f(x_0)}{x - x_0} \text{ exists}$$

Remark A function f is derivable at x_0 if

$$\lim_{x \longrightarrow x_0^+} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{x \longrightarrow x_0^-} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$

Example 5 Let

$$f(x) = \begin{cases} x^2 + x + 1, x < 0 \\ e^x, x \geq 0 \end{cases}$$

Is the function f derivable at $x = 0$?

Solution we have

$$\lim_{x \longrightarrow 0^-} \frac{f(x) - f(0)}{x - 0} = \lim_{x \longrightarrow 0^-} \frac{x^2 + x + 1 - 1}{x - 0} = 1$$
$$\lim_{x \longrightarrow 0^+} \frac{f(x) - f(0)}{x - 0} = \lim_{x \longrightarrow 0^{+-}} \frac{e^x - 1}{x - 0} = 1$$

Hence f is derivable at $x = 0$

Definition 2 (Derivable function on an interval) A function f is derivable on an interval $I \subset R$ if it is derivable at every point of an interval I.

Example 6 Let

$$f(x) = \begin{cases} \frac{x^2-1}{x}, x < 0 \\ 1, x = 0 \\ \ln(x+1), x > 0 \end{cases}$$

Is the function f derivable on R?

Solution: We have $D_f = R$ **.**

First, is f derivable at $x = 0$?

we have

$$\lim_{x \longrightarrow 0^-} \frac{f(x) - f(0)}{x - 0} = \lim_{x \longrightarrow 0^-} \frac{\frac{x^2-1}{x} - 1}{x - 0} = \lim_{x \longrightarrow 0^-} \frac{x^2 - 1 - x}{x^2} = -\infty$$

and by using Hôpital's rule

$$\lim_{x \longrightarrow 0^+} \frac{f(x) - f(0)}{x - 0} = \lim_{x \longrightarrow 0^+} \frac{\ln(x+1) - 1}{x - 0} = \lim_{x \longrightarrow 0^+} \frac{1}{x+1} = 1$$

Hence f is not derivable at $x = 0$ but it is derivable on right at $x = 0$ and more than $\frac{x^2-1}{x}$ is derivable on $]-\infty, 0[$, $\ln(x+1)$ is derivable on $]0, +\infty[$.

implies that f is derivable on R^*.

1.4.2 Properties of the derivative

Theorem 1. (**derivable and continuous function**) If f is derivable at x_0; then f is continuous at x_0:

The converse of this theorem is false, since a function may be continuous at a point without being differentiable at the point.

Example 6. The function

$$f(x) = |x| = \begin{cases} x, x > 0 \\ -x, x < 0 \end{cases}$$

We note that f is continuous at $x = 0$ but we have

$$\lim_{x \longrightarrow 0^-} \frac{f(x) - f(0)}{x - 0} = \lim_{x \longrightarrow 0^-} \frac{-x}{x - 0} = -1$$
$$\lim_{x \longrightarrow 0^+} \frac{f(x) - f(0)}{x - 0} = \lim_{x \longrightarrow 0^-} \frac{x}{x - 0} = 1$$

Hence f is not derivable at $x = 0$.

Theorem 2. If f and g are differentiable at x_0; then so are $f + g, f - g, fg$ and $\frac{f}{g}\,(g \neq 0)$ with

a) $(f + g)'(x_0) = f'(x_0) + g'(x_0)$

b) $(f - g)'(x_0) = f'(x_0) - g'(x_0)$

c)$(fg)'(x_0) = f'(x_0)\, g(x_0) + f(x_0)\, g'(x_0)$

c) $\left(\frac{f}{g}\right)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{(g'(x_0))^2}$.

The Chain Rule

Here is the rule for derivative a composite function.

Theorem 3. (**The Chain Rule**) Suppose that g is differentiable at x_0 and f is derivable at $g(x_0)$: Then the composite function $h = f \circ g$; defined by

$$h(x) = f(g(x))$$

is derivable at x_0; with

$$h'(x) = f'(g(x))\, g'(x)$$

The table of derivative of elementary functions

Function f	Derivative f'
number	0
x^n	nx^{n-1}
$\ln x$	$\frac{1}{x}$
e^x	e^x
e^{ax}	ae^{ax}
$\sin x$	$\cos x$
$\cos x$	$-\sin x$
$\sqrt{x}$	$\frac{1}{2\sqrt{x}}$
$\tan x$	$\frac{1}{\cos^2 x}$
$\arcsin x$	$\frac{1}{\sqrt{1-x^2}}$
$\arccos x$	$\frac{-1}{\sqrt{1-x^2}}$
$\arctan x$	$\frac{1}{x^2+1}$

The table of derivative composite functions

Function	Derivative
f^n	$nf'f^{n-1}$
$\ln f$	$\frac{f'}{f}$
e^f	$f'e^f$
$\sqrt{f}$	$\frac{f'}{2\sqrt{f}}$
$\sin f$	$f'\cos f$
$\cos f$	$-f'\sin f$
$\tan f$	$\frac{f'}{\cos^2 f}$
$\arctan f$	$\frac{f'}{f^2+1}$

Examples 1) $f(x) = x^3 - 2x^2 + 7$, then $f'(x) = 3x^2 - 4x$

2) $f(x) = \ln(x^2+1)$, then $f'(x) = \frac{2x}{x^2+1}$

3) $f(x) = \cos^2(2x)$, by using the chain rule then $f'(x) = -2.2\cos(2x)\sin(2x) = -2\sin(4x)$.

4) $f(x) = \sqrt{x^2 + x - 1}$, by using the chain rule then $f'(x) = \frac{2x+1}{2\sqrt{x^2+x-1}}$.

5) $f(x) = (x^3 + 3)^4$, by using the chain rule then $f'(x) = 3x^2.4(x^3 + 3)^3 = 12x^2(x^3 + 3)^3$

Extreme Values

Definition 3. We say that $f(x_0)$ is a local extreme value of f if

$$f(x) \leq f(x_0)$$

or a local minimum value of f if

$$f(x) \geq f(x_0)$$

The point x_0 is called a local extreme point of f , or, more specifically, a local maximum or local minimum point of f (see Figure 1.5).

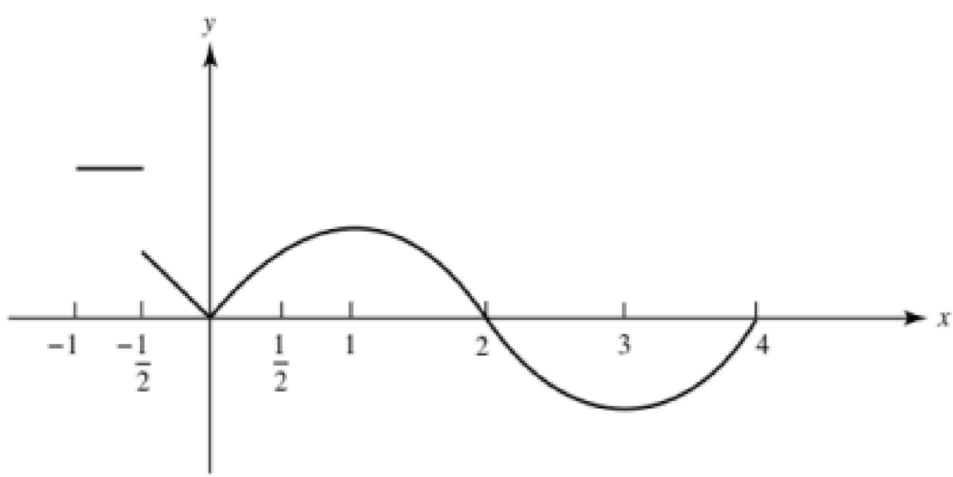


Figure 1.5: Local extreme value of f

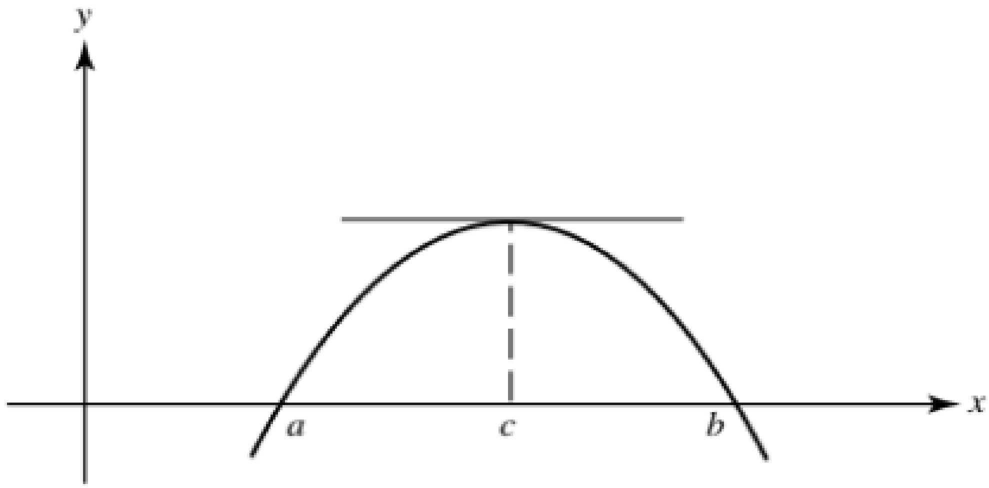
Theorem 4. If f is derivable at a local extreme point $x_0 \in D_f$; then $f'(x_0) = 0$.

If $f'(x_0) = 0$, we say that x_0 is a critical point of f . Theorem 4. says that every local extreme point of f at which f is derivable is a critical point of f . The converse is false. For example, 0 is a critical point of $f(x) = x^3$, but not a local extreme point.

Example 7. Let $f(x) = x^2$, f is derivable on R and we have $f'(x) = 2x$ and also $x_0 = 0$ is a local extreme of f $(x^2 \geq 0)$, then $f'(0) = 0$.

Rolle's Theorem

Theorem 5. Suppose that f is continuous on the closed interval $[a,b]$ and derivable on the open interval $]a,b[$; and $f(a)=f(b)$ Then $f'(c)=0$ for some c in the open interval $]a,b[$.



Example 8. Let $f(x)=\ln(x^2+1)$, f is defined on R. We have $f(2)=f(-2)=\ln 5$ and because f is continuous and derivable on $[-2,2]$, then $f'(c)=f'(0)=0$ $\left(f'(x)=\frac{2x}{x^2+1}, c\in[-2,2]\right)$ (see Figure 1.6)

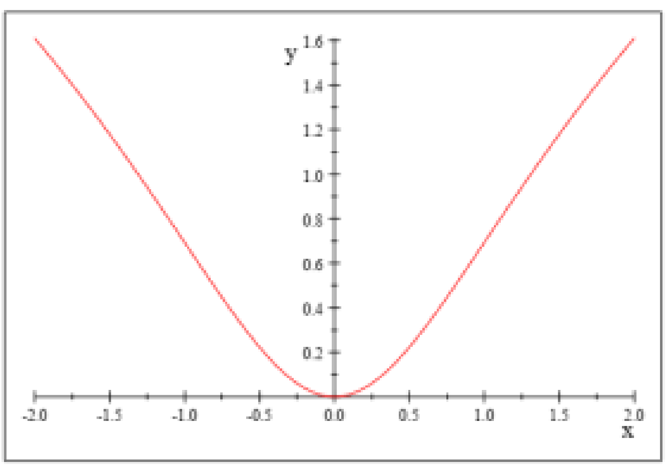


Figure 1.6: Plot of the function $\ln(x^2+1)$

Intermediate Values of Derivatives

A derivative may exist on an interval $[a,b]$ without being continuous on $[a,b]$.

Theorem 4. Suppose that f is differentiable on $[a, b]$; $f'(a) \neq f'(b)$; and β is between $f'(a)$ and $f'(b)$. Then $f'(c) = \beta$ for some c in $]a, b[$.

Example 8 Let the function $f(x) = x^3 - 2x^2 + 1$ and the interval $[1, 2]$.

Solution:

Let us find the values of the derivative function at the $x = 1$ and $x = 2$.

$$f'(x) = 3x^2 - 4x$$

Substitute $x = 1$ in the given function

$f'(1) = 3(1)^2 - 4(1) = -1$

Substitute $x = 2$ in the given function

$f'(2) = 3(2)^2 - 4(2) = 4$

Since the function is a polynomial, it is derivable on R. Thus, applying the intermediate value theorem (theorem 4), there exists a number c between the interval $[1, 2]$ and $\beta \in [-1, 4]$ where $f'(c) = \beta$. There is two solution of a equation $f'(x) = 3x^2 - 4x = 0$, we can choose $c = \frac{4}{3}$ and $\beta = 0$.

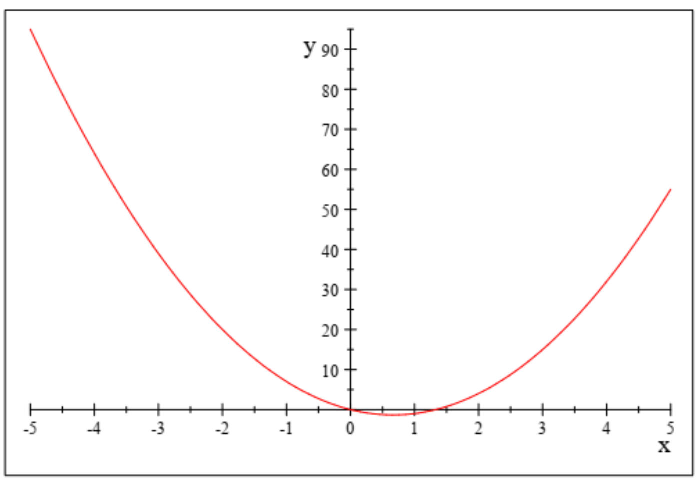


Figure 1.7: Plot of the function $3x^2 - 4x$

Mean Value Theorems

Theorem 5. If f and g are continuous on the closed interval $[a, b]$ and derivable on the open interval $]a, b[$ then

$$(g(b) - g(a)) f'(c) = (f(b) - f(a)) g'(c)$$

for some c in $]a, b[$.

Theorem 6. (**Mean Value Theorem**) If f and g are continuous on the closed interval $[a, b]$ and derivable on the open interval $]a, b[$ then

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

for some c in $]a, b[$.

Theorem 7. If $f'(x) = 0$ for all x in $]a, b[$, then f is constant on $]a, b[$.

1.5 Exercises with solutions

Exercise 1

Let

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x} + 3, x \neq 0 \\ 1, x = 0 \end{cases}$$

1) Is f continuous at $x = 0$?

2) Is f continuous on R?

Solution. 1) a) The function is defined at $x = 0$ and its value is $f(0) = 1$.

b) Now we use the squeeze theorem to find the value of the limit.

Since $-1 \leq \sin \frac{1}{x} \leq 1$ for all values of x, we can multiply by x^2 to get $-x^2 \leq x^2 \sin \frac{1}{x} \leq x^2$ for all values of x. Since $\lim\limits_{x \longrightarrow 0} x^2 = 0$, we conclude that the function between them also approaches

zero. Therefore $\lim\limits_{x \longrightarrow 0} x^2 \sin \frac{1}{x} = 0$, which implies $\lim\limits_{x \longrightarrow 0} f(x) = \lim\limits_{x \longrightarrow 0} x^2 \sin \frac{1}{x} + 3 = 0$.

Since the value of limit does **not** equal the value of the function, $f(x)$ is **not** continuous at $x = 0$.

2) We have $x^2 \sin \frac{1}{x} + 3$ is continuous on R^* and f is **not** continuous at $x = 0$, thus f is continuous on R^*.

Exercise 2 Show that $f(x)$ has a continuous extension to $x = 2$ and find that extension

$$f(x) = \frac{x^2 + x - 6}{x - 2}$$

Solution. Although $f(2)$ is not defined,

If $x \neq 2$ we have

$$f(x) = \frac{x^2 + x - 6}{x - 2} = \frac{(x-2)(x+3)}{x-2} = x + 3$$

The function $g(x) = x + 3$ is equivalent to $f(x) = \frac{x^2+x-6}{x-2}$, but $g(x)$ is continuous at $x = 2$ having a $\lim\limits_{x \longrightarrow 2} g(x) = 5$ and $g(2) = 5$. Thus the extension of the function f

$$\widetilde{f}(x) = \begin{cases} \frac{x^2+x-6}{x-2}, & \text{if } x \neq 2 \\ 5, & \text{if } x = 2 \end{cases}$$

Exercise 3 Let

$$f(x) = \begin{cases} \frac{x^2-9}{x-3}, & \text{if } x < 3 \\ \alpha x^2 + 10, & \text{if } x \geq 3 \end{cases}$$

Find the value of α so that $f(x)$ is continuous at $x = 3$.

Solution. 1) The function is defined at $x = 3$ and its value is $f(3) = \alpha(3)^2 + 10 = 9\alpha + 10$.

2) $\lim\limits_{\substack{< \\ x \longrightarrow 3}} f(x) = \lim\limits_{\substack{< \\ x \longrightarrow 3}} \frac{x^2-9}{x-3} = \lim\limits_{\substack{< \\ x \longrightarrow 3}} \frac{(x-3)(x+3)}{x-3} = 6.$

3) Since $y = cx^2 + 10$ is continuous at $x = 3$, we have:

$$\lim_{\substack{> \\ x \longrightarrow 3}} f(x) = \lim_{x \longrightarrow 3} \left(\alpha x^2 + 10\right) = 9\alpha + 10$$

In order to make all three of these the same, we need $9\alpha + 10 = 6$. Thus, $\alpha = -\frac{4}{9}$.

Exercise 4 Find the derivatives of the following functions

1. $f(x) = \left(x + \frac{1}{x}\right)^2, x \in R, x \neq 0$
2. $f(x) = (\cos 2x)^2, x \in R.$

3. $f(x) = \ln \frac{1+e^x}{e^x}, x \in R$

4. $f(x) = \sqrt{x^2 - x + 1}, x \in R$

5. $f(x) = \sqrt{\frac{\cos 2x}{\sin x}}, x \in R, x \neq (2k+1)\pi$

Solution.

1) $f(x) = x^2 + 2 + \frac{1}{x^2} \Longrightarrow f'(x) = 2x - \frac{1}{x}$.

2) $f'(x) = 2(\cos 2x).(-2\sin 2x) = -4\cos 2x \sin 2x$.

3) $f'(x) = \frac{\frac{e^{2x} - e^x(1+e^x)}{e^{2x}}}{\frac{1+e^x}{e^x}} = \frac{-1}{1+e^x}$

4) $f'(x) = \dfrac{2x-1}{2\sqrt{x^2-x+1}}$.

5) $f'(x) = \dfrac{1}{2\sqrt{\frac{\cos 2x}{\sin x}}} \dfrac{-2\sin 2x \sin x - \cos 2x \cos x}{(\sin x)^2}$.

Exercice 5 Consider the function

$$f(x) = \begin{cases} e^x, & \text{if } x \leq 0 \\ \cos x + x, & \text{if } x > 0 \end{cases}$$

1) Is f derivable at $x = 0$?

2) Is f derivable on R?

3) Calculate $f'(x)$ on its domain.

4) Show that f is (once) continuously differentiable (f' is continuous) at $x = 0$.

5) Is f twice continuously differentiable (f'' is continuous) at $x = 0$?

Solution. 1) We have

$$\lim_{x \longrightarrow 0^+} \frac{f(x) - f(0)}{x - 0} = \lim_{x \longrightarrow 0^+} \frac{\cos x + x - e^0}{x - 0} = +\infty$$
$$\lim_{x \longrightarrow 0^-} \frac{f(x) - f(0)}{x - 0} = \lim_{x \longrightarrow 0^-} \frac{e^x - e^0}{x - 0} = 1$$

Hence f is not derivable at $x = 0$.

2) Since e^x is derivable on $]-\infty, 0]$ and $\cos x + x$ is derivable on $]0, +\infty[$, thus f is derivable on R^*.

3) Since f is derivable on R^*,we have

$$f'(x) = \begin{cases} e^x, & \text{if } x \leq 0 \\ 1 - \sin x, & \text{if } x > 0 \end{cases}$$

4) We have

$$\lim_{x \longrightarrow 0^+} f'(x) = \lim_{x \longrightarrow 0^-} f'(x) = 1$$

Thus f' is continuous at $x = 0$ or f is (once) continuously differentiable at $x = 0$.

5) For $x \neq 0$

$$f''(x) = \begin{cases} e^x, & \text{if } x \leq 0 \\ -\cos x, & \text{if } x > 0 \end{cases}$$

Since

$$\lim_{x \longrightarrow 0^+} f''(x) = -1 \neq \lim_{x \longrightarrow 0^-} f''(x) = 1$$

Thus f'' is not continuous at $x = 0$ or f is not (twice) continuously differentiable at $x = 0$.

Exercise 6 Let

$$f(x) = \begin{cases} ax + b, & \text{if } x < 0 \\ 2\sin x + 3\cos x, & \text{if } x \geq 0 \end{cases}$$

Find the values of a and b so that f is derivable at $x = 0$.

Solution. First of all, $f(x)$ must be continuous (f derivable implies f continuous) at $x = 0$.

Hence $\lim_{x \longrightarrow 0^-} f(x) = f(0) \Longrightarrow b = 2\sin 0 + 3\cos 0 = 3$

Second, find $f'(x)$:

$$f'(x) = \begin{cases} a, & \text{if } x < 0 \\ 2\cos x - 3\sin x, & \text{if } x \geq 0 \end{cases}$$

Since $f(x)$ is differentiable at $x = 0$. $\lim_{x \longrightarrow 0^-} f'(x) = f'(0) \Longrightarrow a = 2\cos 0 - 3\sin 0 = 2.$

Therefore $a = 2, b = 3$.

Chapter 2

INTEGRATION

2.1 Primitive functions

Definition 1 (**Primitive function**). If $I \subseteq R$ is a non-empty open interval and

$$F, f : I \longrightarrow R$$

are functions satisfying $F' = f$ on I, we call F a primitive function of f on I. The notation we use to denote this relationship is either

$$\int f(x)\,dx = F(x)$$

If F is a primitive of f on $[a, b]$ and f is integrable on $[a, b]$, then by fundamental theorem of calculus we have

$$\int_a^b f(x)\,dx = [F(x)]_a^b = F(b) - F(a)$$

Theorem 1 (Set of primitive functions). Let F be a primitive function of f on I. Then the set of all primitive functions of f on I is

$$F + c, c \in R$$

e.g

$$\int f(x)\,dx = F(x) + c$$

Therefore, all primitive functions of f are obtained by shifting any primitive function of f by a constant.

2.1.1 Properties of primitive functions

1) Linearity : If F is a primitive function of f, and G is a primitive function of g on an interval I and $\alpha, \beta \in R$, then the function

$$\alpha F + \beta G$$

is a primitive function of $\alpha f + \beta g$ on I.

For limit and derivative, the result of the operation is unique if it exists, but a primitive function is not unique. We will soon see that the function either does not have any primitive function or has infinitely many.

2) Monotonicity: If $f(x) \geq g(x)$ for all $x \in [a, b]$ then

$$\int_a^b f(x)\,dx \geq \int_a^b g(x)\,dx$$

3) If $f(x)$ is of bounded variation on $[a; b]$, then so is $|f(x)|$ and

$$\left| \int_a^b f(x)\,dx \right| \leq \int_a^b |f(x)|\,dx$$

4) If $a < c < b$ then

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx$$

5) If $m \leq \inf\{f(x)\}$ and $M \geq \sup\{f(x)\}$ where $x \in [a, b]$ then

$$(b - a)\,m \leq \int_a^b f(x)\,dx \leq (b - a)\,M$$

6) Mean value theorem for integrals: If f is continuous and of bounded variation on $[a;b]$ then there is a c with $a < c < b$ such that

$$f(c) = \frac{1}{b-a}\int_a^b f(x)\,dx$$

2.1.2 Primitive function and continuity

Theorem 3 (**Continuity of a primitive function**). If F is a primitive function of f on I, then F is continuous on I.

Proof. We know from the winter term that the existence of the proper derivative of a function at a point implies its continuity at the given point. Since $F'(\alpha)$ exists and is equal to $f(\alpha)$ for each $\alpha \in I$, F is continuous on I.

Theorem 4 (Continuous function has a primitive function). If f is continuous on I, then f has a primitive function F on I.

Can a discontinuous function have a primitive function? Yes.

Example 1 (Discontinuous function with primitive function.). The function
$f : R \longrightarrow R$ defined as

$$f(x) = 2\sin\frac{1}{x^2} - 2\cos\frac{1}{x^2} \text{ for } x \neq 0, f(0) = 0$$

has a primitive function on R, even though it is not continuous at 0.

Solution. Consider $F : R \longrightarrow R$ defined for $x \neq 0$ as $F(x) = x^2 \sin\frac{1}{x^2}$ and for $x = 0$ as $F(0) = 0$. For $x \neq 0$ we have $F' = f$ by standard calculations. At zero by the definition of the derivative we calculate that

$$F'(0) = \lim_{x \longrightarrow 0}\frac{F(x) - F(0)}{x - 0} = \lim_{x \longrightarrow 0} x^2 \sin\frac{1}{x^2} = 0$$

because $\left|x^2 \sin\frac{1}{x^2}\right| < |x|$ for every $x \neq 0$. Thus $F'(0)$ exists and again $F'(0) = f(0)$. Therefore, $F' = f$ on R and F is a primitive function of f on R. Function f is not continuous at 0, in every neighborhood of zero it is even unbounded from both above and below, for $x \longrightarrow 0$ the graph oscilates with increasing amplitude and frequency.

2.1.3 Primitive of elementary and composite functions

By reversing the direction of formulas for derivatives of elementary functions we get the following table of primitive functions.

function	primitive function	on interval
$x^\alpha, \alpha \in R - \{-1\}$	$\frac{1}{\alpha+1}x^{\alpha+1}$	$(0, +\infty)$
$\frac{1}{x} = x^{-1}$	$\log\lvert x\rvert$	$(0, +\infty)$ and $(-\infty, 0)$
e^x	e^x	R
$\sin x$	$-\cos x$	R
$\cos x$	$\sin x$	R
$\frac{1}{\cos^2 x}$	$\tan x$	$\left(\left(k-\frac{1}{2}\right)\pi, \left(k+\frac{1}{2}\right)\pi\right), k \in Z$
$\frac{1}{\sin^2 x}$	$-\cot x$	$(k\pi, (k+1)\pi), k \in Z$
$\frac{1}{1+x^2}$	$\arctan x$	R
$\frac{1}{\sqrt{1-x^2}}$	$\arcsin x$	$(-1, 1)$

Examples:

1) $\int (x^3 + 3x^2 - 10)\, dx = \frac{1}{4}x^4 + x^3 - 10x + c$

2) $\int_2^4 \sqrt{x}dx = \int_2^4 x^{\frac{1}{2}}dx = \left[\frac{2}{3}x^{\frac{3}{2}}\right]_2^4 = \left[\frac{2}{3}x\sqrt{x}\right]_2^4 = \frac{16}{3} - \frac{4}{3}\sqrt{2}$

3) $\int \left(\frac{1}{x^3} + \cos x\right) dx = \int (x^{-3} + \cos x)\, dx = -\frac{1}{2}x^{-2} + \sin x + c = -\frac{1}{2x^2} + \sin x + c$

Primitive of composite functions

function	primitive function
$\frac{f'}{f}\,(f \neq 0)$	$\ln\lvert f\rvert$
$\frac{f'}{2\sqrt{f}}\,(f \neq 0)$	$\sqrt{f}$
$f'\exp(f)$	$\exp(f)$
$2f' \times f$	f^2
$nf'f^{n-1}$	f^n
$\frac{f'}{1+f^2}$	$\arctan f$
$\exp(ax+b)$	$\frac{1}{a}\exp(ax+b)\,(a \neq 0)$
$\cos(ax+b)$	$\frac{1}{a}\sin(ax+b)\,(a \neq 0)$
$\sin(ax+b)$	$-\frac{1}{a}\cos(ax+b)\,(a \neq 0)$

Examples 1) $\int \frac{2x+1}{x^2+x+2}dx = \ln\lvert x^2+x+2\rvert + c$

2) $\int \frac{\ln x}{x}dx = \int \ln x d(\ln x)\,dx = \frac{1}{2}\ln^2 x + c$

3)$\int \frac{2x}{\sqrt{x^2+7}}dx = 2\int \frac{2x}{2\sqrt{x^2+7}}dx = 2\sqrt{x^2+7} + c$

4)$\int_0^1 \frac{x}{x+1}dx = \int_0^1 \frac{x+1-1}{x+1}dx = \int_0^1 \left(1+\frac{1}{x+1}\right)dx = [x+\ln\lvert x+1\rvert]_0^1 = \ln 2.$

2.2 Methods for computing primitive functions

To calculate the derivative of the product of two functions, we have the Leibniz formula $(fg)' = f'g + fg'$. By inverting it, we will get the following important result for primitive functions.

2.2.1 Integration by parts

Suppose f', g' are both continuous (so all integrals below exist) we have

$$(fg)' = f'g + fg'$$

or

$$fg' = (fg)' - f'g$$

Hence

$$\int fg' = fg - \int f'g$$

or

$$\int f(x)\, g'(x)\, dx = f(x)\, g(x) - \int f'(x)\, g(x)\, dx$$

From fundamental theorem of calculus we get that, as long as $[a, b]$ is fully contained in the domains of both f and g,

$$\int_a^b f(x)\, g'(x)\, dx = [f(x)\, g(x)]_a^b - \int_a^b f'(x)\, g(x)\, dx$$

Examples 1) Compute the integral $\int xe^x dx$

We have $\begin{cases} f = x \longrightarrow f' = 1 \\ g' = e^x \longrightarrow g = e^x \end{cases}$ then

$$\begin{aligned} \int xe^x dx &= xe^x - \int e^x dx \\ &= xe^x - e^x + c \end{aligned}$$

2) With $(x)' = 1$ and $(\ln x)' = \frac{1}{x}$ on the interval $]0, +\infty[$ we have

$$\int \ln x dx = \int x' \ln x dx = x \ln x - \int x\, (\ln x)'\, dx = x \ln x - \int 1 dx = x\, (\ln x - 1) + c$$

By taking derivative, we can easily check the correctness of the derived formula.

3) Compute $\int_0^\pi x^2 \sin x dx$.

We have $\begin{cases} f = x^2 \longrightarrow f' = 2x \\ g' = \sin x \longrightarrow g = -\cos x \end{cases}$ then

$$\begin{aligned}\int_0^\pi x^2 \sin x dx &= \left[-x^2 \cos x\right]_0^\pi + \int_0^\pi 2x \cos x dx \\ &= \pi^2 + \int_0^\pi 2x \cos x dx\end{aligned}$$

we have $\begin{cases} f = 2x \longrightarrow f' = 2 \\ g' = \cos x \longrightarrow g = \sin x \end{cases}$ then

$$\begin{aligned}\int_0^\pi 2x \cos x dx &= [2x \sin x]_0^\pi - \int_0^\pi 2 \sin x dx \\ &= [2x \sin x]_0^\pi + [2 \cos x]_0^\pi \\ &= -4\end{aligned}$$

Finally we obtain

$$\int_0^\pi x^2 \sin x dx = \pi^2 - 4$$

2.2.2 Integration by substitution

The chain rule leads to an integration principle, integration by substitution.

If f', g' are both continuous (so all integrals below exist) and if F is a primitive of f, then (since $(F \circ g)'(x) = F'(g(x))\, g'(x) = f(g(x))\, g'(x)$) we have

$$\int f(g(x))\, g'(x)\, dx = (F \circ g)(x) = F(g(x))$$

Examples 1) Calculate the integral $\int \frac{(\ln x)^2}{x} dx$

We pose $t = \ln x$ or $x = e^t$, then $dt = \frac{1}{x} dx$ or $dx = e^t dt$

Hence

$$\int \frac{(\ln x)^2}{x} dx = \int t^2 dt = \frac{1}{3} t^3 = \frac{1}{3} (\ln x)^3 + c$$

2) Compute $\int e^{\sqrt{x}}dx$

We pose $t=\sqrt{x}$, then $dt=\frac{1}{2\sqrt{x}}dx$ or $dx=2tdt$

Hence

$$\int e^{\sqrt{x}}dx=\int 2te^{t}dt$$

By using integration by parts we have $\begin{cases} f=2t \longrightarrow f'=2 \\ g'=e^{t} \longrightarrow g=e^{t} \end{cases}$ then

$$\begin{aligned}\int 2te^{t}dt &= 2te^{t}-\int 2e^{t}dt \\ &= 2te^{t}-2e^{t}\end{aligned}$$

Finally we obtain

$$\int e^{\sqrt{x}}dx=2\sqrt{x}e^{\sqrt{x}}-2e^{\sqrt{x}}+c$$

3) Compute $\int \cos^{3}x\sin^{4}xdx$

We pose $t=\sin x$, then $dt=\cos xdx$

Hence

$$\int \cos^{3}x\sin^{4}xdx=\int \cos^{2}x\sin^{4}x\cos xdx=\int \left(1-t^{2}\right)t^{4}dt=\frac{1}{5}t^{5}-\frac{1}{7}t^{7}=\frac{1}{5}\sin^{5}x-\frac{1}{7}\sin^{7}x+c$$

2.2.3 Integration by partial fractions

A relatively wide class of functions to which primitive functions can be computed are rational functions, which are fractions of polynomials.

Theorem 9. Let $P(x)$ and $Q(x)\neq 0$ be polynomials with real coefficients and $I\subset R$ is an open interval not containing no roots of $Q(x)$. Primitive function

$$F\left(x\right)=\int \frac{P\left(x\right)}{Q\left(x\right)}dx \quad (\text{on } I\ )$$

can be expressed using elementary functions, namely using rational functions, logarithms and arcustangent.

Case 1. We have $\frac{P(x)}{Q(x)} = \frac{ax+b}{(x+c)(c+d)} = \frac{A}{x+c} + \frac{B}{x+d}$ then

$$\int \frac{ax+b}{(x+c)(c+d)}dx = \int \left(\frac{A}{x+c} + \frac{B}{x+d}\right) dx = \ln|x+c| + \ln|x+d| + c$$

Example 9. Compute $\int \frac{2x+1}{(x-1)(x+3)}dx$

We have

$$\frac{2x+1}{(x-1)(x+3)} = \frac{A}{x-1} + \frac{B}{x+3} = \frac{(A+B)x - B + 3A}{(x-1)(x+3)}$$

Hence

$$\begin{cases} A+B=2 \\ -B+3A=1 \end{cases} \Longrightarrow \begin{cases} A=\frac{3}{4} \\ B=\frac{5}{4} \end{cases}$$

This yields

$$\int \frac{2x+1}{(x-1)(x+3)}dx = \int \left(\frac{3/4}{x-1} + \frac{5/4}{x+3}\right) dx = \frac{3}{4}\ln|x-1| + \frac{5}{4}\ln|x+3| + C$$

Case 2. We have $\frac{P(x)}{Q(x)} = \frac{ax+b}{x^2+cx+d}$ we know that

$$\begin{cases} \Delta = c^2-4d=0, x^2+cx+d = \left(x+\frac{c}{2}\right)^2 \\ \Delta = c^2-4d<0, \text{ not exist factoristion} \\ \Delta = c^2-4d>0, x^2+cx+d = (x-x_1)(x-x_2) \end{cases}$$

where $x_1 = \frac{-c-\sqrt{\Delta}}{2}, x_2 = \frac{-c+\sqrt{\Delta}}{2}$

Examples. 1) $\int \frac{2}{x^2-2x+1}dx = \int \frac{2}{(x-1)^2}dx = \int 2(x-1)^{-2}dx = 2\frac{(x-1)^{-1}}{-1} + C = -\frac{2}{(x-1)} + C$

2) $\int \frac{x+1}{x^2-3x+2}dx = \int \frac{x+1}{(x-2)(x-1)}dx = \int \left(-\frac{2}{x-1} + \frac{3}{x-2}\right)dx = -2\ln|x-1| + 3\ln|x-2| + C$ $(\Delta = 1, x_1 = 1, x_2 = 2)$

where

$$\frac{x+1}{(x-2)(x-1)} = \frac{A}{x-1} + \frac{B}{x-2}$$

After calculation we obtain

$$\begin{cases} A = -2 \\ B = 3 \end{cases}$$

3) $\int \frac{1}{x^2-2x+5}dx = \int \frac{1}{(x-1)^2+4}dx = \int \frac{1}{4\left((2(x-1))^2+1\right)}dx$ $(\Delta = -1)$

We pose $t = 2x - 2 \longrightarrow dt = 2dx$ then

$$\int \frac{1}{4\left((2(x-1))^2+1\right)}dx = \frac{1}{8}\int \frac{dt}{t^2+1} = \frac{1}{8}\arctan t = \frac{1}{8}\arctan(2x-2) + C$$

or

$$\int \frac{1}{x^2-2x+5}dx = \frac{1}{8}\arctan(2x-2) + C$$

Case 3. We have $\frac{P(x)}{Q(x)} = \frac{1}{e^{2x}+ce^x+d}$ we can write this equation by using the change of variable $t = e^x$ on the form

$$\frac{1}{e^{2x}+ce^x+d} = \frac{1}{t^2+ct+d}$$

Examples. 1) $\int \frac{2}{e^x+1}dx = \int \frac{2}{t(t+1)}dt$ where $t = e^x$ then $dt = tdx$

We have

$$\frac{2}{t(t+1)} = \frac{A}{t} + \frac{B}{t+1} \Longrightarrow \begin{cases} A = 2 \\ B = -2 \end{cases}$$

$$\int \frac{2}{e^x+1}dx = \int \frac{2}{t(t+1)}dt = \int \left(\frac{2}{t} - \frac{2}{t+1}\right)dt = 2\ln|t| - 2\ln|t+1| = 2x - 2\ln|e^x+1| + K$$

2) $\int \frac{3}{e^{2x}-4e^x+3}dx = \int \frac{3}{t(t-1)(t-3)}dt$ where $t = e^x$ then $dt = tdx$

We have

$$\frac{3}{t(t-1)(t-3)} = \frac{A}{t} + \frac{B}{t-1} + \frac{C}{t-3} = \frac{(A+B+C)t^2 + (-4A-3B-C)t + 3A}{t(t-1)(t-3)}$$

Hence

$$\begin{cases} A+B+C=0 \\ -4A-3B-C=0 \\ 3A=1 \end{cases} \Longrightarrow \begin{cases} A=1 \\ B=-\frac{3}{2} \\ C=\frac{1}{2} \end{cases}$$

This yields

$$\int \frac{3}{t\left(t-1\right)\left(t-3\right)}dt = \int \left(\frac{1}{t} - \frac{3/2}{t-1} + \frac{1/2}{t-3}\right) dt = \ln|t| - \frac{3}{2}\ln|t-1| + \frac{1}{2}\ln|t-3|$$

or

$$\int \frac{3}{e^{2x}-4e^x+3}dx = \ln|e^x| - \frac{3}{2}\ln|e^x-1| + \frac{1}{2}\ln|e^x-3| + K = x - \frac{3}{2}\ln|e^x-1| + \frac{1}{2}\ln|e^x-3| + K$$

3) $\int \frac{1}{e^{2x}+4e^x+4}dx = \int \frac{1}{(e^x+2)^2}dx = \int \frac{dt}{t(t+2)^2}$ where $t=e^x$ then $dt = tdx$

we have

$$\begin{aligned} \frac{1}{t\left(t+2\right)^2} &= \frac{A}{t} + \frac{Bt+C}{\left(t+2\right)^2} = \frac{At^2+4At+4A+Bt^2+Ct}{(t^2+2t)} \\ &= \frac{\left(A+B\right)t^2 + \left(4A+C\right)t + 4A}{t\left(t+2\right)^2} \end{aligned}$$

Hence

$$\begin{cases} A+B=0 \\ 4A+C=0 \\ 4A=1 \end{cases} \Longrightarrow \begin{cases} A=\frac{1}{4} \\ B=-\frac{1}{4} \\ C=-1 \end{cases}$$

We get

$$\begin{aligned}\int \frac{1}{e^{2x}+4e^x+4}dx &= \int \frac{1}{(e^x+2)^2}dx = \int \frac{dt}{t\,(t+2)^2} \\ &= \int \left(\frac{A}{t}+\frac{Bt+C}{(t+2)^2}\right)dt \\ &= \int \left(\frac{\frac{1}{4}}{t}+\frac{\frac{-1}{4}t-1}{(t+2)^2}\right)dt \\ &= \int \left(\frac{1}{4t}-\frac{1}{8}\frac{2t+8}{(t+2)^2}\right)dt \\ &= \int \left(\frac{1}{4t}-\frac{1}{8}\left(\frac{2t+4}{(t+2)^2}+\frac{4}{(t+2)^2}\right)\right)dt \\ &= \frac{1}{4}\ln t-\frac{1}{8}\ln (t+2)^2+\frac{1}{2}\frac{1}{t+2} \\ &= \frac{1}{4}x-\frac{1}{4}\ln (e^x+2)+\frac{1}{2}\frac{1}{e^x+2}+c\end{aligned}$$

2.3 Exercises with solutions

Exercise1 Calculate the following integrals

$$1) \int \ln\left(1+\sqrt{x}\right)dx, 2) \int_1^2 \frac{1}{1+\sqrt{x+1}}dx, 3) \int \sqrt{1-x^2}dx, 4) \int_1^\pi \cos(\ln x)\,dx, 5) \int \cos\left(\sqrt{x}\right)dx$$

Solution 1) $\int \ln(1+\sqrt{x})\,dx$, An abvious substitution is $t=\sqrt{x}$, with $dt=\frac{1}{2\sqrt{x}}dx$ or $dx=2tdt$ leading to

$$\int \ln\left(1+\sqrt{x}\right)dx = \int 2t\ln(1+t)\,dt$$

By using integration by parts $\begin{cases} u = \ln(1+t) \longrightarrow u' = \dfrac{1}{1+t} \\ v' = 2t \longrightarrow v = t^2 \end{cases}$ we get

$$\begin{aligned}
\int \ln\left(1+\sqrt{x}\right) dx = \int 2t \ln(1+t)\, dt &= t^2 \ln(1+t) - \int \frac{t^2}{1+t} dt \\
&= t^2 \ln(1+t) - \int \frac{t^2 - 1 + 1}{1+t} dt \\
&= t^2 \ln(1+t) - \int \left(t - 1 + \frac{1}{1+t}\right) dt \\
&= t^2 \ln(1+t) - \frac{t^2}{2} + t - \ln(1+t) \\
&= (x-1)\ln\left(1+\sqrt{x}\right) - \frac{x}{2} + \sqrt{x} + c
\end{aligned}$$

2) $\int_1^2 \frac{1}{1+\sqrt{x+1}}$, Here An abvious substitution is $t = \sqrt{1+x}$, which gives $dt = \frac{1}{2\sqrt{1+x}} dx$ or $dx = 2tdt$ leading to

$$\begin{aligned}
\int \frac{1}{1+\sqrt{x+1}} &= \int \frac{2tdt}{1+t} \\
&= 2\int \frac{t+1-1}{1+t} dt \\
&= 2\int \left(1 - \frac{1}{1+t}\right) dt \\
&= 2\left(t - \ln(1+t)\right) \\
&= 2\left(\sqrt{1+x} - \ln\left(1+\sqrt{1+x}\right)\right)
\end{aligned}$$

So

$$\begin{aligned}
\int_1^2 \frac{1}{1+\sqrt{x+1}} &= \left[2\left(\sqrt{1+x} - \ln\left(1+\sqrt{1+x}\right)\right)\right]_1^2 \\
&= 2\sqrt{3} - 2\ln\left(1+\sqrt{3}\right) - 2\sqrt{2} + 2\ln\left(1+\sqrt{2}\right) \\
&= 0.388
\end{aligned}$$

3) $\int \sqrt{1-x^2}dx$, we pose $x=\sin t$, then $dx=\cos t dt$, we get

$$\begin{aligned}\int \sqrt{1-x^2}dx &= \int \cos^2 t dt \\ &= \int \frac{1+\cos 2t}{2}dt \\ &= \frac{1}{2}t+\frac{1}{4}\sin 2t \\ &= \frac{1}{2}t+\frac{1}{2}\cos t \sin t \\ &= \frac{1}{2}\arcsin x+\frac{1}{2}x\sqrt{1-x^2}+c\end{aligned}$$

We know that $\cos^2 x+\sin^2 x=1$ and $\cos 2x=\cos^2 x-\sin^2 x, \sin 2x = 2\cos x \sin x$.

Or we can use integration by parts $\begin{cases} u=\cos t \longrightarrow u'=-\sin t \\ v'=(\sin t)' \longrightarrow v=\sin t \end{cases}$

$$\begin{aligned}\int \cos^2 t dt &= \int \cos t\,(\sin t)'\,dt \\ &= \cos t \sin t + \int \sin^2 t dt \\ &= \frac{1}{2}\sin 2t + \int \left(1-\cos^2 t\right) dt \\ &= \frac{1}{2}\sin 2t + t - \int \cos^2 t dt\end{aligned}$$

Hence

$$\int \cos^2 t dt = \frac{1}{4}\sin 2t + \frac{1}{2}t$$

4) $\int_1^\pi \cos(\ln x)\,dx$, An abvious substitution is $t=\ln x$, with $dt=\frac{1}{x}dx$ or $dx=e^t dt$, we get

$$\int \cos(\ln x)\,dx = \int e^t \cos t dt$$

By using integration by parts $\begin{cases} u=e^t \longrightarrow u'=e^t \\ v'=\cos t \longrightarrow v=\sin t \end{cases}$ we get

$$\int e^t \cos t dt = e^t \sin t - \int e^t \sin t dt$$

By using integration by parts again $\begin{cases} u = e^t \longrightarrow u' = e^t \\ v' = \sin t \longrightarrow v = -\cot t \end{cases}$ we get

$$\begin{aligned} \int e^t \cos t dt &= e^t \sin t - \int e^t \sin t dt \\ &= e^t \sin t + e^t \cos t - \int e^t \cot t dt \end{aligned}$$

This yields

$$\int e^t \cot t dt = \frac{e^t \sin t + e^t \cos t}{2}$$

Or

$$\begin{aligned} \int_1^\pi \cos(\ln x)\, dx &= \left[\frac{e^{\ln x} (\sin(\ln x) + \cos(\ln x))}{2} \right]_1^\pi \\ &= \frac{e^{\ln \pi} (\sin(\ln \pi) + \cos(\ln \pi))}{2} - \frac{e^{\ln 1} (\sin(\ln 1) + \cos(\ln 1))}{2} \\ &\approx 1.58 \end{aligned}$$

5) $\int \cos(\sqrt{x})\, dx$, we pose $t = \sqrt{x}$, with $dt = \frac{1}{2\sqrt{x}} dx$ or $dx = 2tdt$ leading to

$$\int \cos(\sqrt{x})\, dx = \int 2t \cos t dt$$

By using integration by parts $\begin{cases} u = 2t \longrightarrow u' = 2 \\ v' = \cos t \longrightarrow v = \sin t \end{cases}$ we get

$$\begin{aligned} \int \cos(\sqrt{x})\, dx &= \int 2t \cos t dt \\ &= 2t \sin t - \int 2 \sin t dt \\ &= 2t \sin t + 2 \cos t \\ &= 2\sqrt{x} \sin \sqrt{x} + 2 \cos \sqrt{x} + c \end{aligned}$$

Exercise 2 Compute the following integrals

$$1) \int x^2 \cos x, \;\; 2) \int \frac{x+1}{x^2-4} dx, \;\; 3) \int \frac{e^x+1}{e^{2x}-5e^x+6} dx$$

Solution. 1) $\int x^2 \cos x$, By using integration by parts $\begin{cases} u = x^2 \longrightarrow u' = 2x \\ v' = \cos x \longrightarrow v = \sin x \end{cases}$ we get

$$\int x^2 \cos x = x^2 \sin x - \int 2x \sin x dx$$

Also we have $\begin{cases} u = 2x \longrightarrow u' = 2 \\ v' = \sin x \longrightarrow v = -\cos x \end{cases}$ thus

$$\begin{aligned} \int x^2 \cos x &= x^2 \sin x - \int 2x \sin x dx \\ &= x^2 \sin x + 2x \cos x + \int 2 \cos x dx \\ &= x^2 \sin x + 2x \cos x + 2 \sin x + c \end{aligned}$$

2) $\int \frac{x+1}{x^2-4} dx = \int \frac{x+1}{(x-2)(x+2)} dx = \int \left(\frac{A}{x-2} + \frac{B}{x+2} \right) dx$

$$\frac{A}{x-2} + \frac{B}{x+2} = \frac{(A+B)\,x + 2A - 2B}{x^2-4}$$

We get $A = \frac{3}{4}, B = \frac{1}{4}$ so

$$\begin{aligned} \int \frac{x+1}{x^2-4} dx &= \int \left(\frac{A}{x-2} + \frac{B}{x+2} \right) dx \\ &= \frac{3}{4} \ln|x-2| + \frac{1}{4} \ln|x+2| + c \end{aligned}$$

3) $\int \frac{e^x+1}{e^{2x}-5e^x+6}dx$, we pose $t = e^x$ with $dt = e^x dx$ or $dx = \dfrac{dt}{t}$. so

$$\begin{aligned}\int \frac{e^x+1}{e^{2x}-5e^x+6}dx &= \int \frac{t+1}{t^2-5t+6}\frac{dt}{t}\\ &= \int \frac{t+1}{(t-2)(t-3)}\frac{dt}{t}\\ &= \int \left(\frac{A}{t-2}+\frac{B}{t-3}+\frac{C}{t}\right)dt\end{aligned}$$

We have

$$\begin{aligned}\frac{A}{t-2}+\frac{B}{t-3}+\frac{C}{t} &= \frac{(A+B)\,t-3A-2B}{t^2-5t+6}+\frac{C}{t}\\ &= \frac{(A+B+C)\,t^2-(3A+2B+5C)\,t+6C}{(t^2-5t+6)\,t}\end{aligned}$$

So $\begin{cases} A+B+C=0 \\ -(3A+2B+5C)=1 \\ 6C=1 \end{cases} \Longrightarrow \begin{cases} A=-\frac{9}{6} \\ B=\frac{8}{6} \\ C=\frac{1}{6} \end{cases}$ then

$$\begin{aligned}\int \frac{e^x+1}{e^{2x}-5e^x+6}dx &= \int \left(\frac{A}{t-2}+\frac{B}{t-3}+\frac{C}{t}\right)dt\\ &= -\frac{9}{6}\ln|t-2|+\frac{8}{6}\ln|t-2|+\frac{1}{6}\ln|t|\\ &= -\frac{3}{2}\ln|e^x-2|+\frac{4}{3}\ln|e^x-2|+\frac{1}{6}x+c\end{aligned}$$

Exercise 3 Consider the integral

$$A_n = \int \frac{dx}{(x^2+1)^n}, n \geq 0$$

1) Find A_0 and A_1.

2) For $n \geq 2$, find the relationship between A_{n+1} and A_n.

3) Compute this integral $\int \frac{dx}{(x^2+1)^2}$

Solution 1) $A_0 = \int dx = x + c$

$A_1 = \int \frac{dx}{x^2+1} = \arctan x + c$

2) By using integration by parts, we have

$$\begin{cases} u = \frac{1}{(x^2+1)^n} \longrightarrow u' = \frac{-2nx}{(1+x^2)^{n+1}} \\ v' = 1 \longrightarrow v = x \end{cases}$$

We get

$$\begin{aligned} A_n &= \frac{x}{(x^2+1)^n} + \int \frac{2nx^2}{(1+x^2)^{n+1}} \\ &= \frac{x}{(x^2+1)^n} + 2n \int \frac{x^2+1-1}{(1+x^2)^{n+1}} \\ =& \frac{x}{(x^2+1)^n} + 2nA_n - 2nA_{n+1} \end{aligned}$$

So

$$A_{n+1} = \frac{x}{2n\,(x^2+1)^n} + \frac{2n-1}{2n} A_n$$

3) Last equation (valid for $n \geq 1$), For example $n = 1$ we get

$$\int \frac{1}{(1+x^2)^2} dx = \frac{x}{2\,(x^2+1)} + \frac{\arctan x}{2}$$

Chapter 3

DESCRIPTIVE STATISTICS

3.1 Introduction

Statistics is concerned with the scientific method by which information is collected, organised, analysed and interpreted for the purpose of description and decision making.

Examples using statistics are: Hang Seng Index, Life or car insurance rate, Unemployment rate, Consumer Price Index, etc.

There are two subdivisions of statistical method.

(a) Descriptive Statistics - It deals with the presentation of numerical facts, or data, in either tables or graphs form, and with the methodology of analysing the data.

(b) Inferential Statistics - It involves techniques for making inferences about the whole population on the basis of observations obtained from samples.

3.2 Some Basic Definitions

(a) **Population** - A population is the group from which data are to be collected.

(b) **Sample** - A sample is a subset of a population.

(c) **Variable** - A variable is a feature characteristic of any member of a population differing in quality or quantity from one member to another.

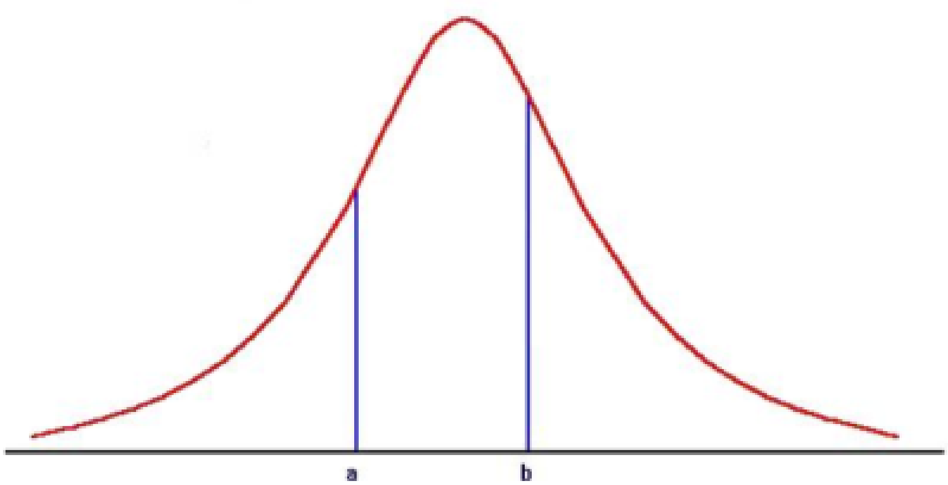
(d) **Frequency-** The frequency of a value is the number of observations taking that value, denoted by n_i.

(e) **Relative Frequency-** Relative frequency of a class f_i is defined as:

$$f_i = \frac{n_i}{N}$$

Where $N = \sum_{i=1}^{m} n_i$ total number of observations. We noted that $\sum_{i=1}^{m} f_i = 1$.

If the frequencies are changed to relative frequencies, then a relative frequency histogram, a relative frequency polygon and a relative frequency curve can similarly be constructed. Relative frequency curve can be considered as probability curve if the total area under the curve be set to 1. Hence the area under the relative frequency curve between a and b is the probability between interval a and b.



Cumulative Frequency (c.f)- In statistics, the frequency of the first-class interval is added to the frequency of the second class, and this sum is added to the third class and so on then, frequencies that are obtained this way are known as cumulative frequency **(c.f.)**. A table that displays the cumulative frequencies that are distributed over various classes is called a cumulative frequency distribution or cumulative frequency table. There are two types of cumulative frequency - lesser than type and greater than type. Cumulative frequency is used to know the number of observations that lie above (or below) a particular frequency in a given data set.

Lesser Than Cumulative Frequency (less than c.f)

Lesser than cumulative frequency is obtained by adding successively the frequencies of all the previous classes including the class against which it is written. The cumulate starts from the lowest to the highest size. In other words, when the number of observations is less than the upper boundary of a class that's when it is called lesser than cumulative frequency.

Greater Than Cumulative Frequency (more than c.f)

Greater than cumulative frequency is obtained by finding the cumulative total of frequencies starting from the highest to the lowest class. It is also called more than type cumulative frequency. In other words, when the number of observations is more than or equal to the lower boundary of the class that's when it is called greater than cumulative frequency.

Let us look at example below to understand the two types.

Variable Type

Before analyzing any data set, one should be familiar with different types of variables.

(d) **Quantitative variable** - A variable differing in quantity is called quantitative variable, for example, the weight of a person, number of people in a car, tall of people.

(e) **Qualitative variable** - A variable differing in quality is called a qualitative variable or attribute, for example, color, the degree of damage of a car in an accident.

(f) **Discrete variable** - A discrete variable is one which no value may be assumed between two given values, for example, number of children in a family.

(h) **Continuous variable** - A continuous variable is one which any value may be assumed between two given values, for example, the time for 100-meter run.

In the next section, we are concerned only with a quantity variable. We will study two types: descrete variable and continuous variable.

Modality- modality of data is the number of different types of data are included in the data set.

Classification of class intervals of grouped data. Class interval is a term that is used to denote the numerical width of a class in a frequency distribution. In a grouped frequency distribution, data is arranged in the form of a class. The difference between the upper-class limit and the lower limit gives the class interval.

C.I= $\dfrac{L-S}{R}$, $L =$*Largest value*

$S =$ *Smallest value*

$R =$The *number of classes.*

In statistics, there are two types of class intervals, namely exclusive and inclusive class intervals. Based on these, a frequency distribution table can be constructed.

We need the number of classes, it will use two methods.

a) *Yule's formula*: the number of classes is $R = 2.5\sqrt[4]{N}$.

b) *Sturges formula*: the number of classes $R = 1 + 3.3\log_{10} N$.

Where N is the total number of observations.

Class mid point or class marks

The mid value (middle point) or central value of the class interval is called mid point.

$$\textit{Mid point of a class } [a, b] = \frac{a+b}{2}$$

Constructing a frequency distribution

The following guidelines may be considered for the construction of frequency distribution.

a) The classes should be clearly defined and each observation must belong to one and to only one class interval. Interval classes must be inclusive and nonoverlapping.

b) The number of classes should be neither too large nor too small. Too small classes result greater interval width with loss of accuracy. Too many class interval result is complexity.

c) All intervals should be of the same width. This is preferred for easy computations.

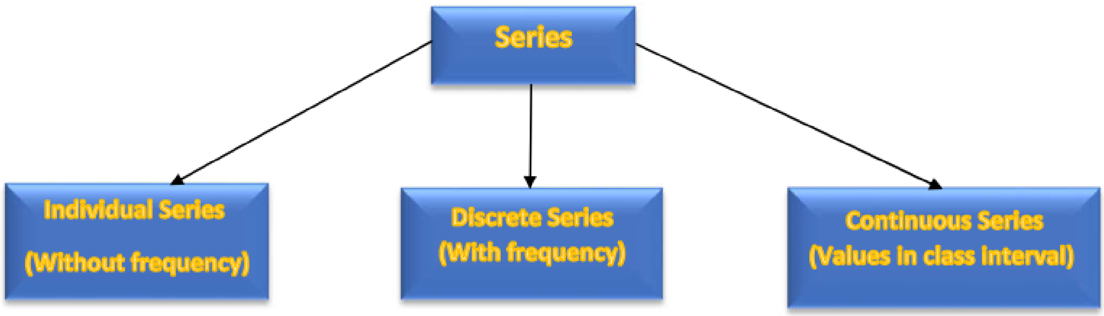
$$\textit{The width of interval } = \textit{Range/Number of classes}$$

d) Open end classes should be avoided since creates difficulty in analysis and interpretation.

e) Intervals would be continuous throughout the distribution. This is important for continuous distribution.

f) The lower limits of the class intervals should be simple multiples of the interval.

Frequency distribution (*Way of presentation of data*)



Example. In a factory, the time during working hours in which a machine is not operating as a result of breakage or failure is called the 'downtime". The following distribution shows a sample of 100 downtimes of a certain machine (rounded to the nearest minute). Calculate relative frequency, lesser Than Cumulative Frequency (Cum.freq less than) and Greater Than Cumulative Frequency (Cum.freq more than).

Downtime	Frequency	Relative frequency	Cum.freq less than	Cum.freq more than
$0 - 9.5$	3	$\frac{3}{100} = 0.03$	3	100
$9.5 - 19.5$	13	$\frac{13}{100} = 0.13$	16	97
$19.5 - 29.5$	30	$\frac{30}{100} = 0.3$	46	84
$29.5 - 39.5$	25	$\frac{25}{100} = 0.25$	71	54
$39.5 - 49.5$	14	$\frac{14}{100} = 0.14$	85	29
$49.5 - 59.5$	8	$\frac{8}{100} = 0.08$	93	15
$59.5 - 69.5$	4	$\frac{4}{100} = 0.04$	97	7
$69.5 - 79.5$	2	$\frac{2}{100} = 0.02$	99	3
$79.5 - 89.5$	1	$\frac{1}{100} = 0.01$	100	1
Total	100	1		

CONCEPT MAP

3.3 Graphical Descriptions of Data

3.3.1 Graphical Presentation

A graph is a method of presenting statistical data in visual form. The main purpose of any chart is to give a quick, easy-to-read-and-interpret pictorial representation of data which is more difficult to obtain from a table or a complete listing of the data. The type of chart or graphical presentation used and the format of its construction is incidental to its main purpose. A well-designed graphical presentation can effectively communicate the data's message in a language readily understood by almost everyone. You will see that graphical methods for describing data are intuitively appealing descriptive techniques and that they can be used to describe either a sample or a population; quantitative or qualitative data sets. Some basic rules for the construction of a statistical chart are listed below:

(a) Every graph must have a clear and concise title which gives enough identification of the graph.

(b) Each scale must have a scale caption indicating the units used.

(c) The zero point should be indicated on the co-ordinate scale. If, however, lack of space makes it inconvenient to use the zero point line, a scale break may be inserted to indicate its omission.

(d) Each item presented in the graph must be clearly labelled and legible even in black and white reprint.

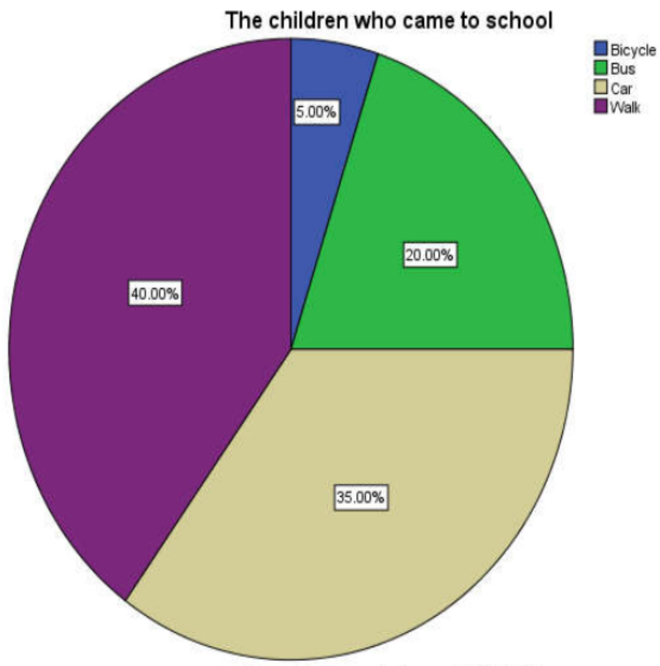
Pie charts

Pie charts (or Angular Circle Diagram) are widely used to show the component parts of a total. They are popular because of their simplicity. In constructing a pie chart, the angles of a slice from the center must be in proportion with the percentage of the total.
Some children survey all the children in school to record how they travelled to school on one day. They record the results in a pie chart. Estimate the percentages of children who came to school each way, explaining how you estimated each answer.

Bar charts

Bar charts is pictorial representation of data having rectangular bars of equal width. These bars are placed on equal spaces on one of the axis i.e bars can be either vertical or horizontal. Height of the bars depends on the given data where lower end of the bar touche the base line such that the height of each bar starts from zero units.

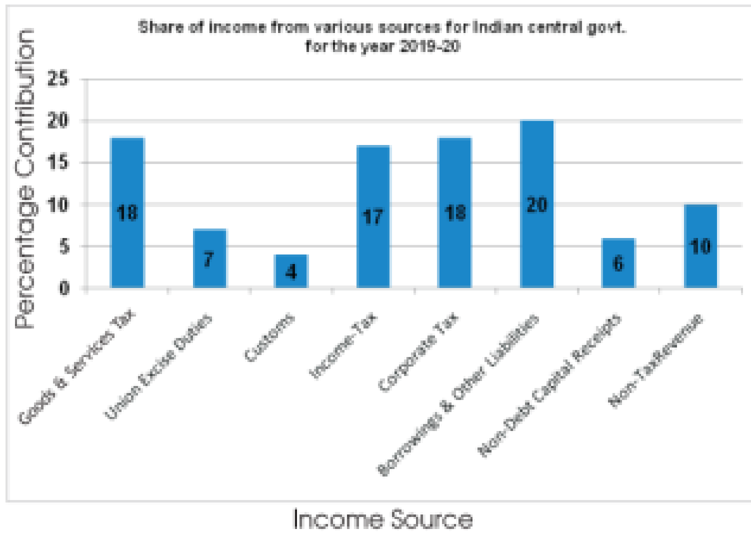
- Every bar depicts only one characteristics of the data.
- The distance between the bars should be equal.

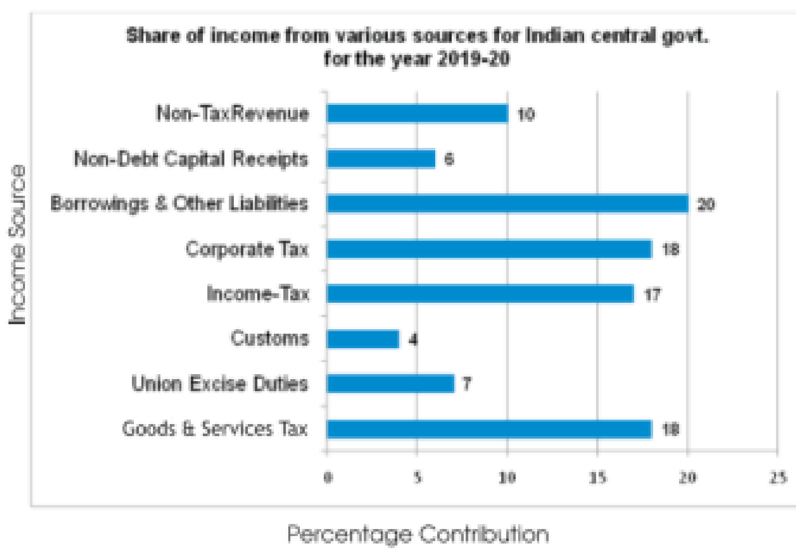


- These bars can be either vertical or horizontal.
- Bars of a bar diagram can be visually compared by their relation height and accordingly data can be comprehended quickly.

Example.1 Let us take disclosed in union budget $2020 - 21$ for percentage shares of earning for a Indian central govt. For the year $2019 - 20$.

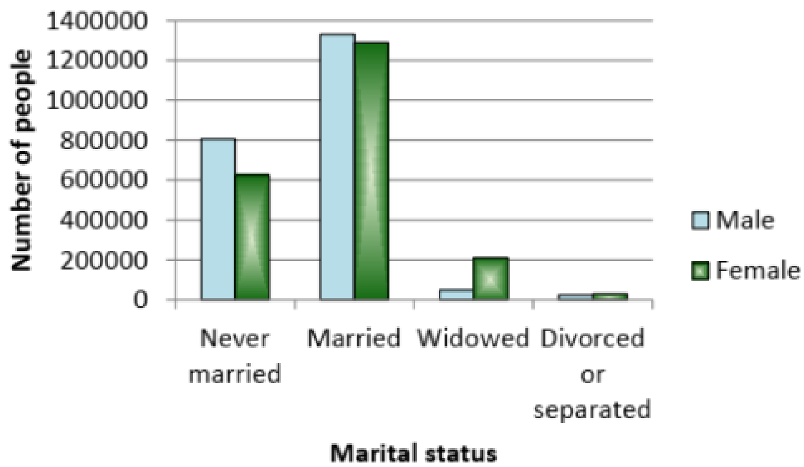
Income source	Percentage contribution
Goods -Services Tax	18%
Union Excise Duties	7%
Customs	4%
Income Tax	17%
Corporate Tax	18%
Borrowings-other liabilities	20%
Non Debt Capital Receipts	6%
Non Tax Revenue	10%


Share of income from various sources for Indian central govt.
for the year 2019-20
Percentage Contribution
25
20
15
10
5
0
18
7
4
17
18
20
6
10
Goods & Services Tax
Union Excise Duties
Customs
Income-Tax
Corporate Tax
Borrowings & Other Liabilities
Non-Debt Capital Receipts
Non-TaxRevenue
Income Source


Share of income from various sources for Indian central govt.
for the year 2019-20
Income Source
Non-TaxRevenue
Non-Debt Capital Receipts
Borrowings & Other Liabilities
Corporate Tax
Income-Tax
Customs
Union Excise Duties
Goods & Services Tax
10
6
20
18
17
4
7
18
0
5
10
15
20
25
Percentage Contribution

Multiple bar charts

A multiple bar charts is particularly useful if one desires to make quick comparison between different sets of data. In the following example, the marital status of male and female in Hong Kong are compared using multiple bar charts.



3.3.2 Histograms

In statistics, data is often represented using a histogram. A histogram is constructed by dividing the data into a number of classes and then number in each class or frequency is represented by a vertical rectangle. The area of the rectangle represents the frequency of each class.

Example.2 A traffic inspector has counted the number of automobiles passing a certain point in 100 successive 20-minute time periods. The observations are listed in table below.

Number of autos per period	Number of periods
$5 - 9$	3
$10 - 14$	9
$15 - 19$	36
$20 - 24$	35
$25 - 29$	12
$30 - 34$	3
$35 - 39$	2
$Total$	100

A histogram of these data has been drawn in Figure 3.5.

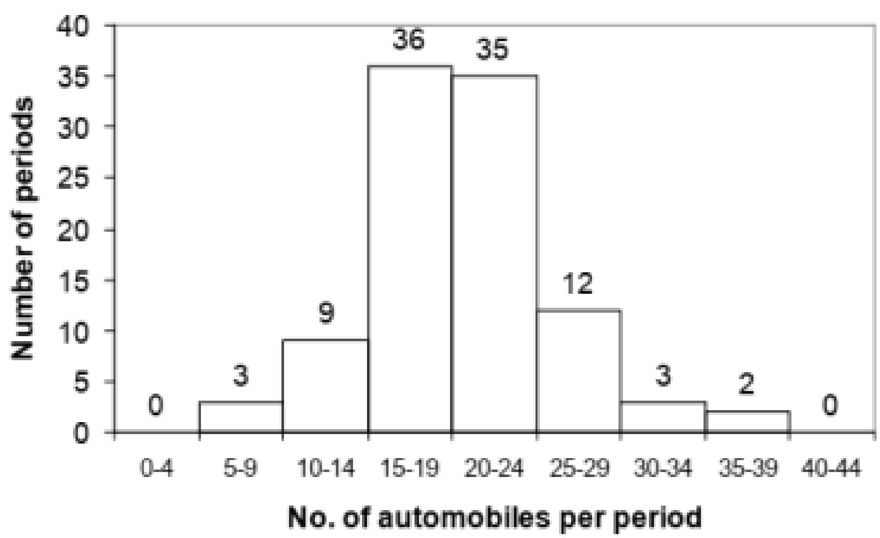


Figure 3.1: Histogram of the traffic data

3.3.3 Frequency Polygon

Another method to represent frequency distribution graphically is by a frequency polygon. As in the histogram, the base line is divided into sections corresponding to the class-interval, but instead of the rectangles, the points of successive class marks are being connected. The frequency polygon is particularly useful when two or more distributions are to be presented for comparison on the same graph.

Exemple.3 Construct a frequency polygon for the traffic data in Example 2.

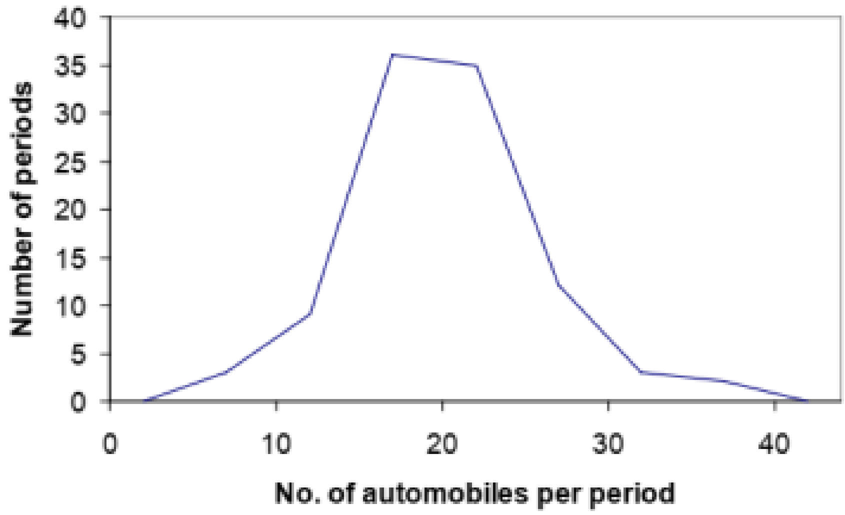


Figure 3.2: Frequency polygon for the traffic data

Cumulative Frequency Distribution and Cumulative Polygon

Sometimes it is preferable to present data in a cumulative frequency distribution, which shows directly how many of the items are less than, or greater then, various values.

less than	cumulative frequency
4.5	0
9.5	3
14.5	12
19.5	48
24.5	83
29.5	95
34.5	98
39.5	100

Example. 4

Construct a "Less-than" ogive of the distribution of traffic data.

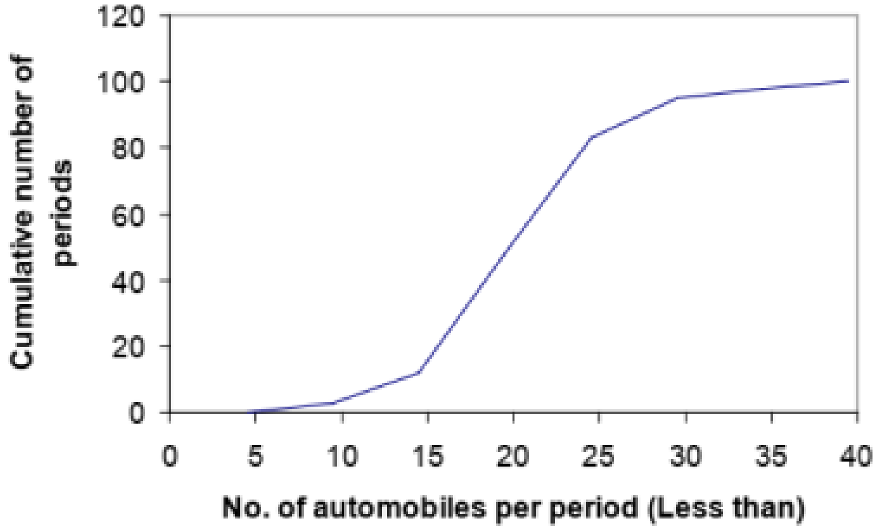


Figure 3.3: Cumulative frequency polygon for the traffic data.

3.4 Discrete variable

3.4.1 Measures of central tendency

Central tendency describes the tendency of the observations to bunch around a particular value, or category. The mean, median and mode are all measures of central tendency. They are all measures of the 'average' of the distribution.

Arithmetic Mean

The arithmetic population mean, $\overline{x}$, or simply called mean, is obtained by adding together all of the measurements and dividing by the total number of measurements taken.

Sample mean

The sample mean of the values is $x_1, x_2, ..., x_n$

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

Arithmetic mean can be used to calculate any numerical data and it is always unique. It is obvious that extreme values affect the mean. Also, arithmetic mean ignores the degree of importance in different categories of data.

Frequency data: suppose that the frequency of the class with midpoint x_i is n_i, for ($i = 1, 2, ..., m$). Then

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + ... + n_mx_m}{N} = \frac{1}{N}\sum_{i=1}^{m} n_ix_i$$

$$\overline{x} = f_1x_1 + f_2x_2 + ... + f_mx_m = \sum_{i=1}^{m} f_ix_i$$

Where $f_i = \frac{n_i}{N}$ and $N = \sum_{i=1}^{m} n_i$ total number of observations.

Example. 5 Given the following set of ungrouped data:

$$20, 10, 15, 2, 3, 25, 8, 14, 6, 9$$

Find the sample mean of the ungrouped data.

$$\overline{x} = \frac{20 + 10 + 15 + 2 + 3 + 25 + 8 + 14 + 6 + 9}{10} = 11.2$$

Example. 6 Accidents data: find the Arithmetic Mean.

Number of accidents x_i	0	1	2	3	4	5	6	7	8	Total
Frequency n_i	55	14	5	2	0	2	1	0	1	80
n_ix_i	0	14	10	6	0	10	6	0	8	54

$$\overline{x} = \frac{0 + 14 + 10 + 6 + 0 + 10 + 6 + 0 + 8}{80} = \frac{54}{80} = 0.675$$

Median

The median (med) is the central value in the sense that there as many values smaller than it as there are larger than it. All values known: if there are N observations then the median is:

- the $\frac{N+1}{2}$ largest value, if N is odd;
- the sample mean of the $\frac{N}{2}$ largest and $\left(\frac{N}{2}+1\right)$ the largest values, if N is even.

Example. 7 Given the following set of ungrouped data:

$$20, 10, 15, 2, 3, 25, 8, 14, 6, 9$$

We have the grouped data

$$2, 3, 6, 8, 9, 10, 14, 15, 20, 25$$

We noted that the total observations $N = 10$ is the even number, then the median is the mean of the $\frac{N}{2}$ largest and $\left(\frac{N}{2}+1\right)$ the largest values.

$$med = \frac{9+10}{2} = 9.5$$

Mode

The mode, or modal value, is the most frequently occurring value.

Examples. 1) The set of data $20, 12, 12, 10, 10, 27, 20, 30, 10$ has mode 10 (uni-modal)

2) The set of data $2, 3, 4, 5, 6, 8, 10$ has no mode.

3) The set of data $5, 6, 6, 6, 6, 8, 7, 11, 11, 11, 11, 15$ has two mode 6 and 11 (bimodal)

3.4.2 Measures of Dispersion

The mean is the value usually used to indicate the centre of a distribution. If we are dealing with quantity variables our description of the data will not be complete without a measure of the extent to which the observed values are spread out from the average.

We will consider several measures of dispersion and discuss the merits and pitfalls of each.

The range

Range is the difference between two extreme values. e.g.

$$Range = \max - \min$$

The range is easy to calculate but cannot be obtained if open ended grouped data are given.

Example. 8 The range of the following set of data

$$5, 8, 10, 1, 6, 1, 3, 3, 9$$

is $10 - 9 = 1$.

Example. 9 One very simple measure of dispersion is the range. Lets consider the two distributions given in Figures **3.8** and **3.9**. They represent the marks of a group of thirty students on two tests.

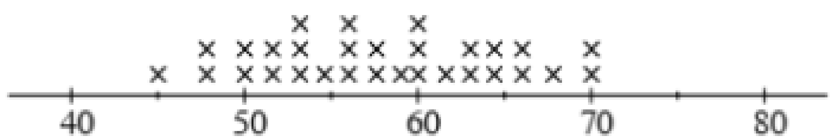


Figure 3.4: Marks of the test A

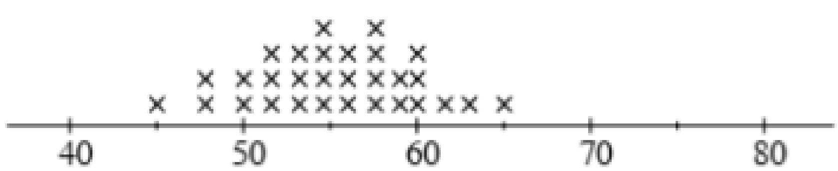


Figure 3.5: Marks of the test B

Here it is clear that the marks on test A are more spread out than the marks on test B, and we need a measure of dispersion that will accurately indicate this.

On test A, the range of marks is $70 - 45 = 25$.

On test B, the range of marks is $65 - 45 = 20$.

Here the range gives us an accurate picture of the dispersion of the two distributions.

However, as a measure of dispersion the range is severely limited. Since it depends only on two observations, the lowest and the highest, we will get a misleading idea of dispersion if these values are outliers.

We want a measure of dispersion that will accurately give a measure of the variability of the observations. We will concentrate now on the measure of dispersion most commonly used, the standard deviation.

Variance and Standard Deviation

The variance and standard deviation are two very popular measures of variation. Their formulations are categorized into whether to evaluate from a population or from a sample. The variance, σ^2 or Var , is the mean of the square of all deviations from the mean. Mathematically it is given as:

$$Var = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} n_i \left(x_i - \overline{x}\right)^2$$

$$Var = \sigma^2 = \sum_{i=1}^{N} f_i \left(x_i - \overline{x}\right)^2$$

or

$$Var = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} n_i \left(x_i\right)^2 - \overline{x}^2$$

$$Var = \sigma^2 = \sum_{i=1}^{N} f_i \left(x_i\right)^2 - \overline{x}^2$$

where: $f_i = \frac{n_i}{N}$

x_i is the value of the ith item;

$\overline{x}$ is the arithmetic mean;

N is the population size.

The standard deviation σ is defined as $\sigma = \sqrt{Var}$

Example. 10 Suppose a class of students conducted a quiz and grades obtained aregiven in following table. Find the variance Var and the standard deviation σ.

Grade	5	10	15	20	25
Number of students	3	7	5	6	4

Solution

x_i	n_i	$x_i n_i$	x_i^2	$n_i x_i^2$
5	3	15	25	75
10	7	70	100	700
15	5	75	225	1125
20	6	120	400	2400
25	4	100	625	2500
	$\sum n_i = 25$	$\sum x_i n_i = 380$		$\sum n_i x_i^2 = 6800$

$\overline{x} = \frac{1}{N} \sum x_i n_i = \frac{380}{25} = 15.2$

$Var = \frac{1}{N} \sum n_i x_i^2 - (\overline{x})^2 = \frac{6800}{25} - (15.2)^2 = 40.96$

$\sigma = \sqrt{Var} = \sqrt{40.96} = 6.4$

3.4.3 Measures of Position

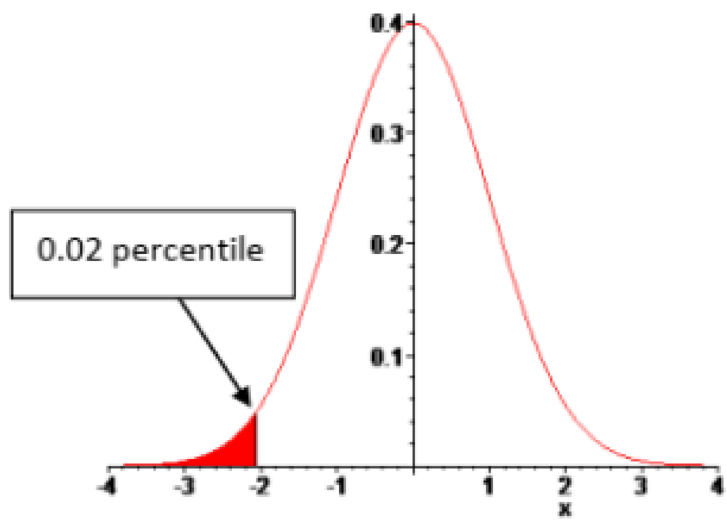
Measures of position identifies the postion of a value, relative to other values in a set of data. The most common measuers are of position are percentiles, quartiles and deciles, Here we shall study only percentiles and quartiles.

Percentile

The Pth percentile is the value corresponding to cumulative relative frequency of $P/100$ on the cumulative relative frequency diagram e.g. the $2nd$ percentile is the value corresponding to cumulative relative frequency 0.02.

Now a days, percentile is very commonly used for declaring the results where number of condidates appeared is very large. It reflect your individual scoring in comparison to all.

Percentiles are denoted by $P_1, P_2, ..., P_{100}$ and divide the distribution into 100 groups.



Quartiles and the interquartile range

First quartile: Q_1 represents the lower quartile, which is the $25th$ percentile denoted by Q_1. Q_1 is the number corresponding to the rank $\frac{N}{4}$ in the data set (in ascending order) or Q_1 contains upto 25% of data.

Second quartile: Q_2 is the second quartile, corresponding to the $50th$ percentile. Q_2 is the number corresponding to the rank $\frac{N}{2}$ in the data set (in ascending order) or Q_2 contains upto 50% of data.

Third quartile : Q_3 is the third quartile, which is the $75th$ percentile.

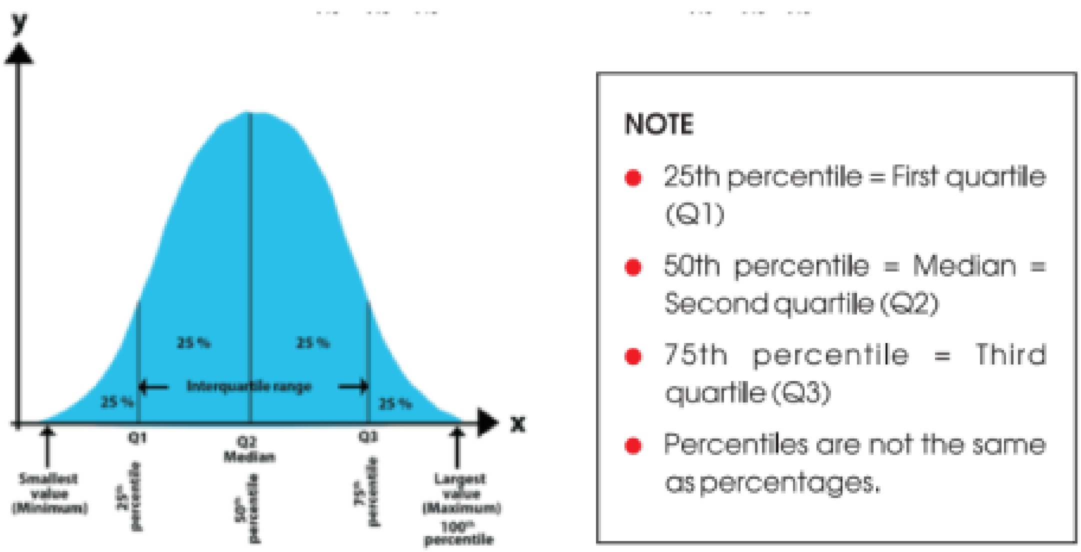
Q_3 is the number corresponding to the rank $\frac{3N}{4}$ in the data set (in ascending order) or Q_3 contains upto 75% of data.

Interquartile range: Interquartile range of a set of data is the difference between the third quartile and the first quartile e.g.

$$IQ = Q_3 - Q_1$$

Quartile Deviation: Quartile deviation is the semi-interquartile range defined by

$$QD = \frac{Q_3 - Q_1}{2}$$



The Box-plot

The box-plot is another way of representing a data set graphically. It is constructed using the quartiles, and gives a good indication of the spread of the data set and its symmetry (or lack of symmetry). It is a very useful method for comparing two or more data sets.

The box-plot consists of a scale, a box drawn between the first and third quartile, the median placed within the box, whiskers on both sides of the box and outliers (if any). This is best illustrated using a diagram such as Figure below.

Example. 11 Given the set of data $\mathbf{10, 2, 18, 7, 20, 12, 15, 16, 6, 4}$.

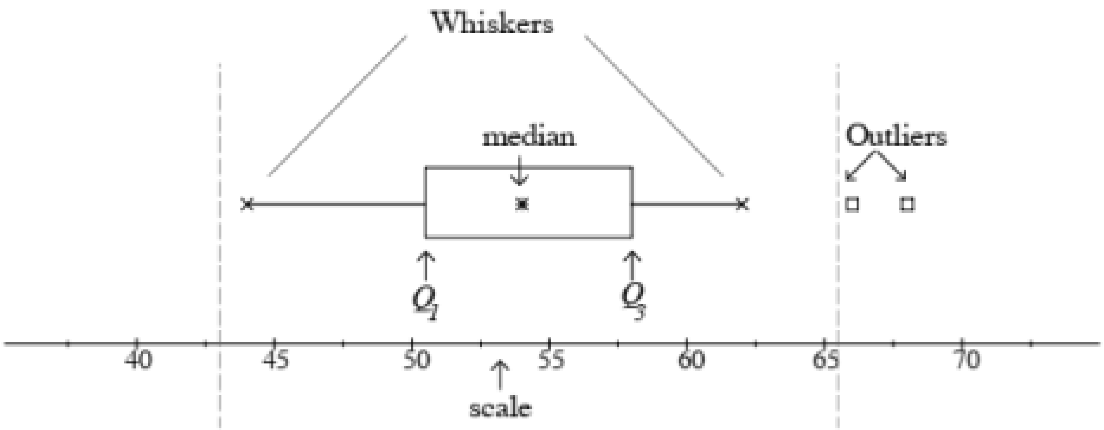


Figure 3.6: Annotated example of a box-plot.

Find the quartiles $Q_1 Q_2, Q_3$ and interquartile range IQ. Draw the box-plot.

Solution. Put the given data in ascending order

$2, 4, 6, 7, 10, 12, 15, 16, 18, 20.$

We have $\frac{N}{4} = \frac{10}{4} = 2.5 \sim 3$ Hence $Q_1 = 6$

$\frac{N}{2} = \frac{10}{2} = 5$ then $Q_2 = 10$

$\frac{3N}{4} = \frac{30}{4} = 7.5 \sim 8$ then $Q_3 = 16.$

$IQ = Q_3 - Q_1 = 16 - 10 = 6$

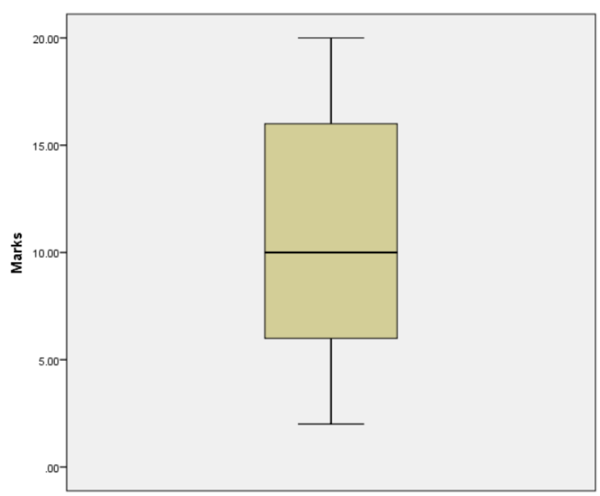


Figure 3.7: Box-plot for Example 11.

2) Given the set of data in the following table. Find the quartiles Q_1, Q_2, Q_3.

x_i	n_i	cumulative frequency (cf)	relative frequency f_i	cumulative relative frequency
10	5	5	0.25	0.25
12	3	8	0.15	0.4
13	2	10	0.1	0.5
16	4	14	0.2	0.7
18	6	20	0.3	1

By using the class of cumulative frequency, we have $0 < \frac{N}{4} = 5 \leq 5$ Hence $Q_1 = 10$ or by using the class of relative cumulative frequency, we have $0 < 0.25 = 25\% \leq 0.25$ i.e. $Q_1 = 10$; $8 < \frac{N}{2} = 10 \leq 10$ Hence $Q_2 = med = 12$ or we have $0.4 < 0.5 = 50\% \leq 0.5$ i.e. $Q_2 = 12$ and $14 < \frac{3N}{4} = 15 \leq 20$ Hence $Q_3 = 18$ or we have $0.7 < 0.75 = 75\% \leq 1$ i.e. $Q_3 = 18$.

Comparing sample standard deviation, interquartile range and range

The range is simple to evaluate and understand, but is sensitive to the odd extreme value and does not make effective use of all the information of the data. The sample standard deviation is also rather sensitive to extreme values but is easier to work with mathematically than the interquartile range.

Mean Absolute Deviation

1) **Mean Absolute Deviation about mean**

Mean absolute deviation is the mean of the absolute values of all deviations from the mean. Therefore it takes every item into account. Mathematically it is given as:

$$M.D\left(\overline{x}\right) = \frac{1}{N}\sum n_i \left|x_i - \overline{x}\right|$$

Where x_i is the value of the ith item;

$\overline{x}$ is the population arithmetic mean;

N is the population size.

2) **Mean Absolute Deviation about median**

$$M.D\left(med\right) = \frac{1}{N}\sum n_i \left|x_i - med\right|$$

Example. 12 Find the Mean Absolute Deviation about median and mean for following data

x_i	5	10	15	20	25
n_i	7	4	6	3	5

Solution.

x_i	n_i	$n_i x_i$	$\lvert x_i - \overline{x}\rvert$	$n_i \lvert x_i - \overline{x}\rvert$	$n_i \lvert x_i - med\rvert$
5	7	35	9	63	70
10	4	40	4	16	20
15	6	90	1	6	0
20	3	60	6	18	15
25	5	125	11	55	50
	$\sum n_i = 25$	$\sum n_i x_i = 350$		$\sum n_i \lvert x_i - \overline{x}\rvert = 158$	$\sum n_i \lvert x_i - med\rvert = 155$

Hence

$\frac{N+1}{2} = 13 \longrightarrow med = 15$

$$\overline{x} = \frac{1}{\sum n_i} \sum n_i x_i = \frac{350}{25} = 14$$

$$M.D\left(\overline{x}\right) = \frac{1}{\sum n_i} \sum n_i \left|x_i - \overline{x}\right| = \frac{158}{25} = 6.32$$

$$M.D\left(med\right) = \frac{1}{\sum n_i} \sum n_i \left|x_i - med\right| = \frac{155}{25} = 6.2$$

3.5 Continuous variable

3.5.1 Measures of Grouped Data

1) Mean suppose that the frequency of the class with midpoint x_i is n_i, for ($i = 1, 2, ..., m$). Then

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{m} n_i x_i = \sum_{i=1}^{m} f_i x_i$$

Where $f_i = \frac{n_i}{N}$ and $N = \sum_{i=1}^{m} n_i$ total number of observations.

2) Median for grouped data, the median can be found by first identify the class containing the median, then apply the following formula:

$$med = a + \frac{\frac{N}{2} - C}{n_m} L$$

Where a is the lower class boundary of the median class;

L is the lenght of median class;

N is the total frequency (i.e. the sample size);

C is the cumulative frequency just before the median class;

n_m is the frequency of the median class.

It is obvious that the median is affected by the total number of data but is independent of extreme values. However if the data is ungrouped and numerous, finding the median is tedious. Note that median may be applied in qualitative data if they can be ranked.

4) Quartiles for grouped data, first and third quartile can be found by the following formula

$$Q_1 = a + \frac{\frac{N}{4} - C}{n_{Q_1}} L$$

and

$$Q_3 = a + \frac{\frac{3N}{4} - C}{n_{Q_3}} L$$

Where a is the lower class boundary of the quartile class;

L is the lenght of quartile class;

N is the total frequency (i.e. the sample size);

C is the cumulative frequency just before the quartile class;

n_{Q_3}, n_{Q_1} are the frequency of the third quartile class and first quartile class respectively.

3) Mode for grouped data, the mode can be found by first identify the largest frequency of that class, called modal class, then apply the following formula on the modal class:

$$mode = a + \frac{d_1}{d_1 + d_2} L$$

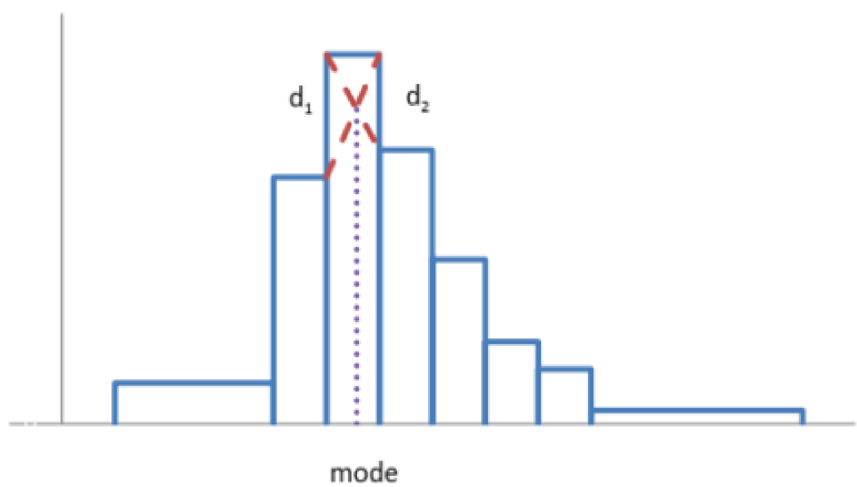
Where a is the lower class boundary of the modal class;

L is the lenght of modal class;

d_1 is the difference of the frequencies of the modal class with the previous class and is always positive;

d_2 is the difference of the frequencies of the modal class with the following class and is always positive.

Geometrically the mode can be represented by the following graph and can be obtained by using similar triangle properties. The formula can be derived by interpolation using second degree polynomial.



6) Variance and Standard Deviation suppose that the frequency of the class with midpoint x_i is n_i, for ($i = 1, 2, ..., m$). Then

$$\sigma^2 = Var = \frac{1}{N}\sum_{i=1}^{m} n_i \left(x_i - \overline{x}\right)^2$$

$$Var = \frac{1}{N}\sum_{i=1}^{m} n_i \left(x_i\right)^2 - \overline{x}^2$$

or

$$\sigma^2 = Var = \sum_{i=1}^{m} f_i \left(x_i - \overline{x}\right)^2$$
$$Var = \sum_{i=1}^{m} f_i \left(x_i\right)^2 - \overline{x}^2$$

and

$$\sigma = \sqrt{Var}$$

Where $f_i = \frac{n_i}{N}$ and $N = \sum_{i=1}^{m} n_i$ total number of observations.

3.6 Coefficient of Variation

The coefficient of variation is a measure of relative importance. It does not depend on unit and can be used to make comparison even two samples differ in means or relate to different types of measurements.

The coefficient of variation gives:

$$CV = \frac{\sigma}{\overline{x}} \times 100\%$$

3.7 Skewness and Kurtosis

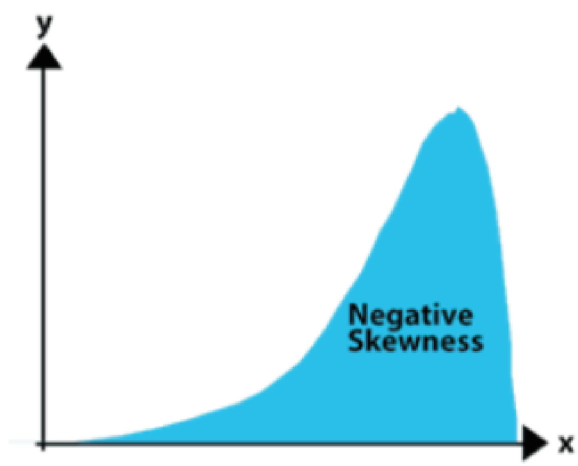
1) Skewness

The skewness is an abstract quantity which shows how data piled-up. A number of measures have been suggested to determine the skewness of a given distribution.

(a) Positive skewness If the longer tail is on the right, we say that it is skewed to the right (positively skew), and the coefficient of skewness is positive. In this case, Mean>Median>mode

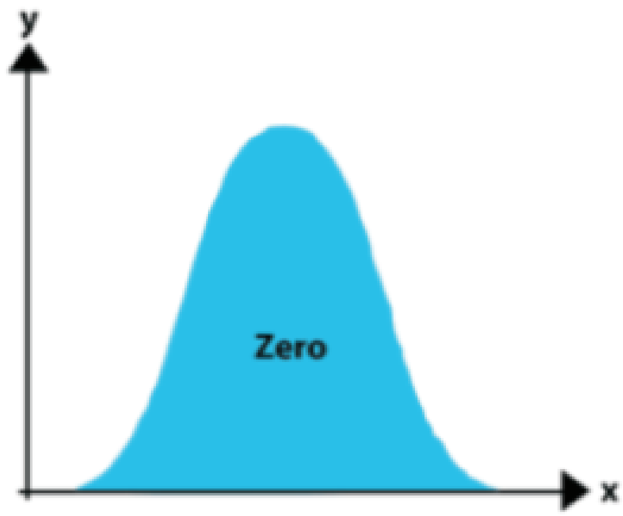
(b) Negative skewness If the longer tail is on the left, we say that is skewed to the left(negatively skew) and the coefficient of skewness is negative. In this case, Mean<Median<mode


y
Positive Skewness
x


y
Negative Skewness
x

(c) Zero skewness

When skewness is zero, distribution is called symmetical



Coefficient of Skewness

Pearson's 1^{st} coefficient of skewness,

$$CK = \frac{\overline{x} - Mode}{\sigma}$$

Pearson's 2^{nd} coefficient of skewness

$$CK = \frac{3\left(\overline{x} - Med\right)}{\sigma}$$

For moderately skewed distribution data, their relationship can be given by

$$\overline{x} - Mode \approx 3\left(\overline{x} - med\right)$$

Hence the Coefficient of Skewness is given by

$$CK = \frac{\overline{x} - Mode}{\sigma} \approx \frac{3\left(\overline{x} - Med\right)}{\sigma}$$

Example. 13

Class interval	Frequency n_i	middle points x_i	cf	class(cf)	$n_i x_i$	$n_i x_i^2$
$9.5 - 19.5$	1	14.5	1	$0 - 1$	14.5	210.25
$19.5 - 29.5$	0	24.5	1	$1 - 1$	0	0
$29.5 - 39.5$	1	34.5	2	$1 - 2$	34.5	1190.25
$39.5 - 49.5$	4	44.5	6	$2 - 6$	178	7921
$49.5 - 59.5$	7	54.5	13	$6 - 13$	381.5	20791.75
$59.5 - 69.5$	16	64.5	29	$13 - 29$	1032	66564
$69.5 - 79.5$	19	74.5	48	$29 - 48$	1415.5	105454.8
$79.5 - 89.5$	20	84.5	68	$48 - 68$	1690	142805
$89.5 - 99.5$	17	94.5	85	$68 - 85$	1606.5	151814.3
$99.5 - 109.5$	11	104.5	96	$85 - 96$	1149.5	120122.8
$109.5 - 119.5$	3	114.5	99	$96 - 99$	343.5	39330.75
$119.5 - 129.5$	1	124.5	100	$99 - 100$	124.5	15500.25
Total	100				7970	671705

Where middle point of $[a, b]$ is $\frac{a+b}{2}$

Hence the mean

$$\overline{x} = \frac{1}{N}\sum n_i x_i = \frac{7970}{100} = 79.7$$

The median class is $[79.5, 89.5]$ $\left(\frac{N}{2} = \frac{100}{2} = 50\,(cf = 48 < 50 < cf = 68)\right)$, the median is given by

$$med = a + \frac{\frac{N}{2} - C}{n_m}L = 79.5 + \frac{50 - 48}{20} \times 10 = 80.5$$

The first quartile class is $[59.5, 69.5]$ $\left(\frac{N}{4} = \frac{100}{4} = 25\,(cf = 13 < 25 < cf = 29)\right)$, the first quartile Q_1 is given by

$$Q_1 = a + \frac{\frac{N}{4} - C}{n_{Q_1}}L = 59.5 + \frac{25 - 13}{16} \times 10 = 67$$

The third quartile class is $[89.5, 99.5]$ $\left(\frac{3N}{4} = \frac{300}{4} = 75\,(cf = 68 < 75 < cf = 85)\right)$, the third quartile Q_3 is given by

$$Q_3 = a + \frac{\frac{3N}{4} - C}{n_{Q_3}} L = 89.5 + \frac{75 - 68}{17} \times 10 = 93.6$$

The modal class is $[79.5, 89.5]$, the largest frequency is $(n_i = 20)$, the mode is given by

$$mode = a + \frac{d_1}{d_1 + d_2} L = 79.5 + \frac{20 - 19}{(20 - 19) + (20 - 17)} \times 10 = 82$$

The variance

$$Var = \frac{1}{N} \sum n_i x_i^2 - \overline{x}^2 = \frac{1}{100} \times 671705 - (79.7)^2 = 364.96$$

The standard deviation $\sigma = \sqrt{Var} = 19.104$

The coefficient of variation $CV = \frac{\sigma}{\overline{x}} \times 100 = \frac{19.104}{79.7} \times 100 \sim 24\%$

The coefficient of skewness

$$CK = \frac{\overline{x} - Mode}{\sigma} = \frac{79.7 - 82}{19.104} = -0.120\,39$$
$$CK = \frac{3\,(\overline{x} - Med)}{\sigma} = \frac{3\,(79.7 - 80.5)}{19.104} = -0.125\,63$$

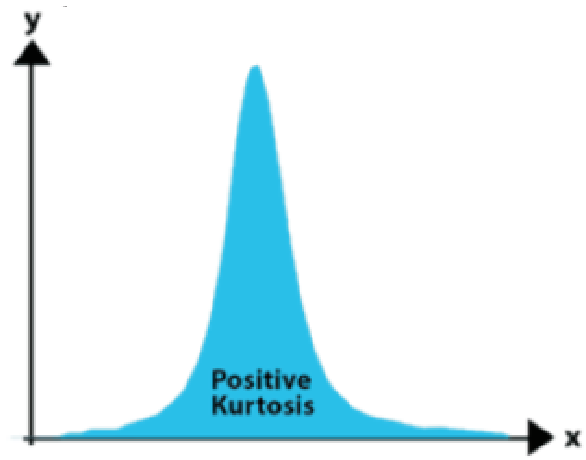
2) Kurtosis

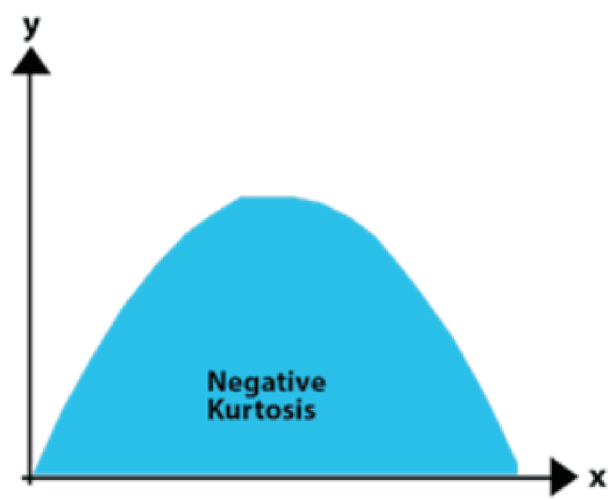
Kurtosis denotes the degree of peakedness of frequency curve. It is used to specify the frequency curve as regards the sharpness of its peak. Kurtosis can be positive, negative or zero.

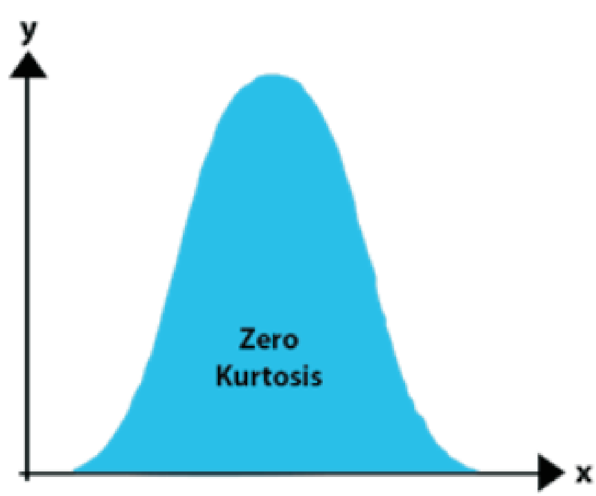
(a) Positive Kurtosis: A frequency distribution of data set having a sharp peak (heavy tailed)/outliers.

(b) Negative Kurtosis: A frequency distribution of data set having a blunt peak (light tailed).

(c) Zero Kurtosis: A frequency distribution of data set having a moderate peak. This means the Kurtosis is same as the normal distribution.


y
Positive
Kurtosis
x


y
Negative
Kurtosis
x


y
Zero
Kurtosis
x

3.8 Exercises with solutions

Exercise 1.

The numbers of accidents experienced by 80 machinists in a certain industry over a period of one year were found to be as shown below.

1) Construct a frequency table and draw a bar chart.

2) Draw the graph of cumulative frequency.

2 0 0 1 0 3 0 6 0 0 8 0 2 0 1

5 1 0 1 1 2 1 0 0 0 2 0 0 0 0

0 0 0 0 0 0 0 0 0 0 0 0 1 0 1

0 0 0 5 1 0 0 0 0 0 0 0 0 1 1

0 3 0 0 1 1 0 0 0 2 0 1 0 0 0

0 0 0 0 0

Solution:

Number of accidents	Frequency	Relative frequency	Cumulative frequency
0	55	$\frac{55}{80}$	55
1	14	$\frac{14}{80}$	69
2	5	$\frac{5}{80}$	74
3	2	$\frac{2}{80}$	76
4	0	0	76
5	2	$\frac{2}{80}$	78
6	1	$\frac{1}{80}$	79
7	0	0	79
8	1	$\frac{1}{80}$	80

This graph is obtained by spss statistics program.

Exercise 2.

A sample is composed by 120 males and 80 females. The following table shows their age in years with the percentage distribution by gender.

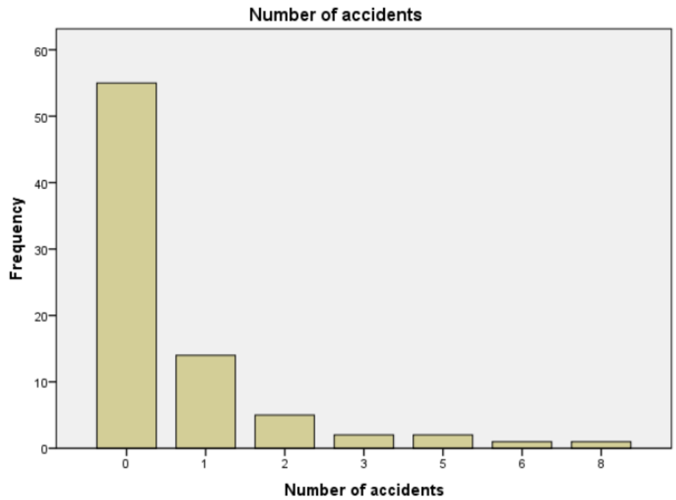


Figure 3.8: Bar charts of Exercise 1

Age(years)	Males(%)	Females(%)
$0-19$	10	20
$20-29$	10	20
$30-49$	30	30
$50-89$	50	30
$Total$	100	100

a. How many people are < 20 years old?

b. What is the percentage of individuals that are ≥ 50 years old?

c. How many males are ≥ 30 years old?

d. Find the modal classes for males and females separately and for the total sample.

e. Identify the median of the total sample.

Solution:

We have $N = 120$ for males or $N = 80$ for females and relative frequency $f_i = \frac{n_i}{N} \rightarrow n_i = N \times f_i$

Age(years)	Males($M_i\%$)	n_i	Females($F_i\%$)	n_i
$0-19$	10	12	20	16
$20-29$	10	12	20	16
$30-49$	30	36	30	24
$50-89$	50	60	30	24
Total	100	120	100	80

a. 28 subjects are < 20 years old

b. $p = \frac{60+24}{120+80} = 0.42 = 42\%$

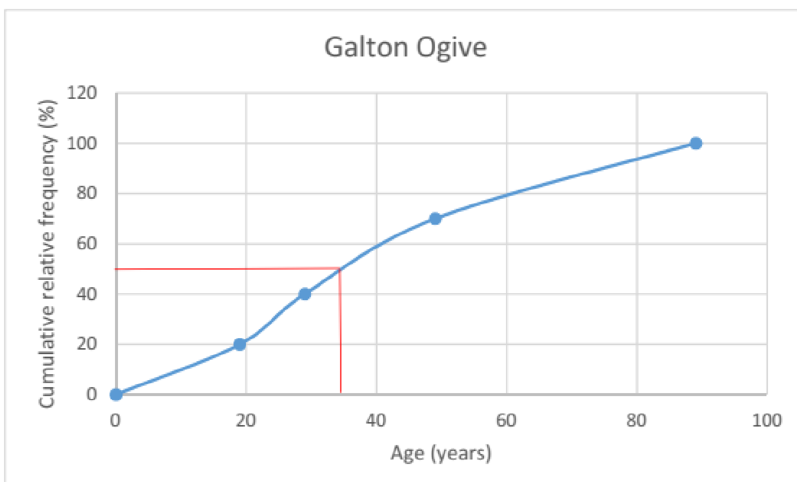
c. 96 males are ≥ 30 years old.

d. Modal class for males: $50-89$ years

Modal class for females: $30-49$ and $50-89$ years $\rightarrow$ bimodal distribution

Modal class overall: $50-89$ years.

e. The median age of the total sample is 35 years (see figure below).



Exercise 3

Five men with obesity have been visited in the same day. The following table shows their weights (kg):

patient ID	Weight (kg)
1	120
2	147
3	132
4	128
5	138

1) Calculate mean and standard deviation

2) Calculate median, Q_1, Q_3.and draw the box-plot.

The scale was later discovered to have been calibrated badly and that all measurements were wrong over estimated by $5kg$.

3) Calculate mean and standard deviation

4) Calculate mean and standard deviation in hg

5) Calculate the coefficient of variation of the weight both in kg and hg

Solution:

1) $\overline{x} = \frac{120+147+132+128+138}{5} = 133kg$

Variance

$$Var = \frac{1}{5}\sum_{i=1}^{5}(x_i - \overline{x})^2 = \frac{416}{5} = 83.2$$

Standard deviation $\sigma = \sqrt{83.2} = 9.12$

2) Median: the set of data $120, 128, 132, 138, 147$ median is $med = 132$

$Q_1 = 128\left(\frac{N}{4} = \frac{5}{4} = 1.2 \sim 2\right), Q_3 = 138\left(\frac{3N}{4} = \frac{3\times5}{4} = 3.8 \sim 4\right)$

The boxplot is obtained by spss statistics program.

3) $\overline{x} = 133 - 5 = 128kg$

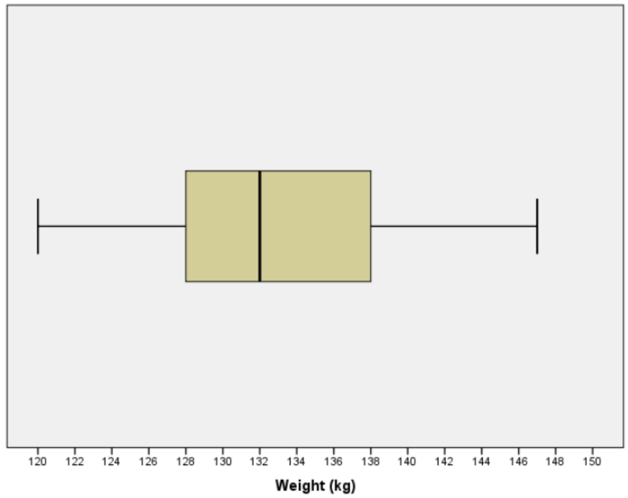
$\sigma = 9.12$ →remain unchanged.

4) $1kg = 10hg$ so

$\overline{x} = 128 \times 10 = 1280hg$

$\sigma = 9.12 \times 10 = 91.2hg$

5) The coefficient of variation



$CV = \frac{9.12}{128} \times 100 = \frac{91.2}{1280} \times 100 = 7.125$

Exercise 4

The following table shows the glycemia (mg/dL) of 500 older adults grouped in 5 classes having the same width:

Class interval	Frequency n_i	Middle point x_i	Cumulative frequency (cf)	$n_i x_i$	$n_i x_i^2$
$[65, 75[$	75				
$[75, 85[$	100				
$[85, 95[$	150				
$[95, 105[$	125				
$[105, 115]$	50				

1) Determine the sample, caracter, nature and complete the table.

2) Calculate mean, median, mode , variance and standard deviation.

3) Calculate Coefficient of Variation, Skewness, mean absolute deviation.

4) Draw the histogram of data with normal curve and frequency polygon.

Solution:

Class interval	Frequency n_i	Middle point x_i	cf	Class (cf)	n_ix_i	$n_ix_i^2$
$[65, 75[$	75	70	75	$0-75$	5250	367500
$[75, 85[$	100	80	175	$75-175$	8000	640000
$[85, 95[$	150	90	325	$175-325$	13500	1215000
$[95, 105[$	125	100	450	$325-450$	12500	1250000
$[105, 115]$	50	110	500	$450-500$	5500	605000
Total	500				44750	4077500

The mean

$$\overline{x} = \frac{1}{N}\sum nix_i = \frac{44750}{500} = 89.5$$

The median: we have $\frac{N}{2} = 250\,(cf = 175 < 250 < cf = 325)$ then class median $[85, 95[$

$$med = a + \frac{\frac{N}{2} - C}{n_m} \times L = 85 + \frac{250 - 175}{150} \times 10 = 90$$

The first quartile: we have $\frac{N}{4} = 125\,(cf = 75 < 125 < cf = 175)$ then class first quartie $[75, 85[$

$$Q_1 = a + \frac{\frac{N}{4} - C}{n_{Q_1}} \times L = 75 + \frac{125 - 75}{100} \times 10 = 80$$

The third quartile: we have $\frac{3N}{4} = 375\,(cf = 325 < 250 < cf = 450)$ then class third quartile $[95, 105[$

$$Q_3 = a + \frac{\frac{3N}{4} - C}{n_{Q_3}} \times L = 95 + \frac{375 - 325}{125} \times 10 = 99$$

The modal class is $[85, 95]$, the largest frequency is $(n_i = 150)$, the mode is given by

$$mode = a + \frac{d_1}{d_1 + d_2}L = 85 + \frac{150 - 100}{(150 - 100) + (150 - 125)} \times 10 = 91.66$$

Variance

$$Var = \frac{1}{N}\sum n_ix_i^2 - (\overline{x})^2 = \frac{4077500}{500} - (89.5)^2 = 144.75$$

Standard deviation

$$\sigma = \sqrt{Var} = 12.03$$

The coefficient of variation $CV = \frac{\sigma}{\overline{x}} \times 100 = \frac{12.03}{89.5} \times 100 \sim 13.44\%$

The coefficient of skewness

$$CK = \frac{\overline{x} - Mode}{\sigma} = \frac{89.5 - 91.66}{12.03} = -0.179$$

The mean Absolute Deviation about mean

$$M.D\left(\overline{x}\right) = \frac{1}{N} \sum n_i \left|x_i - \overline{x}\right| = \frac{4825}{500} = 9.65$$

4) These graphs are obtained by spss statistics program.

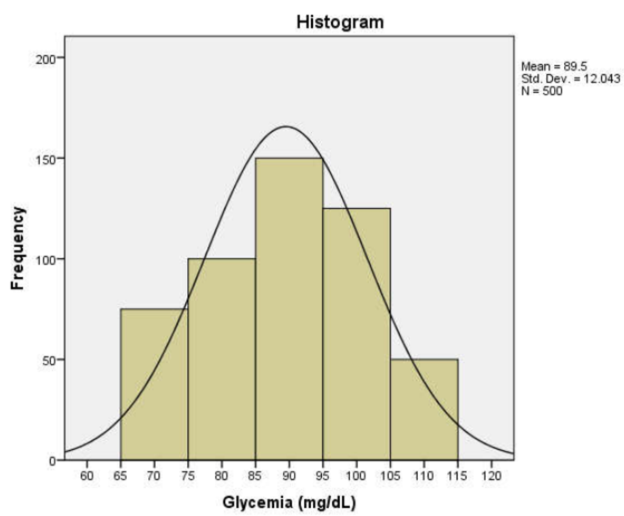


Figure 3.9: Histogram of data with normal curve

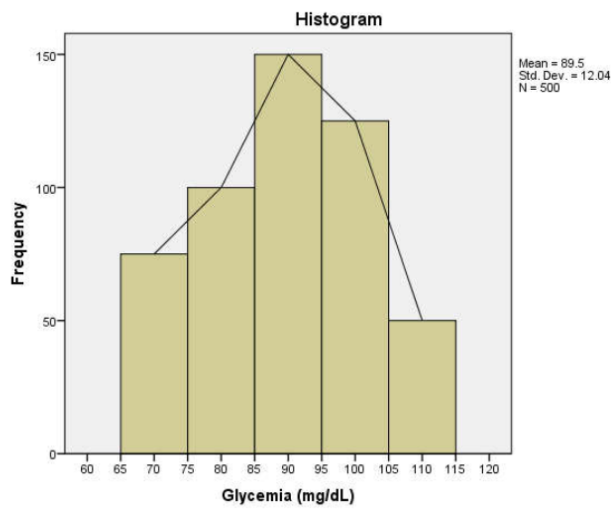


Figure 3.10: Frequency polygon of data with histogram.

Bibliography

[1] Lang, Serge (1997), Undergraduate analysis, Undergraduate Texts in Mathematics (2nd ed.), Berlin, New York: Springer-Verlag.

[2] Kaur P, Stoltzfus J, Yellapu V (2018), Descriptive statistics. Int J Acad Med, 4:3-60.

[3] Bijma, F., Jonker, M., & van der Vaart, A. (2017). Descriptive Statistics. In An Introduction to Mathematical Statistics, Amsterdam University Press, 21-44.

[4] Larsen, R. J., & Marx, M. L. (2017). An introduction to mathematical statistics and its applications (Vol. 5). Pearson.

[5] Witte, R. S., & Witte, J. S. (2017). Statistics. Wiley

[6] Shurman, Jerry (2016). Calculus and Analysis in Euclidean Space (illustrated ed.). Springer, 271–272.

[7] Harper, J.F. (2016), "Defining continuity of real functions of real variables", BSHM Bulletin: Journal of the British Society for the History of Mathematics, 31 (3): 1–16.

[8] Sahai, H., & Ojeda, M. M. (2004). Analysis of Variance for Random Models, Volume 2: Unbalanced Data: Theory, Methods, Applications, and Data Analysis (Vol. 2). Springer Science & Business Media.

[9] Strang, Gilbert (1991). Calculus. SIAM. p. 702.

[10] Stewart, James (2008). Calculus: Early Transcendentals.