Descriptive Statistics and Probability Distribution

Course Work: Quantitative Techniques

Introduction

Statistics may be regarded as the

study of Populations

- aggregates of individuals rather than of individuals
- the totality of individual observations about which inferences are to be made
- If an observation such as a simple measurement be repeated indefinitely the aggregate of the results is a population of measurements

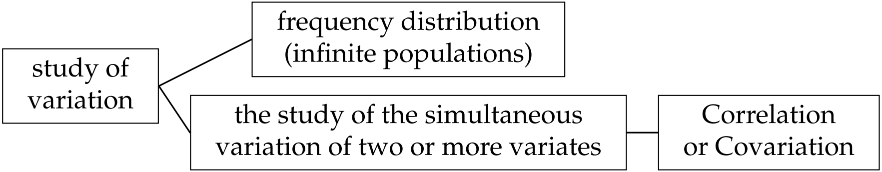
Statistics may be regarded as the

study of variation

The populations which are the object of statistical study always display variation in one or more respects

the study of the causes of variation of any variable phenomenon should be begun by the’examination and measurement of the variation which presents itself

Statistics may be regarded as the



Statistics may be regarded as the

methods of the
reduction of data:
(the sets of values
before us
are interpreted
as a random sample)

Problems of Specification
which arise in the choice of
the mathematical form of the population
(includes parameters)

problems of Estimation:
These estimates which are termed statistics
are calculated from the observations

Problems of Distribution:the exact nature
of the distributions in random samples of our
estimates of the parameters

Frequency Distribution (Variable is discrete)

Frequency Distribution

The following table 2.1 show hospital record of number of days 60 patients stayed in ICU.

Table 2.1: Number of days stayed in ICU of 60 patients

5	2	2	3	1	2	4	2	4	3
1	3	5	5	5	5	5	1	3	5
2	1	5	1	3	4	4	5	4	4
5	2	4	1	3	3	2	2	2	4
5	4	2	4	2	5	4	2	5	2
4	3	5	3	1	4	4	4	4	2

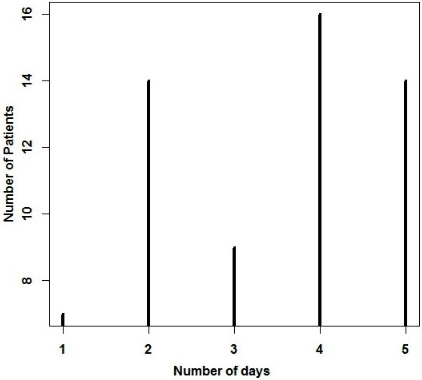
Frequency Table

Table 2.2: Frequency and Relative Frequency

Number of days	Number of patients	Relative Frequency(%)
1	7	12%
2	14	23%
3	9	15%
4	16	27%
5	14	23%
Total	60	100%

Histogram

The Histogram for Number of days stayed in ICU (60 patients)



Frequency Distribution of Continuous Variable

Comparison of theoretical with the observed frequency distribution

The ordinate of the curve gives the density for a given value of the variable shown along the abscissa. By density we mean the relative concentration of variates along the Y-axis.The following table represents raw data containing hemoglobin levels for 90 high – altitude miners in grams per cubic centimeter.

Table 3.1: hemoglobin levels

18.5	16.8	23.2	19.4	19.5	20.6	22	17.8	16.2
23.3	19.7	21.6	24.2	21.4	20.8	19.7	21.1	23
21.7	18.4	22.7	20.9	20.5	16.1	16.9	24.8	12.2
17.4	17.8	19.3	17.3	18.3	17.8	17.1	18.4	19.7
17.8	19	19.2	15.5	26.2	19.1	20.9	18.0	21
20.2	18.3	19.2	17.2	19.8	19.5	20.0	18.4	15.9
19.9	16.4	18.4	17.8	23	19.4	20.3	18.2	13.1
20.3	18.5	24.1	14.3	17.8	19.9	23.5	19.7	19.3
20.6	18.3	20.8	17.6	18.1	19.7	19.1	19.5	23.5
18.5	20.0	22.4	18.8	16.2	15.6	15.5	18.5	19.0

Comparison of theoretical with the observed frequency distribution

Table 3.2: Frequency Distribution

Hb level	Hb level	Number of Workers
Lower Limit	Upper Limit	Frequency
12.0	13.9	2
14.0	15.9	5
16.0	17.9	17
18.0	19.9	36
20.0	21.9	17
22.0	23.9	9
24.0	25.9	3
26.0	27.9	1
	Total	90

Comparison of theoretical with the observed frequency distribution

Table 3.3: Boundaries and Relative Frequency

LB	UB	Frequency	Relative
11.95	13.95	2	2%
13.95	15.95	5	6%
15.95	17.95	17	19%
17.95	19.95	36	40%
19.95	21.95	17	19%
21.95	23.95	9	10%
23.95	25.95	3	3%
25.95	27.95	1	1%
	Total	90	100%

Comparison of theoretical with the observed frequency distribution

Table 3.4: Cumulative Frequency Distribution: Hemoglobin Level

	Cumulative Frequency			Cumulative Frequency	
LB	More than type	Relative	UB	Less than Type	Relative
11.95	90	100%	13.95	2	2%
13.95	88	98%	15.95	7	8%
15.95	83	92%	17.95	24	27%
17.95	66	73%	19.95	60	67%
19.95	30	33%	21.95	77	86%
21.95	13	14%	23.95	86	96%
23.95	4	4%	25.95	89	99%
25.95	1	1%	27.95	90	100%

Comparison of theoretical with the observed frequency distribution

Using table values (See Table 3.4) answer the questions in following example.

Example 3.1

Determine the percentage of workers whose hemoglobin level,

(a) *less than 15.95*

(b) *less than 21.95*

(c) *more than 13.95*

(d) *more than 21.95*

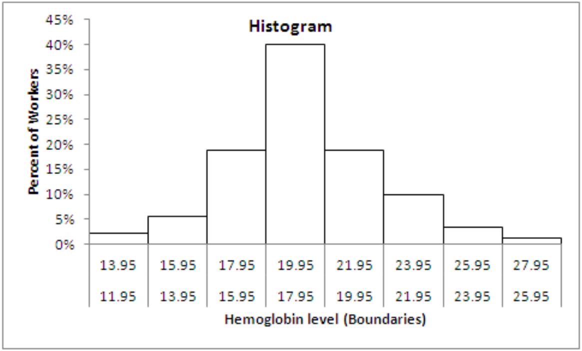
(e) *Between 17.95 and 21.95*

Solution 3.1

(a) 8% *(b)* 86% *(c)* 98% *(d)* 14% *(e)* $86\% - 27\% = 59\%$ *or* $73\% - 14\% = 59\%$

Comparison of theoretical with the observed frequency distribution

Figure 1: Histogram:Relative Frequency

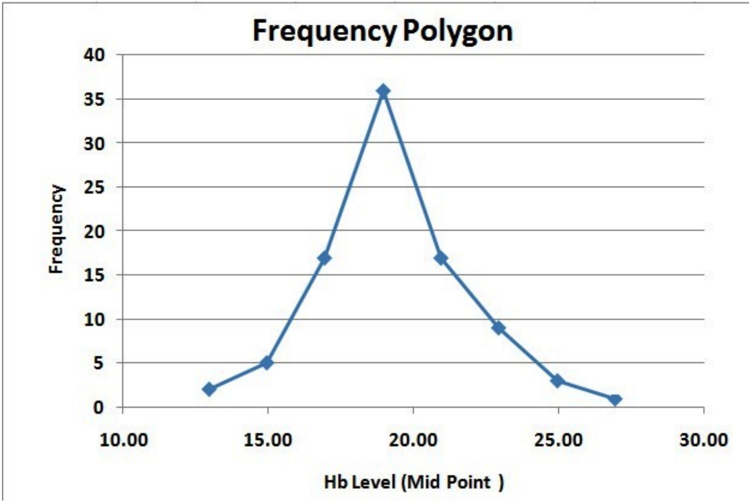


Comparison of theoretical with the observed frequency distribution

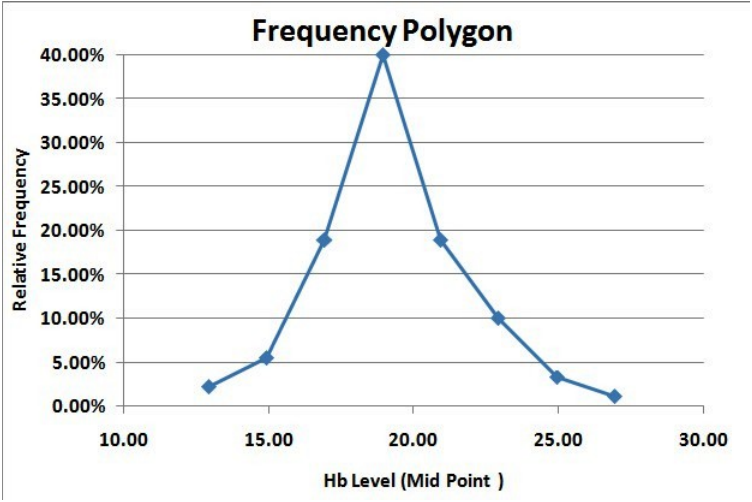
Table 3.5: Mid Point for Polygon

LB	UB	Frequency	Relative	Mid-Point
11.95	13.95	2	2.22%	12.95 (11.95+13.95)/2
13.95	15.95	5	5.56%	14.95 (12.95+2)
15.95	17.95	17	18.89%	16.95 (14.95+2)
17.95	19.95	36	40.00%	18.95 (16.95+2)
19.95	21.95	17	18.89%	20.95 (18.95+2)
21.95	23.95	9	10.00%	22.95 (20.95+2)
23.95	25.95	3	3.33%	24.95 (22.95+2)
25.95	27.95	1	1.11%	26.95 (24.95+2)
	Total	90	100.00%	

Comparison of theoretical with the observed frequency distribution



Comparison of theoretical with the observed frequency distribution



Describing Data

Quantitative Variable

After frequency distribution, the next step is the calculation of certain values which may be used as descriptive of the characteristics of that distribution. These values will enable comparisons to be made between one series of observations and another.

Summary

Two principal characteristics of the distribution

- average value of distribution

Summary

Two principal characteristics of the distribution

- ▷ average value of distribution
- ▷ the degree of scatter of the observations round that average value.

Average Values

Arithmetic Mean

The arithmetic mean of a variable is obtained by dividing the sum of its given values by their number. If the variable is denoted by x and if n values of x are given: $x_1, x_2, \ldots, x_n$, then arithmetic mean of x is

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

Average Values

Median

If the given values of x are arranged in an increasing or decreasing order of magnitude, then middle-most value in this arrangement is called median of x. The median may alternatively be defined as a value of x such that half of the given values of x are smaller than or equal to it and half are greater than or equal to it.

Average Values

Median

When the number of values, n is odd, the middle-most value- that is $\frac{(n+1)}{2}$th value in arrangement will be the unique median of x.
When n is even, there will be no unique median. Any number between $\frac{n}{2}$th and $\left(\frac{n}{2}+1\right)$st values of x in the arrangement, being regarded as middle-most. The arithmetic mean of $\frac{n}{2}$th and $\left(\frac{n}{2}+1\right)$st values is accepted as the median of x.

Average Values

Mode

The mode of a variable is the value of the variable having the highest frequency.

Illustration

Computations of descriptive statistics

Example 5.1

A Random sample of 9 patients with BMI values is given in following table (7.1), Compute descriptive statistics that you know.

Table 5.1: Random Sample and BMI of patient, $n = 9$

42.10	47.78	33.23	36.42	42.10
24.54	25.21	27.78	54.33	

Summary:Arithmetic Mean

Solution 5.1

Let us define a variable as BMI of Selected patients and denoted by x say. The computation of some of descriptive statistics are shown below

Arithmetic Mean:

$$
\begin{aligned}
\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
&= \frac{333.49}{9} \\
&= 37.05444 kg/cm^2
\end{aligned}
$$

Summary:Median

Solution 5.2

Median: Arranging BMI in increasing order of values we obtain

Table 5.2: Ascending order:BMI values

24.54	25.21	27.78	33.23	36.42
42.10	42.10	47.78	54.33	

(1) *Determine* $\dfrac{n+1}{2} = \dfrac{10}{2} = 5$

(2) *The Median is located (table 5.2) at position 5 (from left)*

(3) *Median* $= 36.42\ kg/cm^2$

Summary:Mode

Solution 5.3

Mode: The BMI value 42.10 is repeated maximum number of times (table 5.2)
$Mode = 42.10\ kg/cm^2$

Measuring Variability

Quartiles

The mean and median provide two different measures of the center of a distribution.The simplest useful numerical description of a distribution requires both a measure of center and a measure of spread.The quartiles mark out the middle half. To calculate the quartiles:

(1) Arrange the observations in increasing order and locate the median M in the ordered list of observations.

(2) The first quartile Q1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.

(3) The third quartile Q3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

Inter Quartile Range

The distance between the quartiles (the range of the center half of the data) is a more resistant measure of spread. This distance is called the interquartile range.

Inter Quartile Range IQR and Quartile Deviation

The interquartile range IQR is the distance between the first and third quartiles,

$$IQR = Q_3 - Q_1$$

and $QD = \dfrac{Q_3 - Q_1}{2}$ is called Quartile deviation.

Illustration

Example 7.1

From the data shown in table (7.1) compute quartiles

Computations of descriptive statistics

Example 7.1

A Random sample of 9 patients with BMI values is given in following table (7.1), Compute descriptive statistics that you know.

Table 7.1: Random Sample and BMI of patient, $n = 9$

42.10	47.78	33.23	36.42	42.10
24.54	25.21	27.78	54.33	

Summary:Quartiles

Solution 7.1

(1) *Determine* $\dfrac{j(n+1)}{4} = \dfrac{10j}{4} = 2.5j, j = 1, 2, 3$

(2) *Find Integer part (I) and fraction part (f) from* $\dfrac{j(n+1)}{4}$

(3) *Use Formula* $Q_j = x_{(I)} + f \times (x_{(I+1)} - x_{(I)})$*, where* $x_{(i)}$ *are values given in table 5.2*

(4) *The Calculations:(See table 5.2)*

Summary:Quartiles

Solution 7.2

$$
\begin{aligned}
j = 1, \frac{j(n+1)}{4} &= 2.5 \\
I &= 2 \\
f &= 0.5 \\
Q_1 &= x_{(2)} + 0.5 \times (x_{(3)} - x_{(2)}) \\
&= 25.21 + 0.5 \times (27.78 - 25.21) \\
&= 26.495 kg/cm^2
\end{aligned}
$$

Summary:Quartiles

Solution 7.3

$$
\begin{aligned}
j = 2, \frac{j(n+1)}{4} &= 5 \\
I &= 5 \\
f &= 0 \\
Q_2 &= x_{(5)} + 0 \times \left(x_{(6)} - x_{(5)}\right) \\
&= 36.42kg/cm^2
\end{aligned}
$$

Summary:Quartiles

Solution 7.4

$$j = 3, \frac{j(n+1)}{4} = 7.5, I = 7, f = 0.5$$

$$\begin{aligned} Q_3 &= x_{(7)} + 0.5 \times (x_{(8)} - x_{(7)}) \\ &= 42.10 + 0.5 \times (47.78 - 42.10) \\ &= 44.94 kg/cm^2 \end{aligned}$$

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 44.94 - 26.495 \\ &= 18.445 kg/cm^2 \end{aligned}$$

$$QD = \frac{18.445}{2} = 9.2225 kg/cm^2$$

Summary:Range

Range

The simplest measure of dispersion of a variable is its range, which is defined as the difference between its highest and lowest given values.

Summary:Mean Deviation

Mean Deviation

If A is the chosen average value of the variable x, then $x_i - A$ is the deviation of the i^{th} given value of x from the average. Clearly the higher the deviations $x_1 - A, x_2 - A, \ldots, x_n - A$ in magnitude, the higher is the dispersion of x. The arithmetic mean of absolute deviations $|x_1 - A|, |x_2 - A|, \ldots, |x_n - A|$ may be taken as the measure of dispersion. It is referred to as the *mean deviation* of x *about* A. Denoting this mean deviation by MD_A, we have $\mathrm{MD}_A = \dfrac{\sum\limits_{i=1}^{n} |x_i - A|}{n}$.

Summary:Root Mean Square Deviation

Root Mean Square Deviation

If A is the chosen average value of the variable x, then $x_i - A$ is the deviation of the i^{th} given value of x from the average. Clearly the higher the deviations $x_1 - A, x_2 - A, \ldots, x_n - A$ in magnitude, the higher is the dispersion of x. By taking positive square root of the arithmetic mean of squares of the deviations $(x_i - A)^2$, i.e. $\sqrt{\dfrac{\sum\limits_{i=1}^{n}(x_i - A)^2}{n}}$ is called the *root-mean-square deviation* about A.

Summary:Standard Deviation

Standard Deviation

The measure of dispersion obtained by putting $\bar{x}$ for A above is called the standard deviation of x and is denoted by σ. We have therefore

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

For sample data we denote s or S_x, that is

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

Illustration

Example 8.1

From the data shown in table (7.1) compute mean deviation and standard deviation.

Summary:Mean Deviation

Solution 8.1

Table 8.1: Computation for Mean Deviation

x	$x-\bar{x}$	$\lvert x-\bar{x}\rvert$
42.10	5.0456	5.0456
47.78	10.7256	10.7256
33.23	-3.8244	3.8244
36.42	-0.6344	0.6344
42.10	5.0456	5.0456
24.54	-12.5144	12.5144
25.21	-11.8444	11.8444
27.78	-9.2744	9.2744
54.33	17.2756	17.2756
	Total	76.1844

Summary:Mean Deviation

Solution 8.2

$$MD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$
$$= \frac{76.1844}{9}$$
$$= 8.4649kg/cm^2$$

Summary:Standard Deviation

Solution 8.3

Table 8.2: Computation Standard deviation

x	$x-\bar{x}$	$(x-\bar{x})^2$
42.10	5.0456	25.4576
47.78	10.7256	115.0375
33.23	-3.8244	14.6264
36.42	-0.6344	0.4025
42.10	5.0456	25.4576
24.54	-12.5144	156.6113
25.21	-11.8444	140.2909
27.78	-9.2744	86.0153
54.33	17.2756	298.4448
	Total	862.3440

Summary:Standard Deviation

Solution 8.4

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$
$$= \sqrt{\frac{862.3440}{8}}$$
$$= \sqrt{107.7930}$$
$$= 10.38234 kg/cm^2$$

Comparison of theoretical with the observed frequency distribution

The theoretical frequency distributions in earlier problem were discrete. Their variables assumed values that changed in integral steps (that is, they were meristic variables). Thus, the number of infected insects per sample could be 0 or 1 or 2 but never an intermediate value between these. Similarly, the number of yeast cells per hemacytometer square is a meristic variable and requires a discrete probability function to describe it. However, many variables encountered in biology are continuous (such as the aphid femur lengths or the infant birth weights).

Comparison of theoretical with the observed frequency distribution

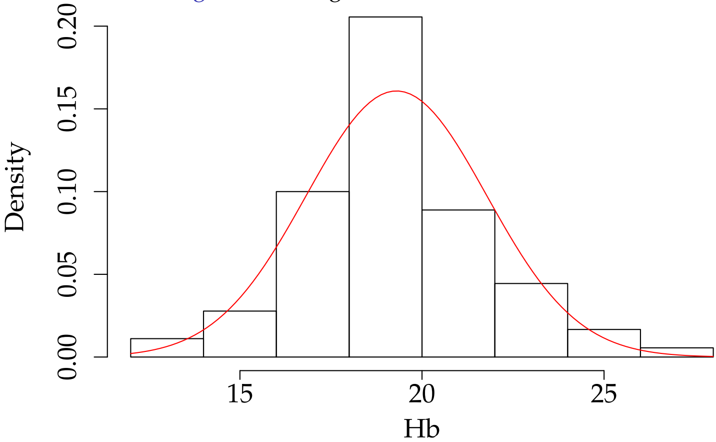
When you form a frequency distribution of observations of a continuous variable, your choice of class limits is arbitrary, because all values of a variable are theoretically possible. In a continuous distribution, one cannot evaluate the probability that the variable will be exactly equal to a given value such as 3 or 3.5. One can only estimate the frequency of observations falling between two limits.

Comparison of theoretical with the observed frequency distribution

Probability density functions are defined so that the expected frequency of observations between two class limits (vertical lines) is given by the area between these limits under the curve. The total area under the curve is therefore equal to the sum of the expected frequencies (1.0 or n, depending on whether relative or absolute expected frequencies have been calculated).(Table 3.3)

Comparison of theoretical with the observed frequency distribution

Figure 2: Histogram and Normal Curve Fit

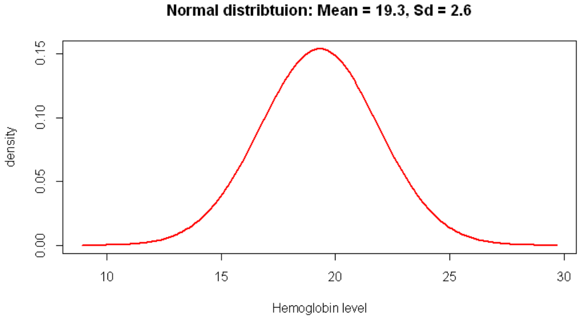


Comparison of theoretical with the observed frequency distribution

From Table (3.3) and Normal Curve (2) it may be observed that the shape of Histogram and shape of Normal Curve are approximately equal. Hence variable hemoglobin level of worker is distributed accordingly as Normal distribution.

Comparison of theoretical with the observed frequency distribution

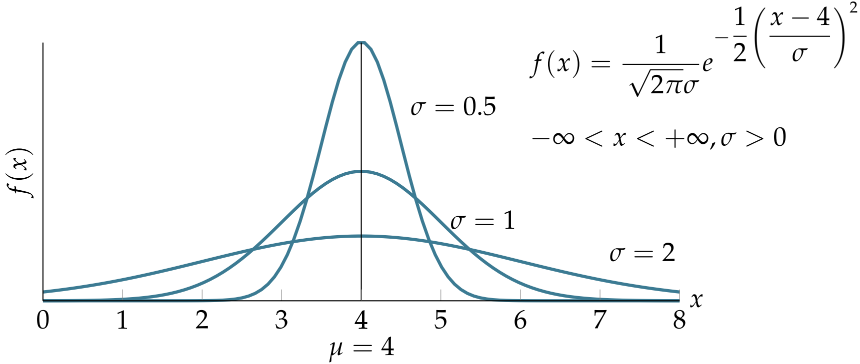
Figure 3: Normal Curve for Specified value of $\mu = 19.3$ and $\sigma = 2.6$



Probability Density Function

For continuous variables, the theoretical probability distribution, or probability density function, can be represented by a continuous curve, as shown in Figure

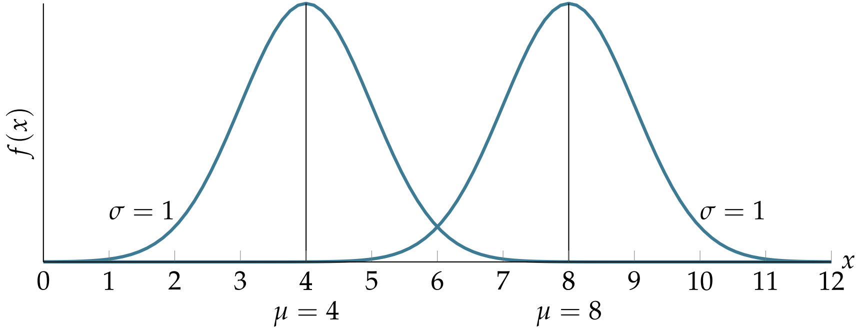
Figure 4: Normal density Curves



Probability Density Function

For continuous variables, the theoretical probability distribution, or probability density function, can be represented by a continuous curve, as shown in Figure

Figure 5: Normal density Curves



Normal Probability Distribution

Introduction

The normal distribution is expressed mathematically as

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x, \mu < +\infty, \sigma > 0 \qquad (9.1)$$

The graph of $f(x)$ is called Normal curve.The Normal curve is symmetrical about μ and the greater the value of σ the greater the spread of the curve, as shown in figure 4 and figure 5.

Mean and Variance of Normal Distribution

The variable X with function $f(x)$ (See 9.1) is called Normal Random variable.The mean and variance of random variable X is given by

$$\boxed{\begin{aligned} \text{Mean} &= \mu \\ \text{Variance} &= \sigma^2 \end{aligned} \qquad (9.2)}$$

We can write symbolically $X \sim N(\mu, \sigma^2)$ to denote random variable X follows Normal Probability distribution with parameters μ and σ^2

Standard Normal Probability Distribution

Standardized Variable

By taking $Z = \dfrac{X - \mu}{\sigma}$ in formula of Normal distribution, the function $f(X)$ in (9.1) is transformed to

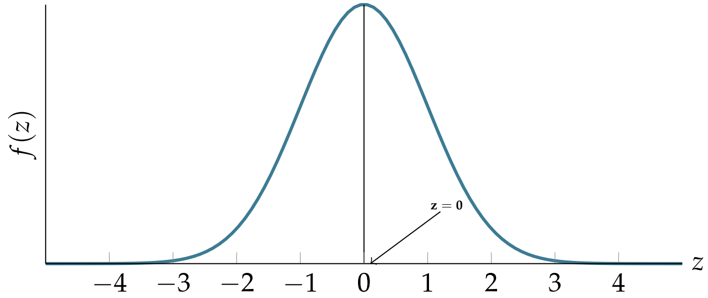
$$\boxed{f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2} \quad -\infty < z < +\infty \qquad (10.1)}$$

The variable Z is called standardized random variable. The mean of random variable Z is 0 and variance is 1 and we can write $Z \sim N(0, 1)$

Standard Normal Probability Density Function

The Graph of $f(z)$ is called Standard Normal Curve. Given below

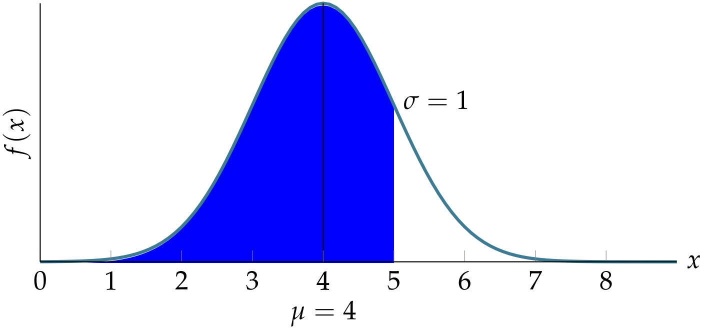
Figure 6: Standard Normal density Curve



The variable Z is called standardized random variable. The mean of random variable Z is 0 and variance is 1 and we can write $Z \sim N(0, 1)$.Note that The curve is symmetrical around the vertical line where $z = 0$.

Area under the curve

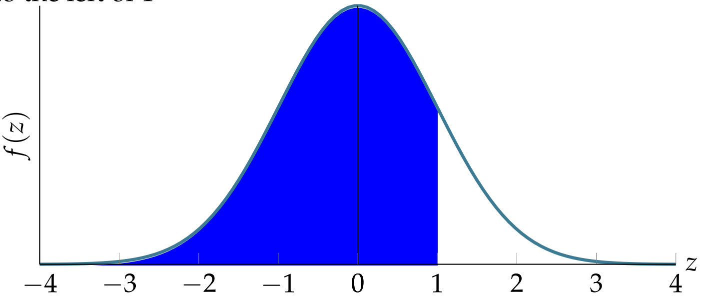
Cumulative Probability $P(X \leq 5)$:Area under the normal curve to the left of 5



By taking transformation $z = \dfrac{5-4}{1}$, we get

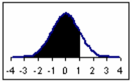
Area under the curve

Cumulative Probability $P(Z \leq 1)$:Area under the standard normal curve to the left of 1



Area under the curve

We note that $P(X \leq 5) = P(Z \leq 1) = 0.8413$ From Statistical Table.The cross of row at 1.0 and column at 0 (See below)



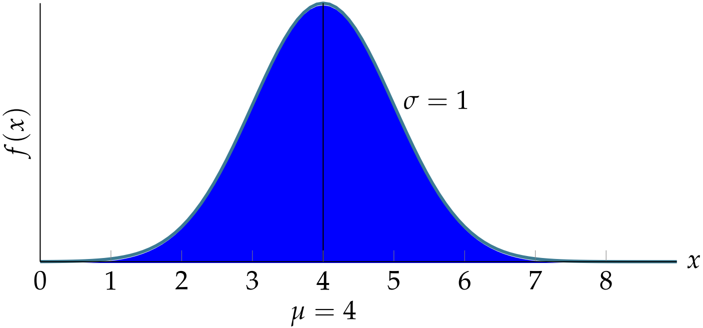
Distribution Function of Standard Normal Random Variable **591**

Table 3 *Distribution Function of Standard Normal Random Variable*

z	0	1	2	3	4	5	6	7	8	9
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5363	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7974	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015

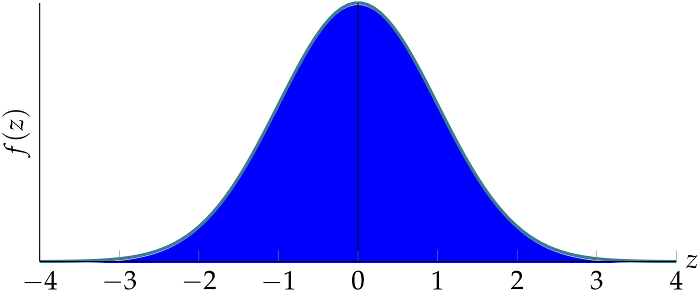
Area under the curve

Total Probability $P(-\infty < X < +\infty) = 1$:
The total area under the normal curve is always equal to 1



Area under the curve

Total Probability $P(-\infty < Z < +\infty) = 1$:
The total area under the standard normal curve is always equal to 1



Determining Area under the curve

Example 10.1

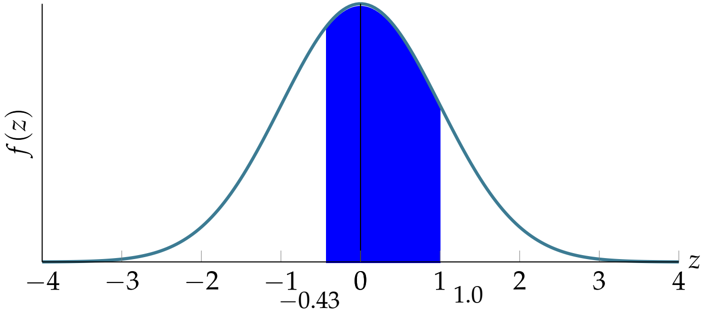
A large number of determinations was carried out on the same sample and the results are known to be normally distributed with $\mu = 215$ *and* $\sigma = 35$ *What percentage of determinations will fall between the boundaries 200 and 250?*

Solution 10.1

First we compute Z values, $z_1 = \dfrac{200 - 215}{35} = -0.43$ *and* $z_2 = \dfrac{250 - 215}{35} = 1.00$

The Area under the curve is shown in figure

Determining Area under the curve



The required probability is
$P(200 \le X \le 250) = P(-0.43 \le Z \le 1.00) = 0.5077$
=Area under the standard normal curve between -0.43 and 1.00
We can conclude that 51% of all data are comprised between 200 and 250.Note that this area is calculated using **Normal Probability Tables**.

Exercise:Computing Summary Statistics

Example 10.2

A person's metabolic rate is the rate at which the body consumes energy. Metabolic rate is important in studies of weight gain, dieting, and exercise. Here are the metabolic rates of 7 men who took part in a study of dieting. (The units are calories per 24 hours. These are the same calories used to describe the energy content of foods.

Table 10.1: Metabolic rate

1792	*1666*	*1362*	*1614*	*1460*	*1867*	*1439*

Exercise:Computing Summary Statistics

Example 10.3

The level of various substances in the blood influences our health. Here are measurements of the level of phosphate in the blood of a patient, in milligrams of phosphate per deciliter of blood, made on 6 consecutive visits to a clinic:

Table 10.2: Phosphate level in blood mg/dl

5.6	5.2	4.6	4.9	5.7	6.4