20

CHAPTER

Probability and Sampling Distribution

Learning Objectives

After the completion of this chapter, you will be able to

- Explain how to generate random sample for an experiment
- Explain how to generate random sample case-wise or according to percentage
- Understand the concept of probability distribution
- Explain the concept of discrete probability distribution including Binomial and Poisson probabilities
- Demonstrate the steps used in SPSS to execute Binomial and Poisson probability distributions
- Display Binomial and Poisson distribution graphically
- Explain the concept of normal density function for continuous random variables
- Demonstrate the steps used in SPSS to execute normal probabilities for normal distributed variables
- Explain the concept of sampling distribution
- Demonstrate the simulation of desired sample through SPSS Syntax programming language
- Use SPSS in simulating data set
- Describe the shape and statistics of sampling distribution derived from simulated data

20.1 Introduction

This chapter starts with explaining how to generate random sample for making inferences in the study. The random sample can be generated either for a particular experiment or in the existing population elements. The chapter also highlights about probability distributions and sampling distribution. The probability distributions are manifested by mathematical functions indicating the probabilities of occurrence of all possible outcomes in a particular experiment. Primarily, the probability distributions are categorized into discrete and continuous based on the nature of random variable as incorporated in the analysis. In this view, the discrete variable reflects the outcomes in the form of discrete (such as tossing a coin or rolling a dice), whereas the continuous variables indicate the outcomes in real number (such as height, weight or temperature of day). The probability distribution of discrete and continuous variables is explained by the *probability mass function* and *probability density function*, respectively. The binomial probability distribution is used for discrete random variable, whereas continuous random variable is explained by Poisson distribution. The chapter also focuses on the application of sampling distribution in order to make inferences about unknown population parameters.

20.2 Generating a Random Sample

Generating a random sample from SPSS is an important application. Sometimes a random data is generated based on the range of pre-defined outcomes for a specific experiment or the data can be generated randomly belonging to the specific number of subjects/cases. The cases can also be drawn randomly from a particular sampling frame of a population elements' list. An outcome and execution of experiment are two important phenomena in generating the random sample. For instance, tossing a coin is an experiment resulting into *heads* and *tails* as two possible outcomes. In this section, we shall discuss how to generate a random sample belonging to fixed number of outcomes of an experiment and also selecting the cases randomly from pre-existing data set.

20.2.1 Generating Random Sample for an Experiment

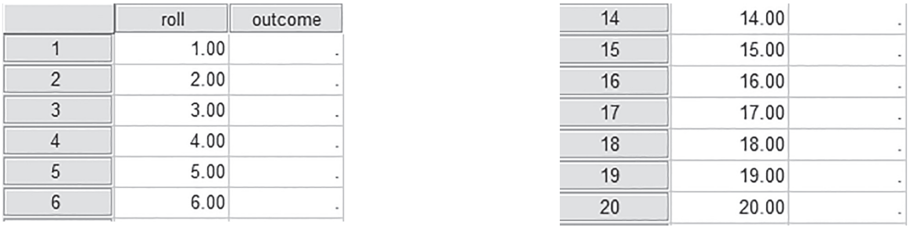
In this section, we will study how SPSS could be used to generate a random sample revealing an outcome of an experiment. For example, rolling a dice generates six possible outcomes (1, 2, 3, 4, 5 and 6) with equal probability of occurrence. Now, we highlight how to execute random numbers of these outcomes by rolling a dice for 20 times.

1. **Generate variables, data and missing values:** The first task is to generate the desired variables according to the outcome of an experiment. We open blank SPSS Data Editor and create two variables—*roll* and *outcome*. In data view, we typed 1 to 20 for *roll*, whereas the data for *outcomes* remains empty and considered as missing value at this point of time. The following steps are used for the same as mentioned in Exhibit 20.1.

Exhibit 20.1. Open blank SPSS Data Editor » Create two variables—roll and outcome » Type 1 to 20 for roll in data view » Outcome remains empty » Consider as missing values » Save as random

The data view appears on the screen as shown in Figure 20.1 by using the steps as mentioned in Exhibit 20.1.

Figure 20.1 Data View along with Missing Values: Random Number



	roll	outcome
1	1.00	.
2	2.00	.
3	3.00	.
4	4.00	.
5	5.00	.
6	6.00	.
14	14.00	.
15	15.00	.
16	16.00	.
17	17.00	.
18	18.00	.
19	19.00	.
20	20.00	.

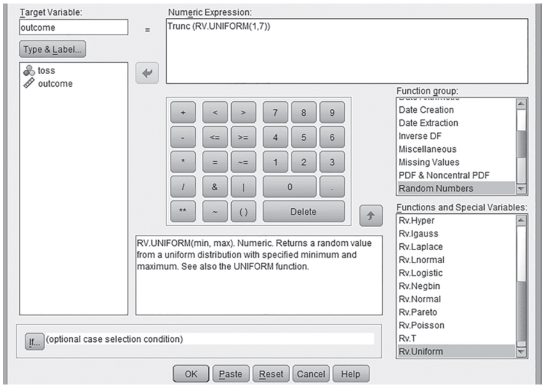
Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

2. **Generate outcome:** Next step is to create an outcome for specific experiment by using the following steps in continuity of Exhibit 20.1. We use RV (random variable) function for the same. This function RV. UNIFORM (min, max) is used to create random values for a uniform probability distribution. We also use the TRUNC function along with RV. This function is basically used to truncate the number for rounding them to the nearest integer. In our case, as minimum and maximum values for a dice are 1 and 6, respectively, we mention (1, 7) for generating the numbers higher than 1 and lower than 6 along with rounding the same. The steps used in SPSS to generate the outcome are shown in Exhibit 20.2.

Exhibit 20.2. Use data set random » Transform » Compute variable » Type outcome in Target variable box » Select and click on Random Numbers » Select RV. Uniform in Functions and Special Variables box » Type 1, 6 in the numerical expression box » Type Trunc in front of RV. Uniform as Trunc (RV.UNIFORM(1,7)) » Press OK » Change existing variable » Press OK

The dialog box appears on the screen as shown in Figure 20.2.

Figure 20.2 Main Dialog Box: Compute Variable



Target Variable: outcome

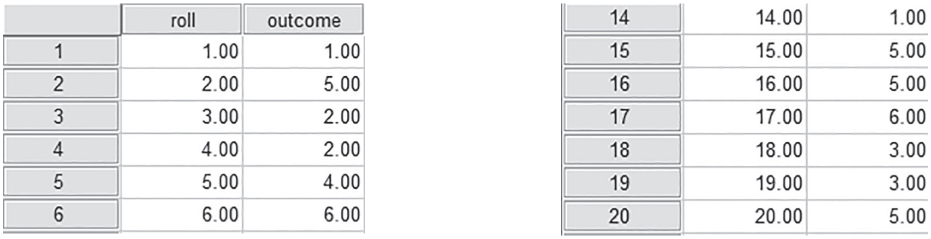
Numeric Expression: Trunc (RV.UNIFORM(1,7))

RV.UNIFORM(min, max). Numeric. Returns a random value from a uniform distribution with specified minimum and maximum. See also the UNIFORM function.

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

3. **Random number output:** The SPSS output viewer creates random values for all possible outcomes of dice ranging from 1 to 6 for 20 cases. These random values are generated based on the uniform probability distribution. The same is shown in Figure 20.3.

Figure 20.3 Random Values of Outcomes



	roll	outcome
1	1.00	1.00
2	2.00	5.00
3	3.00	2.00
4	4.00	2.00
5	5.00	4.00
6	6.00	6.00
14	14.00	1.00
15	15.00	5.00
16	16.00	5.00
17	17.00	6.00
18	18.00	3.00
19	19.00	3.00
20	20.00	5.00

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

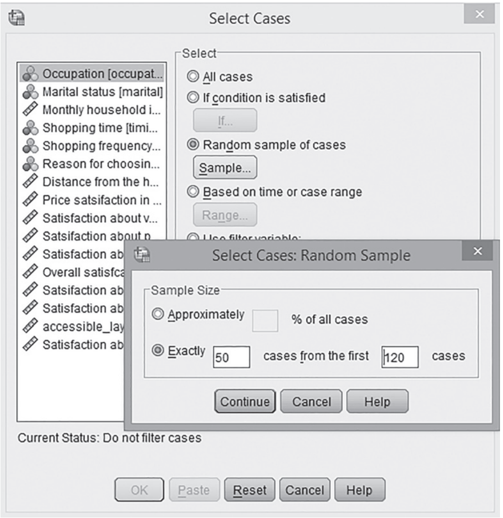
20.2.2 Generating Random Sample Case-wise

SPSS can also be used to generate the random sample from a sampling frame. It is the most basic method of collecting a sample from a target population. The sample is selected by SPSS either as percentage or number of cases as per the specific requirement (Gordon and Johnson 2005). We use data set retail.sav for selecting a random sample of 50 cases out of total 120 elements. The steps mentioned in Exhibit 20.3 are used to create the same.

Exhibit 20.3. Use data set retail.sav » Data » Select cases » Select random sample of cases » Click on sample » Select exactly and type 50 and 120 in the given options » Continue » Press OK

The dialog box appears on the screen as shown in Figure 20.4.

Figure 20.4 Main Dialog Box: Random Selection of Cases



Select Cases

Occupation [occupat...
Marital status [marital]
Monthly household i...
Shopping time [timi...
Shopping frequency...
Reason for choosin...
Distance from the h...
Price satsifaction in ...
Satisfaction about v...
Satsifaction about p...
Satisfaction ab...
Overall satisfca...
Satsifaction ab...
Satisfaction ab...
accessible_lay...
Satisfaction ab...

Select
○ All cases
○ If condition is satisfied
● Random sample of cases
○ Based on time or case range
○ Use filter variable:

Select Cases: Random Sample

Sample Size
○ Approximately □ % of all cases
● Exactly 50 cases from the first 120 cases

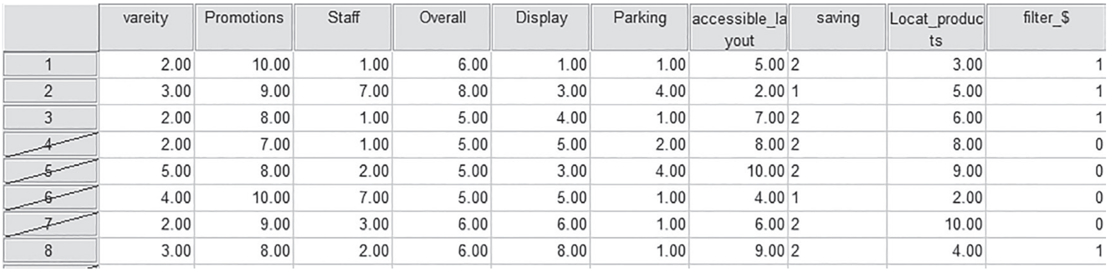
Current Status: Do not filter cases

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

1. **Simple random sample output:** Total 50 cases are selected randomly out of 120 on the basis of simple random method. The selected cases can be easily identified in SPSS Data Editor. The cases that are discarded are crossed diagonally by software. In the last column, one new variable filter_$ is created by SPSS

having values 0 and 1 indicating those cases that are excluded and remained in the data set, respectively. Hence, with this procedure the cases 1, 3, 7, 16, 17, …, 115, 116 and 119 are included in the revised data set. It is important to note that every time the result of random sample output is different. Thus, by following this procedure, you may have different results as compared to shown in Figure 20.5.

Figure 20.5 Random Selection of Cases



	vareity	Promotions	Staff	Overall	Display	Parking	accessible_layout	saving	Locat_products	filter_$
1	2.00	10.00	1.00	6.00	1.00	1.00	5.00	2	3.00	1
2	3.00	9.00	7.00	8.00	3.00	4.00	2.00	1	5.00	1
3	2.00	8.00	1.00	5.00	4.00	1.00	7.00	2	6.00	1
~~4~~	2.00	7.00	1.00	5.00	5.00	2.00	8.00	2	8.00	0
~~5~~	5.00	8.00	2.00	5.00	3.00	4.00	10.00	2	9.00	0
~~6~~	4.00	10.00	7.00	5.00	5.00	1.00	4.00	1	2.00	0
~~7~~	2.00	9.00	3.00	6.00	6.00	1.00	6.00	2	10.00	0
8	3.00	8.00	2.00	6.00	8.00	1.00	9.00	2	4.00	1

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

20.2.3 Generating Random Sample Percentage-wise

SPSS can also be used to generate the random sample percentage-wise. We use the same data set retail.sav for selecting a random sample of 20 per cent cases out of total 120 elements. In continuity of the steps as mentioned in Exhibit 20.3, we type 20 in the percentage box of *select case* area and find out the results similar to Figure 20.4 and consisting approximately 20 per cent of cases.

20.3 Probability Distribution

The probability distributions are categorized as either discrete or continuous on the basis of whether using discrete or continuous variables in the analysis. Discrete variable is countable in nature, whereas the continuous is always in the measurable form. For instance, tossing a coin or rolling of dice is a discrete variable resulting into integer values only. The values of random variable in case of rolling a dice could be only 1, 2, 3, 4, 5 and 6; hence, it is discrete random variable. Whereas, the temperature of last 5 days is considered as continuous variable as it consists of any measurable values. A discrete distribution reveals the probability of occurrence of all the possible values of discrete random variables in tabular form or in specific shape. The discrete probability distributions are further categorized as binomial, Poisson, multinomial and Bernoulli. The following sections discuss these probability distributions in detail.

20.4 Binomial Distribution

The binomial distribution is discrete probability distribution. It is used to summarize the independent observations revealing one of two outcomes under certain assumptions or parameters (Weisstein 2019). In this probability distribution, the number of observations is fixed as well as the probability of success is also fixed for each trial. The observations are independent to each other and each observation is categorized into *success* (1) or *failure* (0). Hence, the binomial distribution represents the probability for x successes for n trials. This probability distribution of this random variable x is called binomial distribution and is presented by the formula:

$$P(X) = C_x^n \cdot p^x\, q^{n-x}$$

where

n = number of trials,
p = probability of success,
q = probability of failure (1–p),
$P(X)$ = probability of x successes in n binomial trials.
C_x^n = combination and expressed as:

$$C_x^n = \frac{n!}{X!(n-X)!}$$

Hence, binomial distribution can be denoted by the following formula.

$$P(X) = \frac{n!}{X!(n-X)!} \cdot (p)^x (q)^{n-x}$$

20.4.1 Executing Binomial Distribution with SPSS

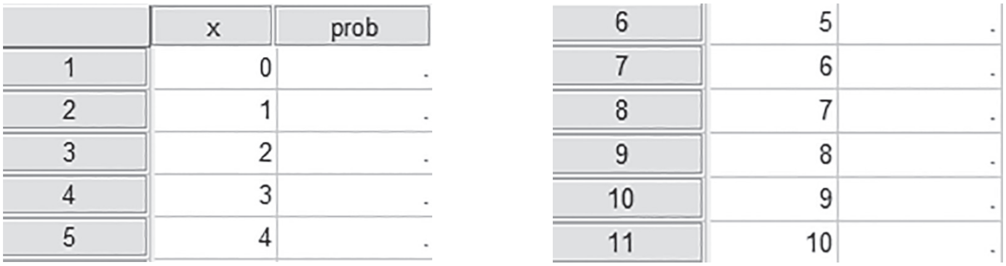
SPSS is used to compute binomial distribution for a specified probability of success (p) for certain values of x successes either through probability density function (PDF) or cumulative density function (CDF). The result of these probability functions appear in a separate column of SPSS Data Editor. The binominal probability distribution is calculated with parameters n and p and denoted as $P\,(x|n,\, p)$, here x represents the number of successes in n trials.

1. **Create data set in SPSS Data Editor:** In order to comprehend binomial distribution, we take an example to find out the distribution associated to 10 women professionals in an organization who might discontinue their job after the birth of their first baby. In this case, we suppose having 10 trials (n) with probability of success (p = do not discontinue the job) 0.40. The steps used to obtain the binomial distribution of random variable x are mentioned in Exhibit 20.4.

Exhibit 20.4. Open SPSS Data Editor » Create two variables » x and prob » Assign values 0, 1, 2, 3, ..., 10 for x variable » Leave blank for prob variable

The SPSS Data Editor appears on the screen as shown in Figure 20.6.

Figure 20.6 SPSS Data Editor: Binomial Variable



	x	prob
1	0	.
2	1	.
3	2	.
4	3	.
5	4	.
6	5	.
7	6	.
8	7	.
9	8	.
10	9	.
11	10	.

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

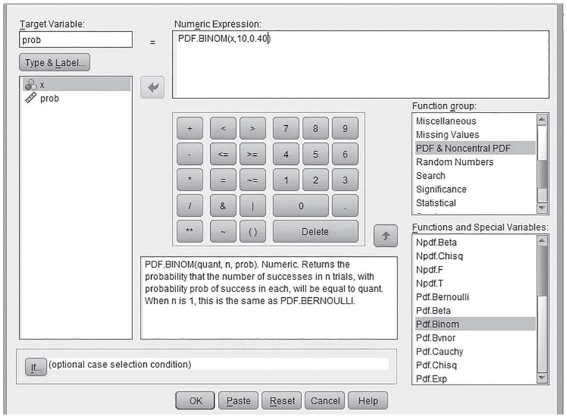
2. **Compute probability density for binomial variable:** We use the function PDF. Binom (quant, n. prob) for the same. The quant represents the number of successes, *n* is the total number of trials and prob is the success probability. The steps with continuity of Exhibit 20.4 are used to compute the PDF for *x* variable as shown in Exhibit 20.5.

Exhibit 20.5. Transform » Compute variable » Type prob in Target variable box » Select and click on PDF & Noncentral PDF in Functions group » Select PDF.Binom in Functions and Special Variables Box and double click » Select x In Type & Label box and transfer into numeric expression box, type 10, and 0.40 in place of ?, ? » Press OK » Change existing variable » Press OK

The following dialog box appears on the screen as shown in Figure 20.7.

The probabilities of *n* trials are shown in SPSS Data Editor under the *prob* column by using the steps mentioned in Exhibit 20.5. SPSS Data Editor consists of the binomial distribution for given case as shown in Figure 20.8.

Figure 20.7 Main Dialog Box: Probability Density Function



Target Variable: prob

Numeric Expression: PDF.BINOM(x,10,0.40)

x
prob

Function group: Miscellaneous, Missing Values, PDF & Noncentral PDF, Random Numbers, Search, Significance, Statistical

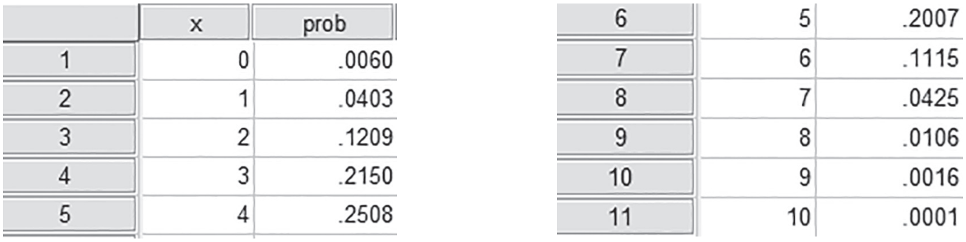
Functions and Special Variables: Npdf.Beta, Npdf.Chisq, Npdf.F, Npdf.T, Pdf.Bernoulli, Pdf.Beta, Pdf.Binom, Pdf.Bvnor, Pdf.Cauchy, Pdf.Chisq, Pdf.Exp

PDF.BINOM(quant, n, prob). Numeric. Returns the probability that the number of successes in n trials, with probability prob of success in each, will be equal to quant. When n is 1, this is the same as PDF.BERNOULLI.

If... (optional case selection condition)

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

Figure 20.8 Binomial Probability Distribution



	x	prob
1	0	.0060
2	1	.0403
3	2	.1209
4	3	.2150
5	4	.2508
6	5	.2007
7	6	.1115
8	7	.0425
9	8	.0106
10	9	.0016
11	10	.0001

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

The column *prob* represents the binomial probability distribution for *x*-values. We can observe that there is .0060 or 0.6 per cent (0.006 × 100) probability, that no women discontinue job after the birth of the first baby. Similarly, there is 12% probability that two women staff discontinue their job in the given condition. Likewise, there is almost no probability (0.01%) for discontinuing the job by all the women staff after the birth of the first baby. We can find out the same probability by using the following formulae by considering $x = 2$.

$$P(X) = \frac{n!}{X!(n-X)!}(p)^x\,(q)^{n-x}$$

$$P(2) = \frac{10!}{2!(10-2)!}(0.40)^{10}(0.60)^{8}$$

$$= \frac{3,628,800}{2(8)!}(0.16)\cdot(0.016)$$

$$= \frac{3,628,800}{80,640}\cdot(0.16)\cdot(0.016)$$

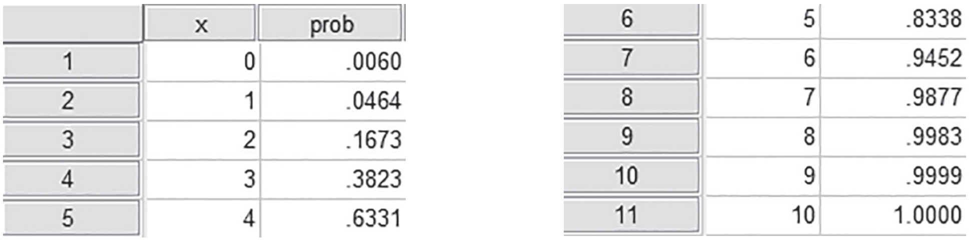
Thus, the probability for $x = 2$, 0.1206 ≈ 12%.

3. **Compute commutative probability density for binomial variable:** SPSS also computes cumulative probability for discrete random variable by using the function CDF.Binom (quant, n, prob). The CDP function computes cumulative probability for each successive trial and the probability of each success is less than the total probability. The steps used to create the same are shown in Exhibit 20.6.

Exhibit 20.6. Transform » Compute variable » Type prob in Target variable box » Select and click on CDF & Noncentral CDF in Functions group » Select CDF.Binom in Functions and Special Variables Box and double click » Select x In Type & Label box and transferred into numeric expression box, type 10, and 0.40 in place of ?, ? » Press OK » Change existing variable » Press OK

The results appear on the screen as shown in Figure 20.9.

Figure 20.9 Cumulative Binomial Probability Distribution



	x	prob
1	0	.0060
2	1	.0464
3	2	.1673
4	3	.3823
5	4	.6331
6	5	.8338
7	6	.9452
8	7	.9877
9	8	.9983
10	9	.9999
11	10	1.0000

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

4. **Displaying PDF and cumulative distribution:** Both distributions are displayed with simple bar chart as shown in Figures 20.10a and 20.10b. Follow the steps of preparing simple bar chart as shown in Chapter 3 in order to display the relative probability for each success for density function and cumulative distribution.

Figure 20.10a Binomial Probability Density Distribution

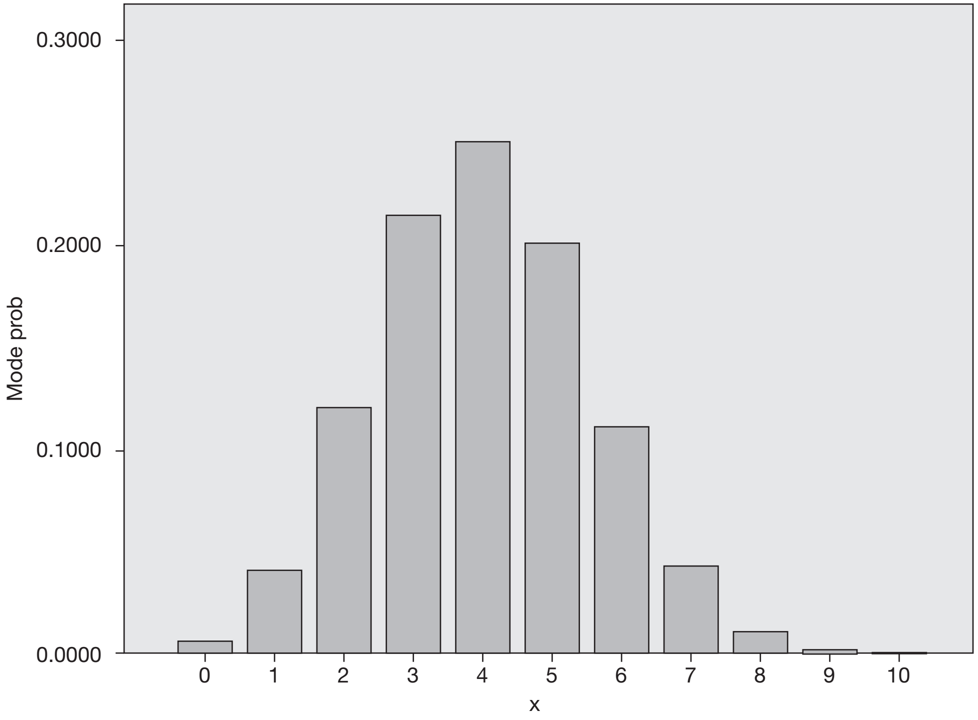
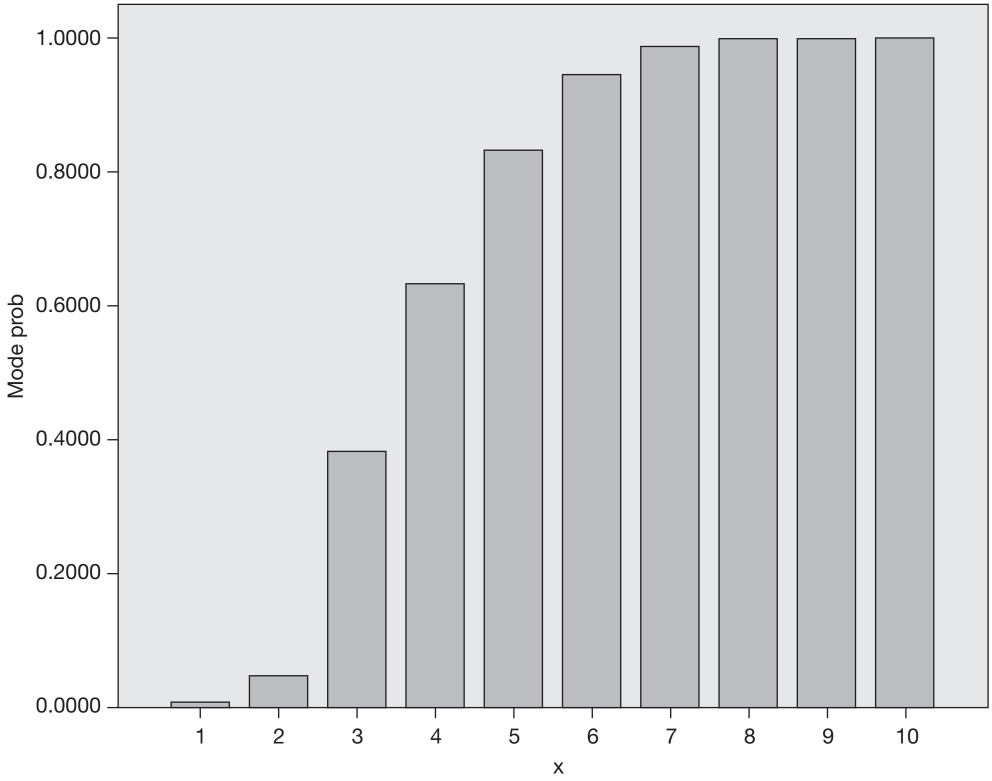


Figure 20.10b Cumulative Distribution: Binomial Probability



Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

20.4.2 Mean and Variance of Binomial Distribution

The mean of binomial distribution (μ_x) reveals the average value of successes in a particular experiment based on *n* number of trials. The following formula is used to compute the mean and variance for binomial distribution:

$$\mu_x = n \times p$$

Thus, in the previous example, the $\mu_x = 10 \times 0.40 = 4$. One would expect that mean of women discontinuing the job is 4 among 10.

The formula for variance is

$$\mu_x^2 = np(1 - p)$$

Thus, the variance is 4 (1–0.40) = 2.4.

20.5 Poisson Distribution

The Poisson distribution indicates the occurrence of events within a specific region. This region could be a certain time period, length, area or volume. The events occurring in the region can also be classified as success or failure. Similar to binomial distribution, the number of success in the region is also known in order to employ the Poisson distribution. For instance, the Poisson distribution is used in modelling the problems such as the number of leakage packets occurring during packaging in a certain time interval, arrivals of heavy vehicles at toll plaza per day and number of phone calls received for promotions per hour. The interval of these regions can also be divided into sub-intervals. These sub-intervals are independent of each other and proportionate to the length of the region. In this distribution, the random variable (*x*) reveals the number of success in the whole intervals or in region and the probability distribution of this random variable is called Poisson distribution (Joseph and Reinhold 2003). Only one parameter named Lambda (λ) is used to compute the Poisson distribution. The λ represents the mean of successes in the interval. We can compute this probability distribution for certain number of λ at given region by using the following formula:

$$P(x;\lambda) = \frac{e^{-\lambda}\mathrm{x}\ \lambda^{\mathrm{x}}}{x!}$$

where

λ = mean number of successes in a given region,
e = Euler's constant (2.71828),
x = random variable denoting number of successes,
P(*x*) = probability of Poisson distribution.

20.5.1 Executing Poisson Distribution with SPSS

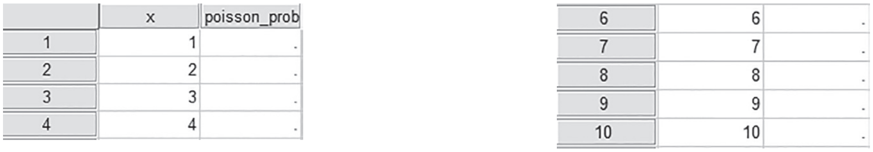
The SPSS computes probability of given number of events occurring in a fixed interval of region as Poisson distribution or in cumulative distribution. Similar to binomial distribution, the result of this distribution also appears in a separate column at SPSS Data Editor.

1. **Creating data set for Poisson distribution:** Similar to binomial distribution, the first step in SPSS is to create an appropriate data set in Data Editor as per the given situation. In order to understand the same, we take an example such as in a super specialty hospital 20 road accident cases arrived in a week. So, we want to ascertain the probability for occurrence of 1, 2, 3, 4, 5, …, 10 accidental cases per day. In this case, the number of accidents is considered as random variable (*x*), those values range from 1 to 10. In SPSS, first we create two variables—*x* as random variable and second *possion_prob* in Data Editor by using the steps mentioned in Exhibit 20.7.

Exhibit 20.7. Open SPSS Data Editor » Create two variables » *x* and *poisson_prob* » Assign values 0, 1, 2, 3, …, 10 for *x* variable » Leave blank for *poisson_prob* variable

The SPSS Data Editor appears on the screen as shown in Figure 20.11.

Figure 20.11 SPSS Data Editor: Random Variable *x* for Poisson Probability



	x	poisson_prob
1	1	.
2	2	.
3	3	.
4	4	.
6	6	.
7	7	.
8	8	.
9	9	.
10	10	.

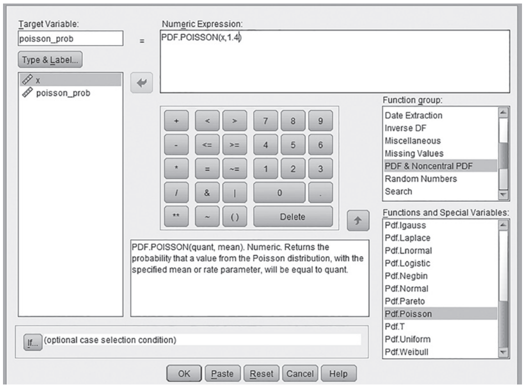
Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

2. Compute probability density for random variable: The function PDF.Poisson (quant, mean) is used to compute Poisson probability. The quant represents the number of successes (occurrence of cases per day) and mean indicates the average number of successes in a specific region. The mean (λ) is computed as 1.4 (λ = total number of accidents/number of days, 20/7 = 1.4). The steps with continuity of Exhibit 20.7 are used to compute the PDF for *x* variable as shown in Exhibit 20.8.

> **Exhibit 20.8.** Transform » Compute variable » Type poisson_prob in Target variable box » Select and click on PDF & Noncentral PDF in Functions group » Select PDF.Poisson in Functions and Special Variables Box and double click » Select x In Type & Label box and transfer into numeric expression box, type 1.4 in place of ?, ? » Press OK » Change existing variable » Press OK

The following dialog box appears on the screen as shown Figure 20.12.

Figure 20.12 Main Dialog Box: Poisson Probability



Target Variable: poisson_prob

Numeric Expression: PDF.POISSON(x,1.4)

Function group: PDF & Noncentral PDF

Functions and Special Variables: Pdf.Poisson

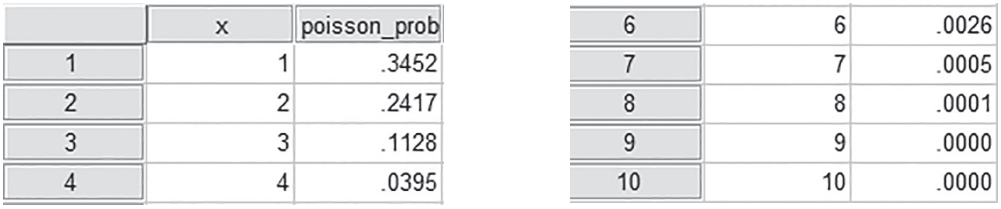
PDF.POISSON(quant, mean). Numeric. Returns the probability that a value from the Poisson distribution, with the specified mean or rate parameter, will be equal to quant.

If... (optional case selection condition)

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

By using the steps mentioned in Exhibit 20.8, the Poisson probabilities of *random variable* (x) are shown in SPSS Data Editor under the poisson_*prob* column. The same is as shown in Figure 20.13.

Figure 20.13 Poisson Probability Distribution



	x	poisson_prob
1	1	.3452
2	2	.2417
3	3	.1128
4	4	.0395
6	6	.0026
7	7	.0005
8	8	.0001
9	9	.0000
10	10	.0000

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

The column *poisson_prob* represents the Poisson probability distribution for *x*-values. We can predict that there is 0.3452 or 34 per cent (0.3452 × 100) probability of arriving 1 accidental case per day in the hospital. Similarly, there is 0.26% (less than 1%) probability of arriving 6 cases per day. This probability can also be computed by using the following formula:

$$P(x;\lambda) = \frac{e^{-\lambda} \times \lambda^{x}}{x!}$$

Hence, the probability of arriving 6 cases per day would be:

$$P(6;1.4) = \frac{2.71828^{-1.4} \times 1.4^{6}}{6!}$$
$$= \frac{0.2465 \times 7.52}{720}$$

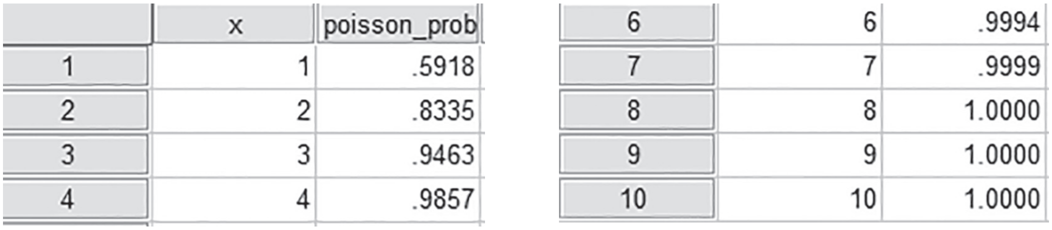
Thus, $P(6;\ 1.4) = 0.0025 \approx 0.26\%$.

3. **Compute commutative probability:** The cumulative probability for discrete random variable (*x*) can be computed by using the function CDF.Poisson (quant, mean). The CDP function computes cumulative probability for each successive number and the probability of each success is less than the total probability as found in binomial distribution.
The steps used to create the same are shown in Exhibit 20.9.

Exhibit 20.9. Transform » Compute variable » Type prob in Target variable box » Select and click on CDF & Noncentral CDF in Functions group » Select CDF.Poisson in Functions and Special Variables Box and double click » Select x In Type & Label box and transfer into numeric expression box, type 1.4 in place of ?, ? » Press OK » Change existing variable » Press OK

The results appear on the screen as shown in Figure 20.14.

Figure 20.14 Cumulative Poisson Probability Distribution



	x	poisson_prob
1	1	.5918
2	2	.8335
3	3	.9463
4	4	.9857
6	6	.9994
7	7	.9999
8	8	1.0000
9	9	1.0000
10	10	1.0000

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

4. **Displaying PDF and cumulative distribution:** Similar to binomial distribution, both probabilities are displayed with simple bar chart as shown in Figures 20.15a and 20.15b.

Figure 20.15a Probability Density: Poisson

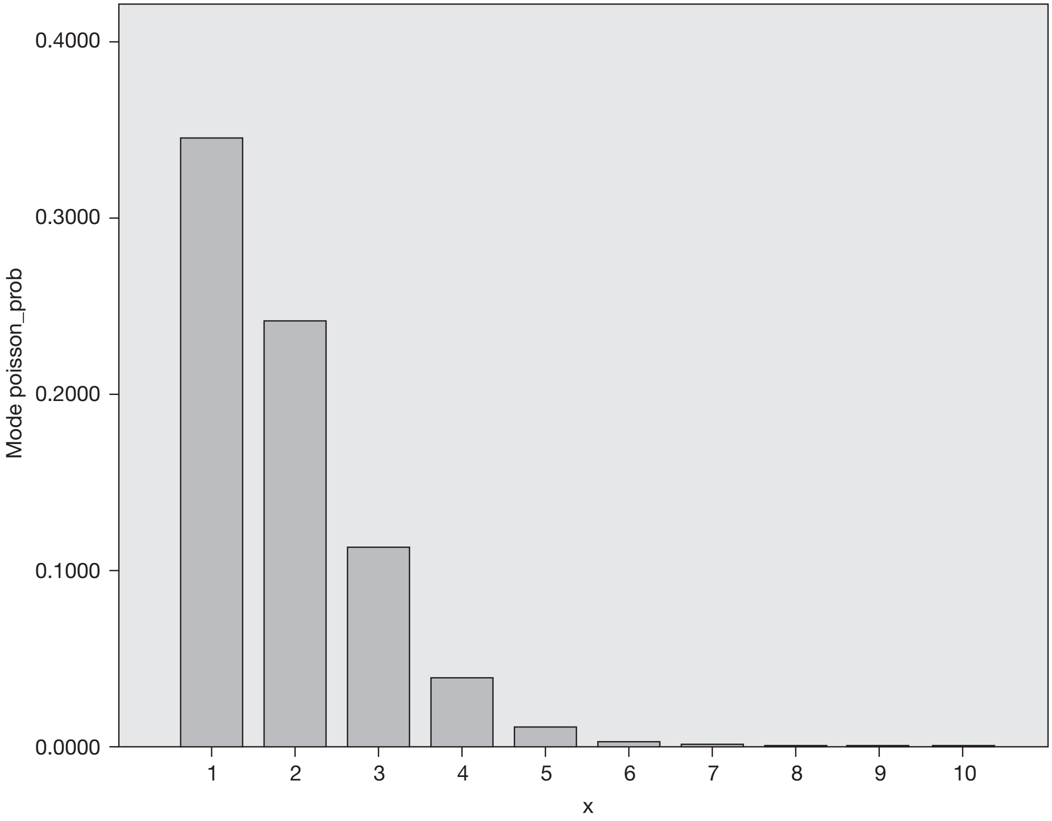
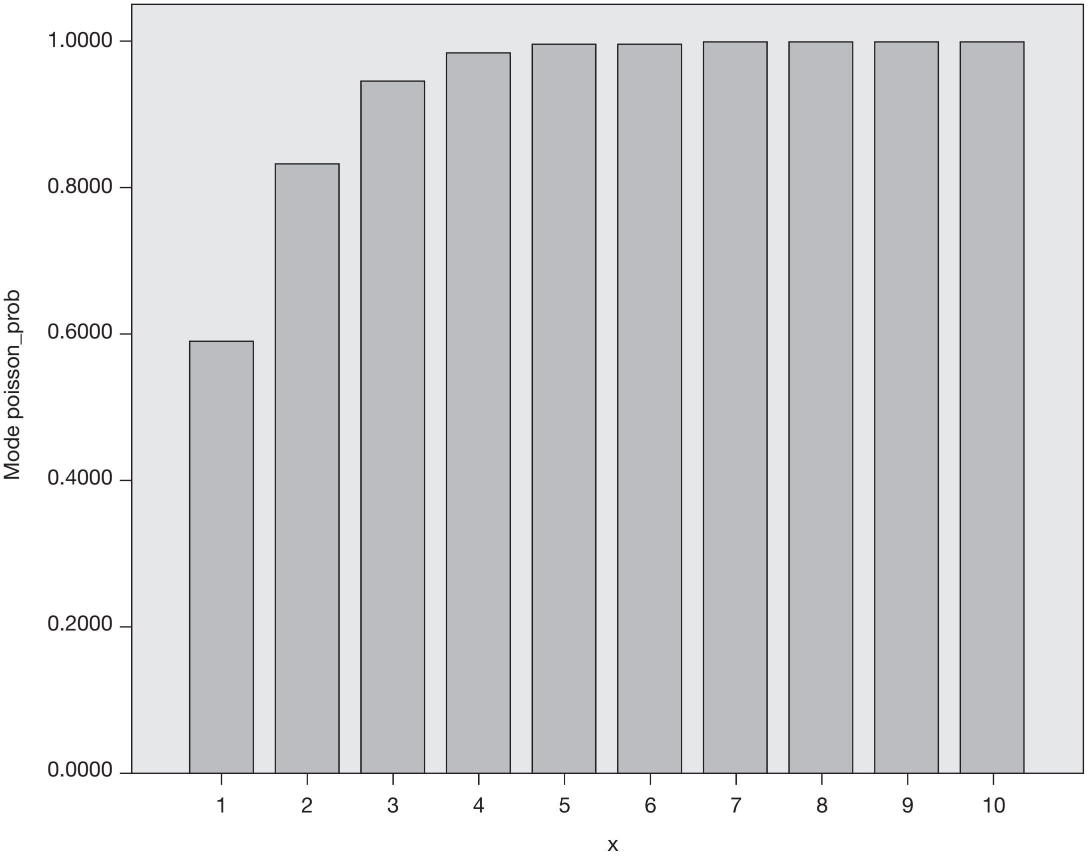


Figure 20.15b Cumulative Distribution: Poisson



Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

20.6 Normal Probabilities

The binomial and Poisson distributions for random variable as we discussed so far are discrete in nature. Till this point, we studied how to compute the respective probabilities of each or specific value of random variable through Sections 20.4 and 20.5. Now, we extend our discussion for computing the probabilities of continuous random variable. Unlike the value (such as 0, 1, 2, 3, …) of discrete random variable, the continuous random variable has infinite number of all possible values such as infinite values between the weighs 50 pounds and 60 pounds. In this scenario, we can compute a distinct probability of each possible value of *x*. Thus, rather than using *probability function*, we use *probability density function*. In this way, the probability is computed for a certain interval of values or specific range of variable *x*. In such cases, the probability is defined under the area of *density function* or under a *curve*. Overall area under the curve is 1 and has no negative values. The most popular and used *density curve* to ascertain the probabilities of continuous random variable is normal distribution curve. Hence, entire discussion in the following section is based on the random variable(s) having property of normal distribution with two parameters mean (μ) and standard deviation (σ). Suppose, if x is considered as normal distribution with 10 and 3 as mean and standard deviation, then it can be denoted as $x \sim$ N (10,3).

20.6.1 Computing Normal Probabilities

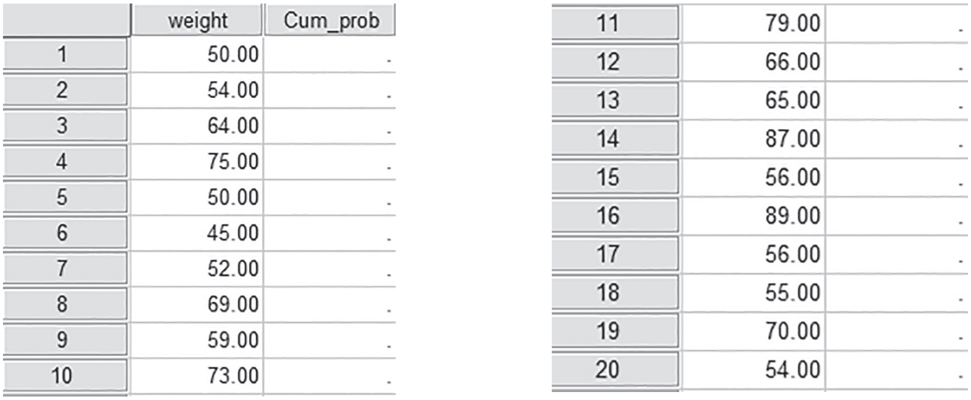
The functions commands as discussed in the previous sections could also be used to compute the probability for normal distributed random variable. In order to comprehend the same, we take an example of hypothetical data, reflecting the weight (kg) of 20 cases. The variable *weight* is normally distributed as found non-significant p-value (test statistics = 0.173, $p > 0.05$, 0.119) at 5% LoS by employing one-sample KS test.

1. **Creating data set for normal density function:** Below mentioned data indicates the weight of 20 cases. The mean and standard deviation of this normally distributed variable *weight* are 63.4 and 12.5, respectively, and is denoted as $x \sim$ N (63.4, 12.5). In SPSS, first we create two variables—*weight* as random variable (x) and second *cum_prob*. The steps used to compute the normal probabilities by using the steps are mentioned in Exhibit 20.10.

Exhibit 20.10. Open SPSS Data Editor » Create two variables » Weight and cum_prob » Assign values of weight in first column » Leave blank for cum_prob variable

The SPSS Data Editor appears on the screen as shown in Figure 20.16.

Figure 20.16 SPSS Data Editor: Random Variable *x* for Normal Probability



	weight	Cum_prob
1	50.00	.
2	54.00	.
3	64.00	.
4	75.00	.
5	50.00	.
6	45.00	.
7	52.00	.
8	69.00	.
9	59.00	.
10	73.00	.
11	79.00	.
12	66.00	.
13	65.00	.
14	87.00	.
15	56.00	.
16	89.00	.
17	56.00	.
18	55.00	.
19	70.00	.
20	54.00	.

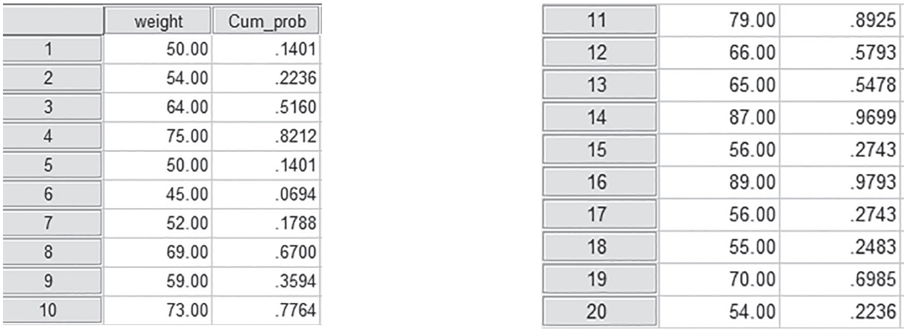
Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

2. Computing probability under the curve: We compute cumulative probability distribution for random variable *weight* by using the steps as shown in Exhibit 20.11.

Exhibit 20.11. Transform » Compute variable » Type cum_prob in Target variable box » Select and click on CDF & Noncentral CDF in Functions group » Select CDF.Normal in Functions and Special Variables Box and double click » Select weight In Type & Label box and transferred into numeric expression box, type 63.4 and 12.5 in place of ?, ? » Press OK » Change existing variable » Press OK

The results appear on the screen as shown in Figure 20.17.

Figure 20.17 Normal Probabilities: *Weight*



	weight	Cum_prob
1	50.00	.1401
2	54.00	.2236
3	64.00	.5160
4	75.00	.8212
5	50.00	.1401
6	45.00	.0694
7	52.00	.1788
8	69.00	.6700
9	59.00	.3594
10	73.00	.7764
11	79.00	.8925
12	66.00	.5793
13	65.00	.5478
14	87.00	.9699
15	56.00	.2743
16	89.00	.9793
17	56.00	.2743
18	55.00	.2483
19	70.00	.6985
20	54.00	.2236

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

We can obtain the probability under the curve for any random variable based on these computed probabilities as shown in Figure 20.15. For instance, the probability for random variable *weight* is less than or equals to 75 would be computed as

$$(P \leq 75 \text{ kg}) \text{ would be } 0.8212 \text{ or } 82\% \ (0.8212 \times 100 = 82).$$

Similarly, the probability for getting weight equals to or more than 75 ($P \geq 75$ kg) would be

$$1–(P \geq 75 \text{ kg}), \text{ thus } 1–0.82 = 0.18 \text{ or } 18 \text{ per cent.}$$

Likewise, we can also ascertain the probability under the curve indicating the interval between two specific values. For instance, the probability of *weight* 50 kg to 70 kg ($P \leq 50$ kg *weight* P ≥ 70 kg) can be computed as

$$P \leq 50 \text{ kg} = 0.1401$$

$$P \geq 70 \text{ kg} = 0.6985$$

Hence, P ≤ 50 kg *weight* $P \geq 70$ kg = 0.6985–0.1401 = 0.2884 or 28%.

20.7 Sampling Distribution

A sampling distribution is the most important phenomenon in data analysis. It refers to a frequency distribution of sample statistics collected from a large number of different samples from a specific population. Each sample is of equal size and independent of each other. In sampling distribution, the random variable is a sample mean $(\bar{x})$ or any other descriptive statistics rather than discrete or continuous random variable as discussed in the previous sections. The distribution of these sample statistics $(\bar{x})$ collected from distinct samples is known as sampling distribution (Pandis 2015). As each sample is individualistic in nature, the sample mean derived from these respective samples also vary. Hence, the standard deviation and variance among these sample means indicate the variability in the sampling distribution. The mean of all these sample statistics or overall mean of sample means indicate the mean of target population. The standard deviation of all the sample statistics is called SEM. The purpose of sampling distribution is to estimate unknown population parameter based on the maximum probability of occurring a particular sample mean from this sampling distribution.

20.7.1 Simulating Sampling from SPSS

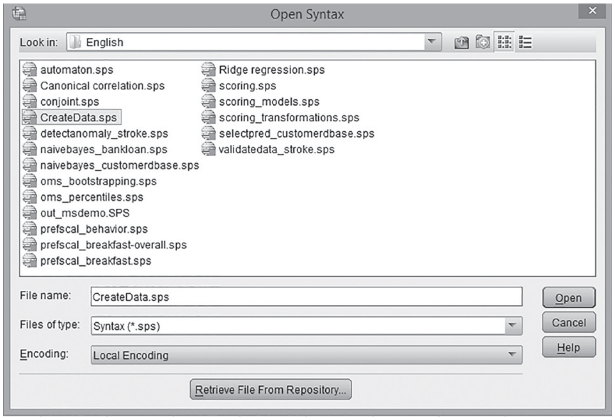
The very first step to comprehend the application of sampling distribution is to create a large number of samples of equal size. In view of this, we use the SPSS syntax as programming language instead of using graphical user interface. The SPSS syntax generates desired number of samples having equal size. The statistics as derived from these samples would be used for drawing sampling distribution. In this regard, we simulate a data set consisting of the weight of 50 cases on random basis belonging to 100 individual samples. We assume that the mean and standard deviation of weight in the target population are 60 kg and 10 kg, respectively. The steps used in SPSS to create sample through syntax are shown in Exhibit 20.12.

Exhibit 20.12. Open blank SPSS Data Editor » File » Open » Syntax » Open syntax dialog box » Select Syntax (*.sps) as file type » Select Local C drive » Program files (x86) » IBM » SPSS » Statistics » 22 » Samples » English » Click Createdata.sps » Open

The dialog box appears on the screen as shown in Figure 20.18.

Once you click on the *Open* tab, the Syntax editor appears on the screen. In order to simulate the samples, mention the following specification in area of INPUT PROGRAM as shown in Exhibit 20.13.

Figure 20.18 Main Dialog Box: Open Syntax



Look in: English

automaton.sps
Canonical correlation.sps
conjoint.sps
CreateData.sps
detectanomaly_stroke.sps
naivebayes_bankloan.sps
naivebayes_customerdbase.sps
oms_bootstrapping.sps
oms_percentiles.sps
out_msdemo.SPS
prefscal_behavior.sps
prefscal_breakfast-overall.sps
prefscal_breakfast.sps
Ridge regression.sps
scoring.sps
scoring_models.sps
scoring_transformations.sps
selectpred_customerdbase.sps
validatedata_stroke.sps

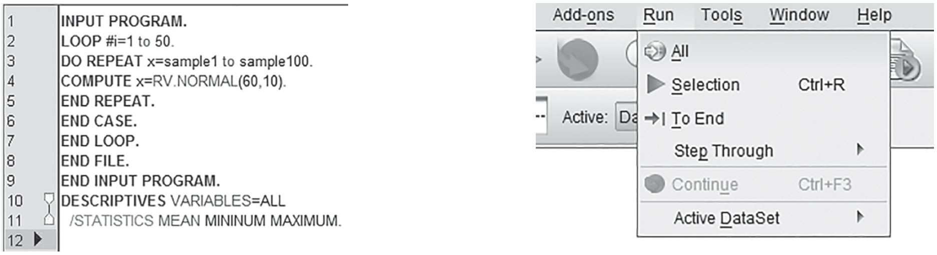
File name: CreateData.sps
Files of type: Syntax (*.sps)
Encoding: Local Encoding

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

Exhibit 20.13. INPUT PROGRAM » Type 1 to 50 for LOOP #i » Type sample1 to sample100 for DO REPEAT x » Type (60,10) for COMPUTE x=RV.NORMAL » Go to menu bar » Run » Click All

The dialog box of SPSS Syntax appears on the screen as shown in Figure 20.19.

Figure 20.19 Main Dialog Box: Simulating Data Set

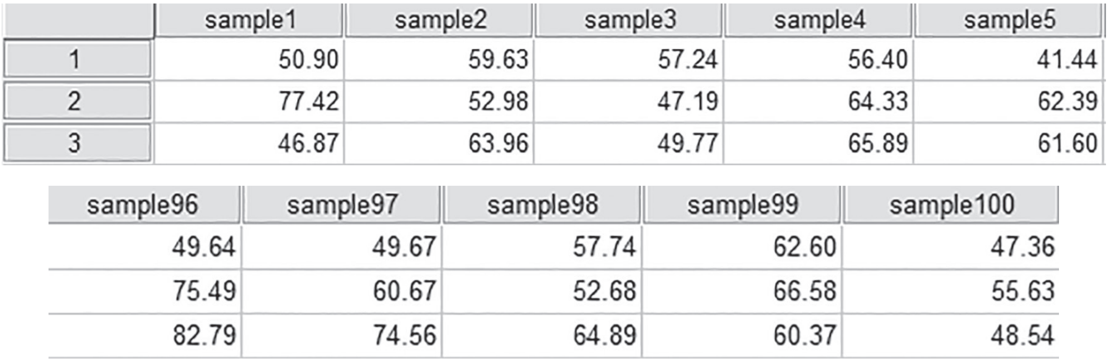


```
INPUT PROGRAM.
LOOP #i=1 to 50.
DO REPEAT x=sample1 to sample100.
COMPUTE x=RV.NORMAL(60,10).
END REPEAT.
END CASE.
END LOOP.
END FILE.
END INPUT PROGRAM.
DESCRIPTIVES VARIABLES=ALL
  /STATISTICS MEAN MININUM MAXIMUM.
```

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

By using the steps mentioned in Exhibit 20.13, 100 samples are simulated and displayed in the data view of SPSS Data Editor as shown in Fig. 20.20.

Figure 20.20 Simulated Data Set: 100 Samples



	sample1	sample2	sample3	sample4	sample5
1	50.90	59.63	57.24	56.40	41.44
2	77.42	52.98	47.19	64.33	62.39
3	46.87	63.96	49.77	65.89	61.60

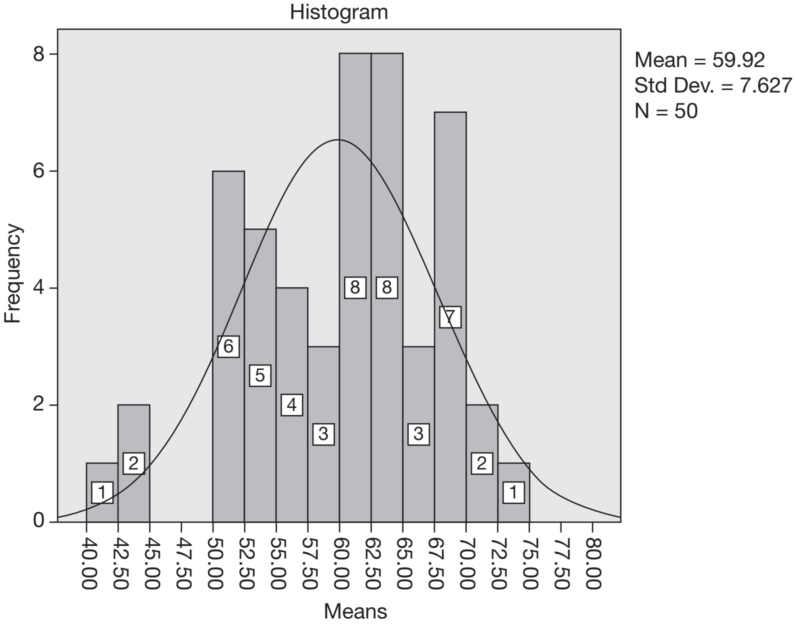
sample96	sample97	sample98	sample99	sample100
49.64	49.67	57.74	62.60	47.36
75.49	60.67	52.68	66.58	55.63
82.79	74.56	64.89	60.37	48.54

Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

20.7.2 Sampling Distribution of Means

The next step is to compute mean of each sample (sample 1 to 100) by using the steps as mentioned in Chapter 4 (Exhibit 4.7). These means are used to prepare the probability distribution. A new variable *Means* ($\bar{\bar{x}}$) is created in Data Editor with having all the sample means obtained from individual sample. Now, we create histogram overlapped with normal distribution curve and find the same as shown in Figure 20.21.

Figure 20.21 Histogram with Normal Distribution



Source: Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

According to Central Limit Theorem, the sampling distribution of means obtained from a large sample is normally distributed irrespective of whether the population is normal. Figure 20.21 resembles a normal distribution curve. The distribution also manifests the sample means that are far away from the mean values, and thus this difference in the sample means is evident while drawing the inferences for the population.

20.7.3 Implications of Sampling Distribution

Following are the main applications of sampling distribution:

1. The most occurring frequency of sample mean ($\bar{\bar{x}}$) in the sampling distribution estimates the population mean (μ); Means ($\bar{\bar{x}}$) = μ. Thus in our case, the overall population mean is 59.92. It is very close to the overall population mean = 60 kg.

2. The standard deviation of all the sample means (SEM) resembles the SEM of population. The SD of sampling distribution is 1.07. It can also verify the same in terms of the population also. The population SEM is:

$$\text{SEM} = \frac{\text{Standard Deviation}}{\sqrt{n}}, \frac{10}{\sqrt{50}} = 1.4.$$

Summary

- **Generating random sample:** A random sample for the analysis can be generated either for a specific experiment or from the population elements. The case-wise sample can be obtained either in proportion or with specific numbers.
- **Probability distribution:** Discrete and continuous are the two main probability distributions. These categories are based on whether to use discrete or continuous variables in the analysis. Discrete variable is countable, whereas the continuous is always measurable in nature.
- **Binomial distribution:** It is discrete probability distribution and used to summarize the independent observations. It resembles one of two outcomes under certain assumptions. The outcomes are in the form of two distinct events.
- **Poisson distribution:** This distribution indicates the occurrence of events within specific region such as particular time period, length, area or volume. The events of this distribution are also classified as success or failure.
- **Normal density function:** In case of normal distribution population, the probability is computed for certain interval of values or specific range of variable x. In such cases, the probability is better explained under certain area of *density function* or under a *curve*.
- **Sampling distribution:** The sampling distribution indicates a probability of a large number of sample means obtained from distinct and independent samples. The standard deviation of sample means is called SEM. The sampling distribution is used to estimate the population mean.

Key Terms

Binomial distribution
Central limit theorem
Commutative probability
Normal density function
Normal distribution
Poisson distribution

Random variable
Sampling distribution
Simulation
SPSS Syntax
Standard error of mean

Solved Example

Q1. The following data indicates monthly mobile bills amount (INR) for randomly selected 20 students in a University campus. The data is presented as follows:

Students	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Bill	500	450	1000	500	900	400	250	750	550	600
Students	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
Bill	700	500	450	500	300	450	800	950	600	350

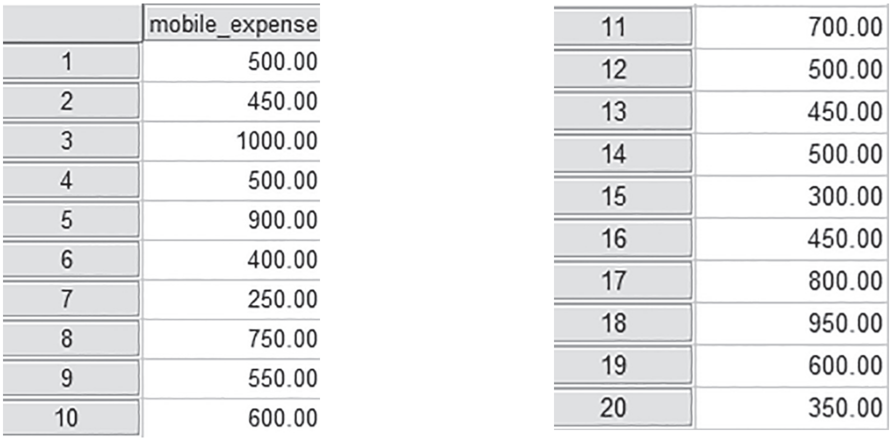
Give answer to the following on the basis of above-mentioned information:

a. Prepare data set in SPSS and examine whether data is normally distributed?
b. Find out the cumulative normal probability for amount of *mobile bills*
   - Less than or equals to 1,000 INR.
   - Equal to or more than 300 INR.

Solution

a. **Preparation of data set for mobile bills:** The variable *mobile_bills* is created in the variable view by using the steps as mentioned in Section 20.2.1. The data set is shown below in Figure 20.22.

Figure 20.22 Data View for Normal Distributed Variable



	mobile_expense
1	500.00
2	450.00
3	1000.00
4	500.00
5	900.00
6	400.00
7	250.00
8	750.00
9	550.00
10	600.00
11	700.00
12	500.00
13	450.00
14	500.00
15	300.00
16	450.00
17	800.00
18	950.00
19	600.00
20	350.00

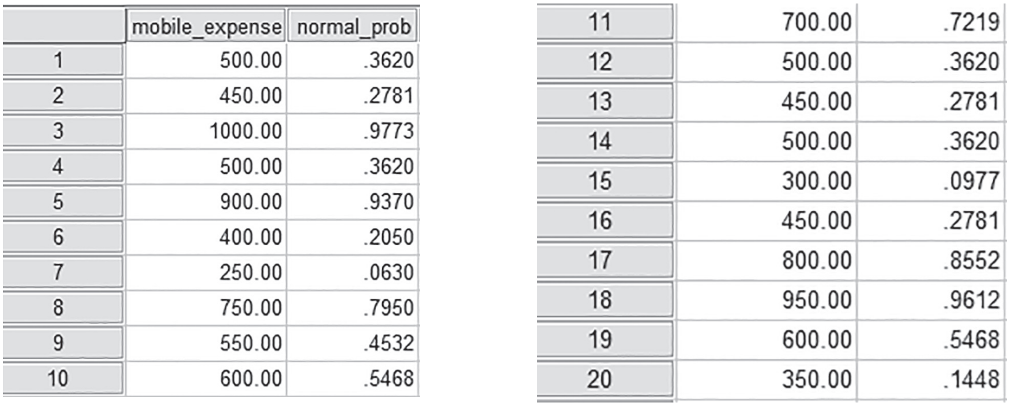
Examine Normality one-sample KS test is used to assess the normality of data as mentioned in Exhibit 8.1. The results appear as shown in Table 20.1. Higher *P*-value (*p* > 0.05, 0.062) at 5% LoS indicates that the mobile bill amount is normally distributed. Hence, we fail to reject the null hypothesis of normal distributed data.

Table 20.1 Normality Test

		mobile_expense
N		*20*
Normal parameters	Mean	575.0
	Std deviation	212.44
Most extreme differences	Absolute	0.188
	Positive	0.188
	Negative	–0.087
Test statistic		0.188
Asymp. Sig. (two-tailed)		0.062

b. **Normal probability distribution:** We create a new variable *normal_prob* in a variable view of SPSS Data Editor. The normal probabilities under the curve can be computed by using the steps as mentioned in Exhibit 20.11. The normal probability distribution of mobile bill is depicted in Figure 20.23.

Figure 20.23 Probability Under the Normal Curve



	mobile_expense	normal_prob
1	500.00	.3620
2	450.00	.2781
3	1000.00	.9773
4	500.00	.3620
5	900.00	.9370
6	400.00	.2050
7	250.00	.0630
8	750.00	.7950
9	550.00	.4532
10	600.00	.5468
11	700.00	.7219
12	500.00	.3620
13	450.00	.2781
14	500.00	.3620
15	300.00	.0977
16	450.00	.2781
17	800.00	.8552
18	950.00	.9612
19	600.00	.5468
20	350.00	.1448

Hence, the cumulative normal probability for billing amount less than or equals to 1,000 ($P \leq$ 1,000 INR) would be 0.9773 or 97 per cent (0.9773 × 100 = 97).

Similarly, the probability for getting equal or more than 300 ($P \geq$ 300 INR) would be:

$$1–(P \geq 300 \text{ INR}), \text{ thus } 1–0.97 = 0.03 \text{ or } 3\%.$$

Hands-on Practice

Q1. The past record from one maternity clinic indicates that 30 per cent deliveries of the pregnant cases are normal, whereas 70 are operated by caesarean. If this claim is true, find out the probability of normal deliveries from 1 to 10 among 10 pregnant cases in that clinic.

a. Prepare data sheet in SPSS as per the given information.
b. Find out the probability distribution according to the claim of clinic.
c. Prepare suitable plot of probability distributions.

Q2. In one residential area, approximately 13 power cuts were executed in one month due to excessive load of Air Conditioners in the summers. Find out the probabilities of following power cut(s) in one month based on the given facts.

Power cut per month	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Probability										

Q3. Use data set student.sav and create random data as per the following instructions:

a. Select approximately 50% cases randomly based on the existing data set.
b. Select 10 cases randomly based on this data set.

Q4. Based on a survey in one departmental store, 2 customers are recorded as the member of *Smart Saving Card Scheme* out of 10. If this analysis is true, find out the probability of the card membership for the following number of customers in a survey consisting of 50 cases.

Number of Customers	*Probability*
5	
12	
17	
20	
22	
27	
34	
45	

a. Prepare data sheet in SPSS as per the case facts.
b. Find out the probability distribution based on the given conditions.
c. Prepare suitable plot of probability distributions.

Problem Based on Real Data

Q1. Data set *insurance* presents information about health insurance of 1,339 persons. Use the Web link https://www.kaggle.com/bmarco/health-insurance-d;ata#insurance.csv from the open data source Kaggle Inc. (2019) and perform the following tasks.

a. Convert insurance.csv file to Excel workbook and import data set into SPSS Data Editor.
b. Select approximately 60 per cent cases randomly based on existing data set twice and compute descriptive statistics for each random data set.
c. Compare descriptive statistics for these two separate cases and analyse the results.

References

Joseph, L., and C. Reinhold. 2003. 'Introduction to Probability Theory and Sampling Distribution'. *American Journal of Roentgenology* 180 (4): 917–923.

Kaggle Inc. 2019. *Health Insurance Data*. Available at: https://www.kaggle.com/bmarco/health-insurance-data#insurance.csv (accessed on 17 September 2019).

Pandis, N. 2015. 'The Sampling Distribution'. *American Journal of Orthodontics and Dentofacial Orthopedics* 147 (4): 517–519.

Weisstein, E. W. 2019. *Binomial Distribution, From MathWorld-A Wolfram Web*. Available at: http://mathworld.wolfram.com/BinomialDistribution.html (accessed on 5 July 2019).