Week 9 :

21.1 Simulation – Basic definitions :

What is Simulation ?

Definition

Simulation is a technique which imitates a real world system's change with the elapse of time. It attempts to duplicate the features, appearance and characteristics of the real world system.

DEFINITION ■ A **system** is a collection of entities that act and interact toward the accomplishment of some logical end. ■

DEFINITION ■ The **state** of a system is the collection of variables necessary to describe the status of the system at any given time. ■

DEFINITION ■ A **discrete system** is one in which the state variables change only at discrete or countable points in time. ■

DEFINITION ■ A **continuous system** is one in which the state variables change continuously over time. ■

There are two types of simulation models: static and dynamic.

DEFINITION ■ A **static simulation model** is a representation of a system at a particular point in time. ■

We usually refer to a static simulation as a **Monte Carlo simulation.**

DEFINITION ■ A **dynamic simulation** is a representation of a system as it evolves over time. ■

- A deterministic simulation model is one that contains no random variables.
- A stochastic simulation model contains one or more random variables.

Why simulate?

- Study of the interactions of elements in a complex system;
- Effect of change in information, organization and environment of a system on system behaviour;
- Provide direction for enhancements / upgrades;
- Verification of analytical results;
- Preparation for possible outcomes;
- Animation helps with visualization;
- Training.

General steps of simulation

1. Define the problem.
2. Determine the important variables.
3. Build the simulation model.
4. Set the parameters for the model to test against.
5. Execute the simulation.
6. Analyze the results.
7. Decide on a plan of action.

Advantages of simulation

- Relativley straightforward and flexible.
- Computers simplifies modelling.
- Enables the analysis of complex systems.
- Analyzes "what if "-scenarios.
- No impact on th real system.
- Enables the study of interaction of system components.
- Time-compression.
- Enables the inclusion of probability distributions which are not possible with analytical methods.

Disadvantages of simulation

- Good models of complex systems can be very expensive.
- Simulation does not generate optimal solutions as might be found with some analytical techniques.
- Conditions and constraints of scenarios must be set my decisionmakers for implementation in a simulation.
- Models tend to be unique and results therefore generally not transferreable to different problems.

For now this is all we need to know on simulation, at week 12 we will continue.

20.12 How to Tell Whether Interarrival Times and Service Times Are Exponential :

With **Goodness-of-fit tests**

- 2 types : **χ^2-test** used for continuous and discrete probability distributions.

  Kolmogorov-Smirnov toetse / tests

  used for continuous probability distributions.

χ^2-test :

The χ^2-statistic

$$\chi^2_{data} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

where k is determined such that $e_i \geq 5$.

$$k = \lceil log_2 n \rceil + 1$$

where n is the number of observations.

o_i is the observed value for the interval i

e_i is the expected value for the interval i

Parameter van die eksponensiële verdeling $\hat{\lambda}$ / Parameter of the exponential distribution $\hat{\lambda}$

Given a series of n interarrival times $\{t_1, t_2, \ldots, t_n\}$, then

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i}$$

Intervals

Let u_i be the upper limit of each interval i. For the exponential distribution then,

$$u_i = -\frac{1}{\hat{\lambda}} \ln\left(1 - \frac{i}{k}\right) \qquad 1 \leq i \leq k-1$$

$$u_k = M$$

$$l_1 = 0$$

$$l_i = u_{i-1} \qquad 2 \leq i \leq k$$

Determining $\chi^2_{k-r-1,\alpha}$

- k is the number of intervals.
- r is the number of parameters estimated from the data.
- α is the level of significance.

If the data is used to estimate a parameter such as $\hat{\lambda}$ for the exponential distribution, then $r = 1$.

In the absence of a specific prescription assume $\alpha = 0.05$.

where ν is referred as the number of degrees of freedom of the distribution, $\nu = k - r - 1$.

H_0: $t_1, t_2, \ldots, t_n$ is a random sample from a random variable with density $f(t)$

H_a: $t_1, t_2, \ldots, t_n$ is not a random sample from a random variable with density function $f(t)$

we accept H_0 if $\chi^2(\text{obs}) \leq \chi^2_{k-r-1}(\alpha)$ and accept H_a if $\chi^2(\text{obs}) > \chi^2_{k-r-1}(\alpha)$.

EXAMPLE 16 Interarrival Times: Exponential or Not Exponential?

The following interarrival times (in minutes) have been observed: 0.01, 0.07, 0.03, 0.08, 0.04, 0.10, 0.05, 0.10, 0.11, 1.17, 1.50, 0.93, 0.54, 0.19, 0.22, 0.36, 0.27, 0.46, 0.51, 0.11, 0.56, 0.72, 0.29, 0.04, 0.73. Does it seem reasonable to conclude that these observations come from an exponential distribution?

Solution There are 25 observations with $\sum_{i=1}^{i=25} t_i = 9.19$. Thus, $\bar{\lambda} = \frac{25}{9.19} = 2.72$ arrivals per minute. We now test whether or not our data are consistent with an exponential random variable (call it **A**) having a density $f(t) = 2.72e^{-2.72t}$. We choose five categories so as to ensure that the probability that an observation from **A** falls into each of the five categories is .20. This yields $e_i = 25(.20) = 5$ for each category. To set the category boundaries, we need to determine the cumulative distribution function, $F(t)$, for **A**:

$$F(t) = P(\mathbf{A} \le t) = \int_0^t 2.72e^{-2.72s}\,ds = 1 - e^{-2.72t}$$

Then we choose the categories to be as follows:

Category 1 $0 \le t < m_1$ minutes

Category 2 $m_1 \le t < m_2$ minutes

Category 3 $m_2 \le t < m_3$ minutes

Category 4 $m_3 \le t < m_4$ minutes

Category 5 $m_4 \le t$ minutes

where $F(m_1) = .20$, $F(m_2) = .40$, $F(m_3) = .60$, and $F(m_4) = .80$.

Since $F(t) = 1 - e^{-2.72t}$, we see that for any number p, the value of t satisfying $F(t) = p$ may be found as follows:

$$1 - e^{-2.72t} = p$$
$$1 - p = e^{-2.72t}$$

Taking logarithms (to base e) of both sides yields

$$t = \frac{\ln(1-p)}{-2.72}$$

$$m_1 = \frac{\ln .80}{-2.72} = 0.08$$

$$m_2 = \frac{\ln .60}{-2.72} = 0.19$$

$$m_3 = \frac{\ln .40}{-2.72} = 0.34$$

$$m_4 = \frac{\ln .20}{-2.72} = 0.59$$

Hence, our categories are as follows:

Category 1 $0 \le t < 0.08$ minute

Category 2 $0.08 \le t < 0.19$ minute

Category 3 $0.19 \le t < 0.34$ minute

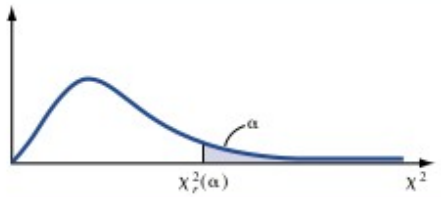
Category 4 $0.34 \le t < 0.59$ minute

Category 5 $0.59 \le t$

After classifying the data into these categories, we find that $o_1 = 6$, $o_2 = 5$, $o_3 = 4$, $o_4 = 5$, and $o_5 = 5$. By the construction of our categories, $e_1 = e_2 = e_3 = e_4 = e_5 = .20(25) = 5$. We now compute $\chi^2(\text{obs})$:

$$\chi^2(\text{obs}) = \frac{(6-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(4-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(5-5)^2}{5}$$
$$= .20 + 0 + .20 + 0 + 0 = .40$$

TABLE 9
Percentiles of Chi-Square Distribution



d.f. ν	α .990	.950	.900	.500	.100	.050	.025	.010	.005
1	.0002	.004	.02	.45	2.71	3.84	5.02	6.63	7.88
2	.02	.10	.21	1.39	4.61	5.99	7.38	9.21	10.60
3	.11	.35	.58	2.37	6.25	7.81	9.35	11.34	12.84
4	.30	.71	1.06	3.36	7.78†	9.49	11.14	13.28	14.86
5	.55	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	.87	1.64	2.20	5.35	10.64	12.59	14.45	16.81	18.55
7	1.24	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.65	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.95
9	2.09	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.56	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	3.05	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.57	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	4.11	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.66	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	5.23	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.81	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	6.41	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	7.01	9.39	10.86	17.34	25.99	28.87	31.53	34.81	37.16
19	7.63	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	8.26	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.90	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	9.54	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	10.20	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	10.86	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	11.52	14.61	16.47	24.34	34.38	37.65	40.65	44.31	46.93
26	12.20	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	12.88	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.64
28	13.56	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	14.26	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	14.95	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	22.16	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	29.71	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	37.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	45.44	51.74	55.33	69.33	85.53	90.53	95.02	100.43	104.21
80	53.54	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	61.75	69.13	73.29	89.33	107.57	113.15	118.14	124.12	128.30
100	70.06	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17

Source: Richard A. Johnson and Dean W. Wichern, *Applied Multivariate Statistical Analysis*, © 1982, p. 583. Reprinted by permission of Prentice Hall, Inc., Englewood Cliffs, New Jersey.

†*Note:* For example, $P(\chi^2_4 > 7.78) = .10$.

We arbitrarily choose $\alpha = .05$. Since we are trying to fit an exponential distribution to interarrival times, $r = 1$. Then $\chi^2_3(.05) = 7.81$, and we see that for $\alpha = .05$, we can accept the hypothesis that the observed interarrival times come from an exponential distribution with $\lambda = 2.72$ arrivals per minute.

Alternatively, we could have found the cutoff point for the chi-square test with the formula

$$=\text{CHINV}(.05,3)$$

This formula yields the value 7.81.

#Here I conveniently generate data which would normally come to the
#researcher from observations.A student is not required (at this stage)
#to generate any test data.

#Data generated.
```
set.seed(101)

A=rexp(100,1/3)

B=rexp(100,1/3)
```

#This is a the chi-squared test applied to the data in the vector B,
#which as stated above, is normally the result of some observation
#exercise.

#Find the number intervals to use in the test by Sturge's rule.

$$k = \lceil log_2 n \rceil + 1$$

```
k=ceiling(log(length(B),2))+1
```

#Estimate the parameter for the theoretical distribution from the
#data (if required in the question).

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i}$$

```
lambdahat=1/mean(B)
```

#Set bounds for the counting intervals

$$u_i = -\frac{1}{\hat{\lambda}} \ln\left(1 - \frac{i}{k}\right) \qquad 1 \le i \le k-1$$

$$u_k = M$$

```
ub_i=c()
for(i in 1:k-1)ub_i[i]=(-1/lambdahat)*log(1-(i/k))
ub_i[k]=max(B)+1
ub_i
```

$$l_1 = 0$$

$$l_i = u_{i-1} \qquad 2 \le i \le k$$

```
lb_i=c(0)
for(i in 2:k)lb_i[i]=ub_i[i-1]
```

#Prepare a structure for the display of the results

```
myDF=data.frame(lb_i,ub_i)
```

#Initialize the vector that will contain
#the result of the counting exercise.

```
o_i=rep(0,k)
```

o_i is the observed value for the interval i

```
o_i
```

#Do the counting

```
for(j in 1:length(B)){
  for(i in 1:k)if(B[j]>=lb_i[i]&&B[j]<=ub_i[i])o_i[i]=o_i[i]+1
}
```

#View the result of the counting

```
o_i
```

#Add the counting result to the display structure

```
myDF=cbind(myDF,o_i)
```

MyDF

#Calculate the expected counts ineach interval

e_i is the expected value for the interval i

```
e_i=rep(length(B)/k,k)
```

```
# Add the expeceted counts to the dispaly structure
myDF=cbind(myDF,e_i)

MyDF

#Calculate the chis squeared statistic for the data.
chisqstat_data=sum((o_i-e_i)^2/e_i)

#calculate the critical chisquared value to compare against
alpha=0.05

r=1

df=k-r-1

chisqstat_critical=qchisq(alpha,df,lower.tail = FALSE)

chisqstat_critical

#Pronounce on the set hypothesis based on the result of the comparisons
if(chisqstat_data<chisqstat_critical)cat("Do not reject H_0")else cat("Reject H_0")
```

$$\chi^2_{data} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

$$\nu = k - r - 1.$$

Kolmogorov-Smirnov toetse / tests

The Kolmogorv test or K-S test as it is also known, is a goodness-of-fit test, just like the χ^2-test, but with fewer applications.

Determining D_n

Given the set $\{X_1, X_2, \ldots, X_n\}$ and the CDF $\hat{F}$

1. Define an empirical CDF $F_n(x)$ where $F_n(x) = \frac{\text{Aantal } X_i's \leq x}{n}$.
2. Now $F_n(x) = \frac{i}{n}$ vir $1 \leq i \leq n$ is a right continuous step-function,
3. from which D_n as determined as

$$D_n = \sup_x \left\{ |F_n(x) - \hat{F}(x)| \right\}.$$

Steps to calculate D_n

1. Rank the original data from small to large such that $X_i < X_{i+1},\ i \in \{1, 2, 3, \ldots, n\}$;
2. Determine

$$D_n^+ = \max_{1 \leq i \leq n} \left\{ \left| \frac{i}{n} - \hat{F}(X_i) \right| \right\};$$

3. Determine

$$D_n^- = \max_{1 \leq i \leq n} \left\{ \left| \hat{F}(X_i) - \frac{i-1}{n} \right| \right\};$$

4. Now

$$D_n = \max \left\{ D_n^+,\ D_n^- \right\}.$$

Voltooiing van die toets / Completion of the test

Verkry die *Aangepaste* $D_n = D_n^{'}$ en vergelyk dit met die gepaste kritieke c uit die tabel.

Find the *Adjusted* $D_n = D_n^{'}$ and compare with the critical c from the table.

Critical Values $c_{1-\alpha}$, $c'_{1-\alpha}$ and $c''_{1-\alpha}$ for $1-\alpha$

Adjusted K-S Statistics (Law and Kelton)

Case	Adjusted test statistic $D_n^{'}$	0.900	0.950	0.975	0.990
All parameters known	$\left(\sqrt{n}+0.12+\frac{0.11}{\sqrt{n}}\right) D_n$	1.224	1.358	1.480	1.628
$N\left(\overline{X}(n), s^2(n)\right)$	$\left(\sqrt{n}-0.01+\frac{0.85}{\sqrt{n}}\right) D_n$	0.819	0.895	0.955	1.035
Exponential $\left(\overline{X}(n)\right)$	$\left(D_n-\frac{0.2}{n}\right)\left(\sqrt{n}+0.26+\frac{0.5}{\sqrt{n}}\right)$	0.990	1.094	1.190	1.308

#Kolmogorov-Smirnov test

#Part1
#This part is only included to serve as a basis for the
#graphical comparison between F_n(x) and Fhat_n(x) for
#an extremely small dataset (only 5 elements). Students
#are not required to do this part - it is not part of
#the KS-test procedure

#Download the small datset by the indicated name from Sunlearn
#and save to some directory that you point to as working
#directory. The following statement then reads the values
#of the dataset.

```
MyData=scan("MydataInClass.txt")

MyData

MyDataSorted=sort(MyData)

MyDataSorted

n=length(MyDataSorted)

mu=30
sigma=10

cps=seq(0,1,1/n)

cps

#The following 4 statements display the 2 graphs.
plot(stepfun(MyDataSorted,cps))
x=seq((min(MyDataSorted)-2),(max(MyDataSorted)+2),0.1)
y=pnorm(x,mu,sigma)
lines(x,y,col="red")
```

```
#Part 2
#The following set of statements are the required steps
#of the KS-test

#Initialize the vectors to contain the calculated differences
DnMinusSet=c()
DnPlusSet=c()


#Sort the data in an ascending order.
MyDataSorted=sort(MyData)

#calculate the absolute value of the differences
for(i in 1:n){
  DnPlusSet[i]=abs((i/n)-pnorm(MyDataSorted[i],mu,sigma))
  DnMinusSet[i]=abs(pnorm(MyDataSorted[i],mu,sigma)-((i-1)/n))
}

#Optionally display the sets containing the differences.
DnPlusSet
DnMinusSet


#Find the maximum from each set
DnPlus=max(DnPlusSet)
DnMinus=max(DnMinusSet)

#Find the maximum difference.
Dn=max(DnMinus,DnPlus)

Dn

#Calulate the adjusted D value relevant to the
#tested hypothesis.

Dprimen=(sqrt(n)+.12+0.11/sqrt(n))*Dn

Dprimen

#Compare the adjusted D value with the
#relevant critical value from the table
#on slide 9. (The hypothesis determines the
#row from which the formula for the adjusted
#D value AND the critical value are read )
```