HIROSAKI UNIVERSITY

Radiation Emergency Medicine 2014 Vol.3, No.1 1–6

REM
Radiation Emergency Medicine

Special Contribution

Basic Epidemiology
—Methods and Their Application to Epidemiology on Cancer and Radiation

Suminori Akiba

Department of Epidemiology and Preventive Medicine, Kagoshima University Graduate School of Medical and Dental Sciences, 8-35-1, Sakuragaoka, Kagoshima 890-8544, Japan

Received 22 August 2013; revised 25 November 2013; accepted 29 November 2013

I. The basic terminology necessary for understanding epidemiological studies

1. What is epidemiology

World Health Organization (WHO) defines epidemiology as the study of the distribution and determinants of health-related states or events (including disease), and the application of this study to the control of diseases and other health problems[1]. The fourth edition of Dictionary of Epidemiology (referred to as DE-4, hereinafter), edited by James Last[2] gives a similar definition, which is as follows: "The study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to control of health problems." The fifth edition of Dictionary of Epidemiology (DE-5), edited by Miquel Porta[3], gives a slightly different definition: "Epidemiology is the study (or the science of the study) of the patterns, causes, and effects of health and disease conditions in defined populations."

Anyway, epidemiology deals with i) the distributions of diseases among human populations, ii) the distributions of factors that are known or suspected to cause a disease, and/or iii) the causal relationship between exposures and diseases. Diseases can be health related problems/conditions.

Suminori Akiba: Department of Epidemiology and Preventive Medicine, Kagoshima University Graduate School of Medical and Dental Sciences, 8-35-1,Sakuragaoka,Kagoshima 890-8544,Japan
E-mail: akiba@m.kufm.kagoshima-u.ac.jp

2. Observational study and interventional (experimental) study

There are observational and experimental epidemiological studies. National Cancer Institute (NCI) Dictionary of Cancer Terms defines the observational study as a type of study in which individuals are observed or certain outcomes are measured. If attempts are made to affect the outcome, such a study is called the interventional study[4]. All the studies described in this paper are observational studies unless otherwise specified. Experimental epidemiological studies are also called interventional studies.

Sir Percvall Pott reported a high incidence of scrotal cancer (later found to be squamous cell carcinoma) in chimney sweepers in an essay published in 1775. He suspected that the cancer was caused by exposure to soot[5]. This is considered to be the first observational epidemiological study to demonstrate that a malignancy could be caused by an environmental carcinogen and occupational exposure.

Scurvy is a disease caused by vitamin C deficiency. It often presents itself initially as symptoms of malaise and lethargy, followed by formation of spots on the skin, spongy gums, and bleeding from the mucous membranes. Citrus fruit was believed to be a cure for scurvy. James Lind conducted an experimental epidemiological study showing this effect. When he was on board Her Majesty's Ship Salisbury of the British Royal Navy as a ship's surgeon in 1747, he divided the 12 patients into 6 groups. One of those groups was provided with two oranges and one lemon every day. This group showed the most sudden and visible good effects.[6]

John Snow is famous for finding the source of a cholera outbreak in Soho, London, in 1854. The pattern of the disease made him convinced that the public water pump on Broad Street (now Broadwick Street) was the source of the outbreak. His approach can be regarded as an observational epidemiological study. After disabling the well pump by removing its handle, the epidemic was gone. This episode is considered as an interventional study or experimental epidemiological study. Lancet, a British medical journal, writes as follows[7]: "Snow's intervention in Broad Street, Soho, in 1854, when he persuaded the authorities to remove the handle from a contaminated pump well, has caught the public imagination, but it was his "Grand Experiment" that same year that secured his huge reputation in epidemiology. During Britain's second cholera epidemic in 1848-49, both the Lambeth and the Southwark and Vauxhall water companies were taking their supplies from the Thames next to where the London sewers were discharged. By the time of the 1854-55 epidemic, however, Lambeth had moved its works up river out of reach of the sewage. Here, Snow saw, was the perfect means of testing his theory. He compared the numbers of cholera victims whose water was supplied by the two different companies in 1848-1849 with the numbers in 1854-55. During 1848-49, the death rates for the two companies were the same, but by 1854, after Lambeth's move, Southwark and Vauxhall's rate was between eight and nine times higher, and in the first 4 weeks of the epidemic, Southwark and Vauxhall customers had a 14-fold higher risk. In 1855, Snow published a much-expanded second edition of On the Mode of Communication of Cholera that included these results. Again he was largely ignored, although by then, the idea that polluted water had some part to play in cholera was gaining ground." The "Grand Experiment" of John Snow is considered to be an experimental epidemiological study.

3. Disease incidence

When investigating the causes of a disease, we are interested in the occurrence of the disease in relation to the exposure of interest. Disease occurrence has two important elements in epidemiology. One is the speed of occurrence, and the other, the proportion of people with newly developed disease. Let me explain this point using Figure 1A. In the upper panel of this figure, there are five exposed subjects, and three of them developed leukemia during the 10-year follow-up. Therefore, the frequency of leukemia cases is 3/5 among the exposed. The lower panel shows the follow-up of the unexposed (n = 5). Among them, there are three leukemia cases, and therefore the frequency of leukemia is 3/5, which is the same as that among the exposed.

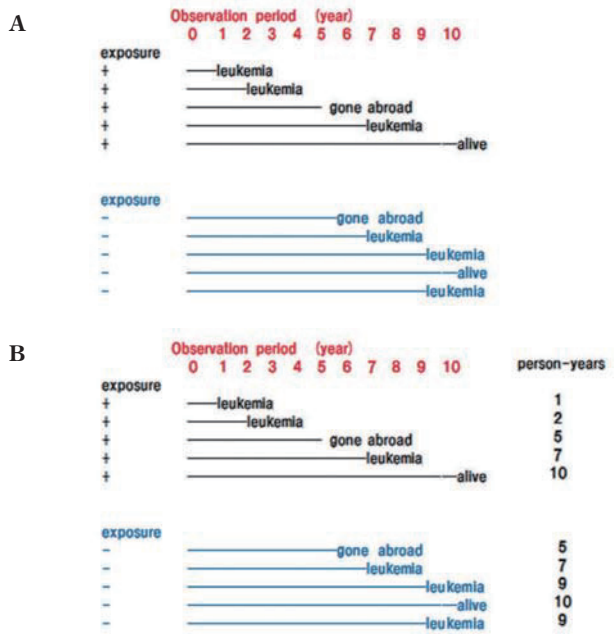

Fig. 1. The results of 10-year follow-up of the exposed and the unexposed.

However, a closer look at the figure will make you realize that leukemia cases among the exposed tended to occur earlier during the follow-up. In the following figure, the years of follow-up for each subject are shown as person-years. For example, the first person among the exposed group developed leukemia at 1 year after follow-up. In the second person, it took 2 years for leukemia development. The third person was gone abroad at 5 year after follow-up. However, in this subject, we know that leukemia was not developed for the first 5 years of follow-up. The total years of follow-up for the exposed group are 25 years. Likewise, the total years of follow-up for the unexposed were 40. Therefore, leukemia incidence is 3/25 among the exposed, and is 3/40 among the unexposed. In this example, the leukemia incidence among the exposed group is 1.6-fold larger than that among the unexposed when person-years are taking into account.

Needless to say, if all leukemia cases develop after three years of follow-up regardless of exposure, the occurrence of this blood malignancy is compared on the basis of the numbers of cases among the two groups.

4. Incidence rate

The frequency of leukemia, which was calculated as 3/5 in both groups in Figure 1B, is called a cumulative incidence (or cumulative incidence rate). It is also called incidence proportion (The second edition of Modern Epidemiology[8]. Hereinafter, it will be abbreviated as ME-II). A proportion is unit-less since its nominator and

Table 1. How to calculate expected numbers of death.

age	mortality rate (/1000) of standard population		population size of town B		Expected cancer deaths
0-39	5	×	1000	=	5
40-59	25	×	1000	=	25
60-79	30	×	1000	=	30
80+	50	×	1000	=	50
total			4000		110

denominator have the same unit, which are the numbers of people, and they are cancelled out by division.

A strict definition of incidence in epidemiology is the number of newly occurred cases during follow-up divided by person-time of follow-up. The unit of its nominator is the number of persons as is the case with a proportion. However, the unit of its denominator is different. Since the denominator of an incidence rate is person-years, the unit is the number of people × time. Therefore, the unit of incidence rate is 1/time when the numbers of people are ignored. Note that the unit of speed, which is distance/time, has the same denominator (= time). In other words, the incidence rate is a kind of a speed of disease occurrence.

In summary, the incidence rate of a disease used in epidemiology has two important elements: the speed of disease occurrence (exemplified by 1/time), and the proportion (= unitless) of newly developed cases among the study subjects. Incidence rate is sometimes referred to as incidence density, because it corresponds to the density of events in the person-time space (ME-II).

5. SMR analysis

1) Example

In the early 1980s, an increased risk of childhood leukemia was noted in the neighborhood of Sellafield nuclear reprocessing plant in the UK, which was commissioned in the 1950s. The nearest village to the plant is Seasclae, which is located at the northern coastal tip of Millom Rural District. In the study of Gardner and Winter[9], SMRs for residents aged 0-24 in Millom Rural District was 0.63 during 1959-76 and 4.35 for the period 1968-78. Later, studies conducted by Gardner *et al.* revealed that the risk increase was limited to the children born in Seascale, and that the children who moved into the area after birth did not experience any increased risk[10, 11].

2) How to calculate an SMR.

Suppose that you followed 4,000 residents in a town in Fukushima for 10 years, and ascertained 100 cancer deaths. Is it possible for you to tell whether there is any excess cancer or not? Needless to say, the answer is NO. You need to have the data from a control group and compare the mortality experiences of the two groups. Another approach is to calculate an expected value using the mortality among a standard population, and calculate an SMR (Standardized Mortality Ratio), which is the ratio between the observed and expected numbers of deaths. It is called external comparison. On the other hand, in an internal comparison, a population is divided into the exposed and unexposed to make comparison.

Theoretically, you can use any standard population for calculating the expected number of death. The most commonly used standard populations are the entire nation, the entire states, and the entire prefectures, in which the study population is included. When you use the expected mortality of standard population shown in the second left column of Table 1, the expected number of cancer deaths is 110. The SMR is approximately 0.91, which is 100 (= the observed number of cancer deaths) divided by 110 (= the expected number of cancer deaths). Not a small number of epidemiologists express an SMR in percentage. In this case, the SMR will be expressed as 91, rather than 0.91.

6. Relative risk and risk difference

When comparing the incidence of two groups, relative risks (RRs) and risk differences (RDs) are used. In Figure 1, RR = (3/5)/(3/5) = 1 if cumulative incidence rates are used. In the case of incidence rate, RR = (3/25)/(3/40) = 1.6. A risk difference (RD) is the difference of cumulative incidence rates (or incidence rates). In this example, RD = 3/25-3/40. Since 3/25 and 3/40 are 12,000/100,000 and 75,000/100,000, respectively, the RD per 100,000 is 12000 - 7500 = 4500. In cancer epidemiology, incidence rate is usually (but not always) expressed as cases per 100,000.

7. Excess relative risk

Frequently, RR is estimated by a log-linear model such as follows:

RR = (risk among the exposed)/(risk among the unexposed) = exp(β dose).

Therefore,

RR = exp (0) = 1.0 at dose = 0,

RR = exp (1) = 2.7 at dose = 1,

RR = exp (2) = 7.4 at dose = 2,

RR = exp (3) = 20.1 at dose = 3, and
RR = exp (4) = 54.6 at dose = 4.

Its dose-response curve is not linear but log-linear as shown in the following figure.

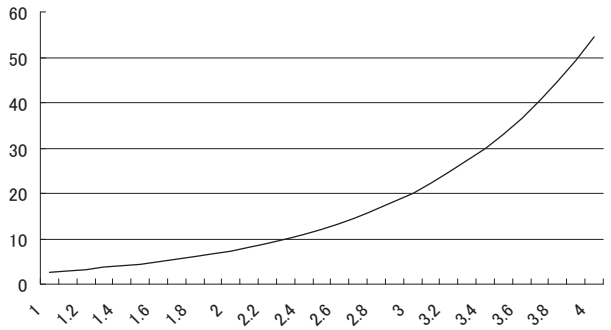

Fig. 2. A log-linear dose-response.

However, we sometimes need a risk parameter showing a linear dose-response. In a linear dose-response relationship, the risk increases (or decreases) in proportion to the dose; and if the dose is zero, the risk parameter should be zero. If we use RRs, even if dose is zero, the corresponding RR is 1. However, if you use RR-1, which is called excess relative risk (ERR), its value is zero when the dose is zero. Instead, you can use RDs. However, it is sometimes difficult to estimate RDs.

RR–1 = 0 at dose = 0,
RR–1 = 1 at dose = 1,
RR–1 = 2 at dose = 2,
RR–1 = 3 at dose = 3,
RR–1 = 4 at dose = 4.

This is a linear relationship.

8. Attributable fraction (exposed)

Attributable risk is the proportion of a disease (or other outcomes) in the exposed individuals that can be attributed to the exposure. This is calculated as follows:
attributable risk = (Ie – Iu) / Ie,
where
Ie is incidence (risk) among the exposed, and
Iu is incidence (risk) among the unexposed.

The synonym of attributable risk is "etiologic fraction" and "attributable fraction (exposed)". It is slightly different and is calculated as follows:
attributable fraction (population) = (Io – Iu) / Io,
where Io is incidence (risk) among the total population (the exposed + the unexposed).
The synonym of attributable fraction (population) is "population attributable risk" and "attributable fraction (population)".

The term "probability of causation (PC)" is also used in epidemiology. It is calculated as follows:
PC = (RR-1)/RR = attributable fraction (population).
Strictly speaking, this is not probability of causation. Some epidemiologists suggest to use the term, "assigned share", rather than probability of causation[12].

9. Major study designs in epidemiology

Major study designs of epidemiological studies are as follows:
a) the ecological study,
b) the cohort study, and
c) the case-control study.
The example shown in Figures 1A and 1B is a cohort study. The most important feature of the ecological study is that its fundamental unit of observation is a population, rather than an individual person. On the other hand, the fundamental unit of observation in a case-control study and a cohort study is an individual.

The fundamental unit of observation = population
→An ecological study

The fundamental unit of observation = individual
→A case-control study
A cohort study
A case-control study nested in a cohort
A case-cohort study

II. Introduction to statistical testing

In order to understand what was conducted by an epidemiological study, basic knowledge on statistics is necessary. In this chapter, vary basic descriptions of statistical approaches used in epidemiological studies will be described.

1. *P* value

You might have found statements similar to the following one in a scientific article:
"We measured systolic blood pressures (SBPs) among men and women. Mean SBPs (standards errors) among men and women were 125 (10) and 120 (8) mmHg, respectively. The observed sex difference of SBP was statistically significant since the *P* value obtained from a Student's t-test was 0.03, and was smaller than 0.05,"

You might also have read the statements similar to the following:
"The proportions of smokers among lung cancer cases and healthy control subjects were 50% and 30%, respectively. We conducted a Pearson's chi-square test

and obtained a *P* value smaller than 0.05, and concluded that the case-control difference with respect to the proportion of smokers was statistically significant."

What is the *P* value used in those statistical tests?

Suppose that you flipped a coin for 5 times and you got 5 heads in a row. If the coin is fair, the probability of getting such a result is $0.5^5 = 0.03125$. Is it a *P* value?

Yes, it is a *P* value. More, exactly, it is a one-sided *P* value. It should be noted, however, that you do not usually distinguish "five heads in a row" from "five tails in a row". Therefore, you double the one-sided *P* value of 0.03125 to get the two-sided *P* value of 0.0625. Since this *P* value is not smaller than 0.05, you will accept the null hypothesis. Then, **what is the null hypothesis?** In this scenario, the null hypothesis is that the coin is fair, and that getting heads and tails have the same the probabilities (= 0.5).

Suppose you flip a coin and get 6 heads in a row. In this case, the one-sided *P* value is 0.5^6 and the two-sided *P* value is $2 \times 0.5^6 = 0.5^5 = 0.03125$. Since this *P* value is smaller than 0.05, you reject the null hypothesis and accept the alternative hypothesis, in which the coin is not fair, and, therefore, the probability of getting a head (and a tail) is not 0.5.

Suppose you flip a coin for 6 times and get 5 heads and 1 tails. The probability of getting such a result is 6×0.5^6. Is it a *P* value?

No, it is not a *P* value. Why? The *P* value is not merely the probability of what you observed. It is the sum of the probabilities of what you observed (5 heads and 1 tail) and of what is more extreme (in this case, 6 heads in a row). Therefore, the *P* value (one-sided *P* value) is $6 \times 0.5^6 + 0.5^6 = 7 \times 0.5^6$.

The definition of *P* value is as follows (DE-IV): The probability that a test statistic would be as extreme as observed or more extreme if null hypothesis is true.

2. Normal distribution

The distribution of human body height is known to follow the Normal distribution. Suppose that there are 10 million Japanese men aged 20–39, and they have a mean height of 170 cm and a standard deviation of 5.5 cm. The standard deviation is a parameter reflecting the average differences (in absolute term) from the mean. When the height of a man is expressed as X, the corresponding Normal score (Z score) is (X–170)/5.5. The Normal scores follow the standard Normal distribution, which has the mean of 0 and the standard deviation of 1. In the standard Normal distribution, the men with Normal scores of 1.96 ($170 + 1.96 \times 5.5 = 180.78$cm) or larger account for 2.5% and the men with Normal scores of –1.96 ($170 - 1.96 \times 5.5 = 159.22$cm) or smaller make up 2.5%.

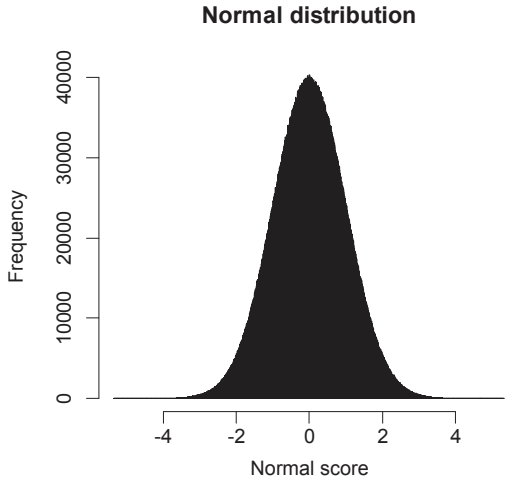

Fig. 3. Normal distribution.
This figure is the histogram of 10 million Normal random values created by R software.

3. Z test and Student's t test

Suppose that you measured indoor radon levels at 12:00 and 24:00 in a house for 1000 days, and calculated their difference for each day. The distribution of indoor radon concentrations is known to be log-normal. Therefore, after log transformation, this distribution becomes Normal. So, you calculated the ratios of measurements at 12:00 and 24:00, and, then, calculated a log-transformed value each of them. Let us call the calculated value as x. The distribution of x is expected to be Normal. In the null hypothesis, the difference is zero, and therefore, you will consider a Normal distribution with the mean value of zero. If you know the standard deviation of this Normal distribution, you can calculate a Normal score for each x. Note that we need to know the standard deviation of the distribution in the theoretical population from which 1000 samples were taken, rather than the standard deviation of the observed data (in statistical testing, we consider the observed dataset as a sample from a theoretical population with an extremely large size). If the mean of Normal scores is 1.96, the corresponding two-sided *P* value is 0.049996. Since this *P* value is less than 0.05, you will reject the null hypothesis, in which the difference is zero. The statistical test described here is sometimes called as the Z test.

Usually, however, we do not know the standard deviation of this Normal distribution (in the theoretical population from which the 1000 samples were taken). Therefore, you have to estimate this value using your own data. In this example, you will estimate it with your data of 1000 measurements. The standard deviation divided by the square-root of the sample size (= 1000) is called a standard error. Using this estimate, you can conduct the Student's t test, in which adjustment for using

the estimated standard deviation is made. The shape of t distribution is similar to the Normal distribution, and the t distribution becomes similar to the Normal distribution when the sample size gets larger.

In the calculation of the standard error, we divided the observed standard deviation by the square root of the number of observations (= 1000 in this example). Why is that?

Before giving an explanation for this, let me introduce the new term, "variance", which is defined as the average of squared difference between each value and its mean. When the value of a standard deviation is S, the variance (V) is S^2. As already shown, when the S is calculated using samples, the standard error is estimated to be $S/n^{0.5}$, where n is the number of observations. Suppose there are two sets of variables X_1, X_2,, X_n and Y_1, Y_2,, Y_n and we define V_x as the variance of Xs (= X_1, X_2,, X_n) and V_y, the variance of Ys. In general, the followings are true:

i) (V_x+V_y) is the variance of (Xs + Ys) – theorem A--;

ii) the variance of (Xs)/n is (Vx)/n^2 – theorem B--.

Let us call the each observed value as x_1, x_2,,, x_n. The number of observation is 1000 (= n) in this example. Here, we assume that x_1, x_2,,, and x_n are taken from theoretical populations, X_1, X_2,,, and X_n, and that each of X_1, X_2,,, and X_n has the mean of M and the standard deviation of S[13]. The mean of X_1, X_2,,, and X_n is (X_1/n + X_2/n,,, + X_n/n). The standard deviation of X_i/n is S/n (, where i = 1,2,,, or n), and the corresponding variance is S^2/n^2 =V/n^2 (derived from theorem B). The variance of M, which is (X_1/n + X_2/n,,, + X_n/n), is the sum of variances of X_i/n over n (derived from theorem B), which is (n times V/n^2) = nV/n^2 = V/n. Since V/n corresponds to $S/n^{0.5}$, the standard error is the standard deviation divided by $n^{0.5}$.

4. Significance levels

In statistical tests, if a *P* value which is calculated on the basis of a null hypothesis is less than 0.05, the null hypothesis will be rejected. The value of 0.05 is sometimes called the significance level. You may wonder why the rejection criterion (the significance level) is 0.05. This value was selected on no theoretical basis but on the tradition and/or experiences. However, based on my experiences, I can tell you that it was not a bad choice.

You may also wonder whether one can change this significance value. You cannot arbitrarily change the significance level. However, you can use a one-sided *P* value rather than two-sided *P* value if you can justify such a choice. In the case of radiation epidemiology, one-sided *P* values are frequently used since scientists are usually interested only in the health risk associated with radiation exposure and any protective effect of health risk is ignored. The significance level of 0.05 for a one-sided test corresponds to the significance level of 0.10 for a two-sided test.

When you examine various cancer sites, you have to carefully interpret the results of statistical testing. If you examine the association of radiation exposure with the risks of 20 different cancer sites, one of them is expected to be significant since 0.05 = 1/20 even if radiation exposure is not related to cancer risk (or randomly related to cancer risk). In such a case, the significance level may be changed. A well-known approach is the Bonferroni correction. Suppose you conducted essentially the same statistical test for n difference cancer sites, in this case, the new significance level can be calculated as $1 – (1 – 0.05)^{1/n}$, which is approximately 0.05/n. This method is known to be most strict.

References

1. WHO. Health topics: Epidemiology [homepage on the internet]. Available from: http://www.who.int/topics/epidemiology/en/
2. Last J (2001) Dictionary of Epidemiology, 4th edition. International Epidemiological Association.
3. Porta M (2008) A Dictionary of Epidemiology.5th edition. Oxford University Press.
4. National Cancer Institute. NCI Dictionary of Cancer Terms: observational study [homepage on the internet] Available from: http://www.cancer.gov/dictionary? CdrID = 286105
5. Brown JR and Thornton JL (1957) Percivall Pott (1714–1788) and Chimney Sweepers' Cancer of the Scrotum. Br J Ind Med. 14(1): 68–70.
6. James Lind: A Treatise of the Scurvy in Three Parts. Containing an inquiry into the Nature, Causes and Cure of that Disease, together with a Critical and Chronological View of what has been published on the subject. Millar, London, 1753.
7. Hempel S (2013) John Snow. The Lancet [serial online]. 381(9874): 1269–1270. Available from: http://www.thelancet.com/journals/lancet/article/PIIS0140–6736(13)60830–2/fulltext
8. Rothman KJ and Greenland S (1998) Modern Epidemiology second edition. Lippincot, Wiliamd and Wilkins.
9. Gardner MJ, Winter PD (1984) Mortality in Cumberland during 1959–78 with reference to cancer in young people around Windscale, Lancet, i, 217.
10. Gardner MJ, et al. (1987) Follow up study of children born to mothers resident in Seascale, West Cumbria (birth cohort), Br Med J 295: 822–827.
11. Gardner MJ, et al. (1987) Follow up study of children born elsewhere but attending schools in Seascale, West Cumbria (schools cohort), Br Med J 295: 819–822.
12. Lagakos SW and Mosteller F. (1986) Assigned shares in compensation for radiation-related cancers. Risk Anal Sep;6(3): 345–357.
13. Kreyszig E (1970) Introductory mathematical statistics. Principles and methods. John Wiley & Sons. Inc. New York.