Simulation study scenario algorithm on gene expression differences of Chickpea data in Indonesia to modeling series with the Bayesian approach

Ani Budi Astuti1,a Solimun2, Adji Achmad Rinaldo Fernandes3 and Nur Iriawan4

1,2,3Department of Statistics, Faculty of Mathematics and Natural Sciences, University of Brawijaya Malang, Jl. Veteran Malang 65145 Indonesia

4Department of Statistics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Sepuluh Nopember Surabaya Kampus ITS Sukolilo, Surabaya 60111 Indonesia

Email: aani_budi@ub.ac.id

Abstract. Data on gene expression differences of chickpeas in Indonesia in healthy and diseased conditions was the result of microarray experiments. The diseased condition in chickpeas are caused by Ascochyta Rabiei pathogenic fungi. To learn more details about the gene expression differences data of chickpeas in Indonesia, especially regarding the distribution and fit modeling of the data, a specific simulation scenario is needed in order to obtain an overview of these data. The purpose of this study is to obtain the right simulation scenario algorithm for data on differences in gene expression of chickpeas in Indonesia in a series of modeling activities using the Bayesian approach. The results showed that with certain simulation scenarios obtained simulation data of gene expression differences of chickpeas in Indonesia that are similar to the original data. In this scenario repetition of 1,000 times with each repetition generated a data sample as many as $3n$ according to the number of samples in the original data. These simulation scenario has specific characteristics because of the strict limitations of the meaning of the data on the gene expression differences that are generated. In these simulation scenario, the value of largest tolerance limit is 0.0999.

Keywords: Simulation Study Scenario Algorithm, Bayesian Modeling Series, Gene Expression, Chickpea Data, Indonesia

1. Introduction

The series of processes of translating genetic information (in the form of base sequences on DNA or RNA) into proteins which then become phenotypes is called gene expression. According to [1] and [2], the meaning carried by genetic material for organisms must be expressed in the form of phenotypes. These gene expression can simultaneously be monitored using an analysis technique called microarray. Therefore, gene expression data is data acquired from microarray experiments.

The main challenge of statistical analysis for the results of the microarray experiments data is the small sample size available and has a complex distribution (multimodal distribution). This
challenge can be overcome by Bayesian analysis which is a statistical analysis method that does not consider the sample size in the data and for any distribution, so this analysis can be used for data that has small or large sample sizes and for any distribution of the data [3]. According to [4] and [5], the advantage of Bayesian analysis is that these methods can still provide inference to an unknown parameter based on posterior distribution while maintaining data conditions as they are through data driven concepts.

Chickpea plants commonly referred to in Indonesia as Arabian bean plants have many nutritional benefits for the human body. In addition, this plant roots can bind nitrogen from the air. Therefore, the production of Chickpea plants needs to be increased so that the price is not too expensive. This plant is often attacked by pathogenic fungi Ascochyta Rabiei so that its production can be very sharply reduced [6]. Modeling based on application data of gene expression differences of Chickpea plants in Indonesia with Bayesian approach has been carried out by Astuti et al. ([7] and [8]).

Data simulation is data that is generated according to the original data pattern [9]. Studying the data characteristics of differences in gene expression of Chickpea plants in Indonesia as a result of microarray experiments healthy and diseased conditions in simulation is very important. Through simulation data will be obtained more detailed and general information about the distribution of data and then this distribution information is used for data modeling, considering the results of the microarray experiment data have a very small sample size (n = 3) with complex distribution (multimodal distribution). In addition, given the strong and detailed limitations of the data characteristics of gene expression differences are related to the meaning of the data, so a certain scenario is needed to generate the simulation data. In this study will be studied about simulation data scenarios algorithm from gene expression differences data on healthy and diseased conditions from Chickpea plants in Indonesia from the results of Harijati's research in 2007 [6].

2. Material and methods

Data generation scenarios of gene expression differences in healthy and diseased conditions of Chickpea plants in Indonesia are based on learning data on gene expression differences in healthy and diseased conditions of Chickpea plants in Indonesia from the research of Harijati in 2007. The data studied are data on gene expression differences in healthy and diseased conditions for gene IDs that have defence functions as many as 15 gene IDs and energy functions as many as 11 gene IDs. Data generation is performed 1,000 times with each repeat generating three data.

Data on gene expression differences in this study are assumed to follow the unimodal normal distribution with the mean value and standard deviation according to the original data. The choice of normal distribution is based on previous studies on gene expression differences data and for simplify of distribution approaches. The existence of strict restrictions in generating data on differences in expression of these genes must be found a certain tolerance value, so that the data pattern still follows the original data and the meaning of the data does not change. To determine whether the original data patterns and meanings are followed by generated simulation data by the algorithm, the test is used the goodness of fit Kolmogorov-Smirnov (KS) with $\alpha = 5\%$. The software used in these research is R software.

In Figure 1.1 presented the results of data exploration of gene expression differences at ID gene of healthy and diseased conditions, on the ID gene of the defence function and energy function of Chickpea plants in Indonesia as the result of original data from Harijati’s research in 2007 [6].
This section will explain the theoretical studies that underlie the research to be conducted. First about the simulation Modeling, second about Bayesian Analysis, third about Unimodal Normal Distribution and fourth about Goodness of Fit Kolmogorov-Smirnov (KS).

3.1. Simulation modeling

According to [9], systems in the real world are often complicated to learn to get realistic models to be evaluated analytically. To do this, a simulation scenario is needed. Imitating or simulation activities from the original data requires a computer to do it. In a simulation work system, to learn it scientifically often has to make a set of assumptions about how it works. These assumptions, which are usually mathematical or logical relationships, are the models used to try to get some understanding of how the appropriate system behaves.

The category of the system is divided into two categories, namely the discrete system and the continuous system. Discrete systems are systems in which the variables studied can change instantly where points are separated in time. On the other hand, a continuous system is a system whose variables can change continuously with time [9]. The following is presented in Figure 1.2 which illustrates about maps out different ways in which a system might be studied [9].

![Figure 1](image_url)

**Figure 1.** Data exploration of gene expression differences to defence function (a) and energy function (b) for original data as the result of Harijati’s research in 2007 [6]
3.2. Bayesian analysis

Bayes’ theorem found by Thomas Bayes in 1702-1761. These theorem is used as a basic concept in Bayesian analysis. In this analysis, the model parameters $\Theta$ is seen as a random variable in the parameter space $\Omega$. Bayesian analysis is a statistical analysis method based on the posterior probability distribution model. Posterior distribution is a combination of two information, namely past data (prior information) and sample observation data (likelihood information) ([10]; [11]; [12]; [13] and [14]).

The concept of Bayesian analysis can be explained as follows: if known an observation data $x$ that has a likelihood function $f(x|\Theta)$, then the information about the parameter $\Theta$ before the observation is done is called a prior $\Theta$, that is $p(\Theta)$. Furthermore, to determine the posterior distribution of the $\Theta$, that is $p(\Theta|x)$ are as follows:

$$p(\Theta|x) = \frac{f(x|\Theta)p(\Theta)}{f(x)} \quad (1)$$

where

$$f(x) = \begin{cases} \int_{\Theta \in R} f(x|\Theta)p(\Theta)d\Theta, & \text{when } \Theta \text{ is continuous,} \\ \sum_{\Theta \in B} f(x|\Theta)p(\Theta), & \text{when } \Theta \text{ is discrete,} \end{cases}$$

and $f(x)$ is a total probability and represents a normalized constant [12], so that equation (1) can be rewritten as equation (2).

$$p(\Theta|x) \propto f(x|\Theta)p(\Theta) \quad (2)$$
3.3. Unimodal normal distribution

Normal distribution was first introduced by Abraham DeMoivre (1733) as a binomial distribution approach for large $n$. Normal distribution is also called the Gauss distribution because this equation of distribution was discovered by Gauss (1777-1855). Unimodal Normal Distribution is one continuous distribution that characterized by two parameters, namely the mean $(\mu)$ and variance $(\sigma^2)$ and this distribution has one peak (unimodal). If a random variable $x$ has a unimodal normal distribution, it is generally written with symbols $x \sim N(\mu, \sigma^2)$. The shape of the curve from a normal distribution is like an armrest that extends infinitely in both its positive and negative directions. The general formula of unimodal normal distribution is as follows ([15] and [16]):

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2} \text{ where } -\infty < x < \infty, -\infty < \mu < \infty \text{ and } \sigma > 0$$ (3)

The shape of the unimodal normal distribution can be described as follows:

![Figure 3. Unimodal Normal Distribution with Two Parameters $\mu$ and $\sigma^2$](image)

3.4. Goodness of fit Kolmogorov-Smirnov (KS)

The Kolmogorov-Smirnov (KS) test is one of the goodness of fit tests that can be used to determine whether a continuous data follows a particular distribution or whether a presumed distribution actually matches the observed data [13]. The theoretical concept of the KS test is to compare the empirical cumulative (CDF) distribution function, namely $F_n(x_i)$ with hypothesis cumulative distribution function (prediction CDF), namely $\hat{F}(x)$. If $X_1, X_2, \ldots, X_n$ is the statistical order of independent random variables with the cumulative distribution of hypotheses $\hat{F}(x)$ then the empirical (observation) cumulative distribution can be defined in equation (4).

$$F_n(x_i) = \frac{\text{the number of data } X_i \leq x_i}{n} \text{ for } i = 1, 2, 3, \ldots, m < n$$ (4)

where is $F_n(x_i)$ the right continuous step function. Furthermore, each value $F_n(x_i)$ will be compared with the distribution value of the hypothesis (prediction) $\hat{F}(x)$, so that the statistical tests is used in the
Kolmogorov-Smirnov test is the largest vertical distance (maximum) between $F_n(x_i)$ and $\hat{F}(x)$, that notated with $D_n$ ([13] and [17]).

Hypothesis testing used in the Kolmogorov-Smirnov theory is as follows:

$H_0$: Data X is an independent random variable that is distributed according to the distribution $\hat{F}(x)$ versus
$H_1$: Data X is an independent random variable that is distributed not in accordance with the distribution $\hat{F}(x)$

Rejection of the hypothesis is done using $D_n$ test statistics which can be written in equation (5).

\[ D_n = \sup \left| F_n(x) - \hat{F}(x) \right| \] (5)

If the value of $D_n > d_n$ or P-value <0.05, then the null hypothesis ($H_0$) is rejected and the other hand if $D_n \leq d_n$ or P-value $\geq$ 0.05, then the null hypothesis ($H_0$) fails to be rejected. The value of $d_n$ is the reference value obtained from the Kolmogorov-Smirnov table.

In this study, $F_n(x_i)$ is the cumulative distribution of original data while $\hat{F}(x)$ is the cumulative distribution of simulation data generated. The following is presented the concept from the goodness of fit test theory of Kolmogorov-Smirnov.

4. Result and Discussion

In this section, the results and discussion of this study will be presented.
4.1. Algorithm of data generation scenarios

There are five steps in the simulation data generation scenario algorithm generated. This algorithm is referred to as Algorithm 1.1 in this study. The five steps are as follows:

Algorithm 1.1. Algorithm of Data Generation Scenarios

Step 1. View \( x \) data as a random variable with \( x \) as univariate observation from Chickpea data with sample size \( n = 3 \), so the observation data is \( x_1, x_2, x_3 \).

Step 2. Sort the observation data so that it can be obtained \( x_{(1)}, x_{(2)}, x_{(3)} \).

Step 3. Calculate the average \( (\bar{x}) \) and standard deviation \( (s_x) \) from data \( x \).

Step 4. Set the difference tolerance \( (\varepsilon) \) amounting to 0.0999 (determination by trial) so that the simulation data obtained remains compatible with the original data. Determination of this tolerance value is done by calculating the absolute difference between the Chickpea data that has been ordered \( x_{(i)} \) and simulation data that has been sorted \( \hat{x}_{(i)} \), that is \( |x_{(i)} - \hat{x}_{(i)}| < \varepsilon \). If the difference value \( (\varepsilon) \) too big or too small the simulation data will deviate from the original data.

Step 5. Perform data generation in the following ways:

- For repetition \( i = 1 \) or \( i = 3 \), then
  1. Generated \( \hat{x}_{(i)} \sim N(\bar{x}, s_x) \)
  2. Calculate the difference tolerance \( |x_{(i)} - \hat{x}_{(i)}| < \varepsilon \)
  3. Repeat generation on (1) until conditions are reached \( \hat{x}_{(1)} \geq x_{(1)} \) and \( \hat{x}_{(3)} \leq x_{(3)} \) and \( |x_{(i)} - \hat{x}_{(i)}| < \varepsilon = 0.0999 \) can be fulfilled.

- For repetition \( i = 2 \), then
  1. Generated \( \hat{x}_{(i)} \sim N(\bar{x}, s_x) \)
  2. Calculate the difference tolerance \( |x_{(i)} - \hat{x}_{(i)}| < \varepsilon \)
  3. Repeat generation on (1) until conditions are reached \( |x_{(i)} - \hat{x}_{(i)}| < \varepsilon = 0.0999 \) can be fulfilled.

Based on Algorithm 1.1, it can be seen that the tolerance value obtained is 0.0999 which with this tolerance value obtained simulation data that is compatible with the original data. In this study, data generation was carried out 1,000 times.

4.2. Data generation simulation

Simulation data as generated data based on Algorithm 1.1 can be shown in Table 1.1 for the defence function of gene IDs and Table 1.2 for the energy function of gene IDs. There are 15 gene IDs with defence function and there are 11 gene IDs with energy function. The data presented are data on differences in gene expression which is the ratio of the expression value of diseased condition genes to the healthy expression values of genes namely: \( \ln \left( \frac{\text{value of diseased genes expression}}{\text{value of healthy genes expression}} \right) \). The results obtained can be three groups, namely negative (-), zero and positive (+). The definition of each group of gene IDs can be described as follows: (1) Negative value (-): the value of the gene expression in a healthy condition is greater than the diseased condition. This means that the gene IDs in this group is included in the Up-regulated group. In this condition, the gene ID is resistant to attack by pathogenic fungus Ascochyta Rabiei and can significantly increase the production of Chickpeas. (2) Positive
value (+): the value of the expression of the gene of the diseased condition is greater than the healthy condition. This means that gene IDs in this group belong to the Down-regulated group. In this condition, gene ID is not very resistant to attack by pathogenic fungus Ascochyta Rabiei and can significantly reduce the production of Chickpea plants. (3) Zero value (0): the expression value of the diseased condition is the same or relatively the same as the healthy condition. This means that the gene ID in this group is in the Regulated group. In this condition, the gene ID is quite resistant to attack by pathogenic fungus Ascochyta Rabiei, but it does not significantly affect the production of Chickpea plants.

4.3. Goodness of fit Kolmogorov-Smirnov test

The suitability of the results of simulation data generation with the original data was tested using the Kolmogorov-Smirnov goodness of fit test method. Test results are presented in Table 1.3 and Table 1.4. Based on the results of the Goodness of fit test Kolmogorov-Smirnov in the Table 1.3 and Table 1.4 can be shown that for all gene IDs, both the resistance function gene IDs and the energy function gene IDs have P-value average of > 0.05. This means that there is a suitability of the generated data pattern (simulation data) with the original data. Therefore it can be shown that Algorithm 1.1. as a simulation data generation scenario algorithm for Chickpea data has been able to generate data in accordance with the Chickpea data pattern. This test result has a test error rate of 5%.

| No | Defence Function of Gene IDs | Repetition | Average |
|----|------------------------------|------------|---------|
|    |                              | U1        | U2      | U3       |         |
| 1  | LS0024                       | -0.6241   | -0.4632 | -0.0880  | -0.3918 |
| 2  | LS0035                       | -0.6251   | -0.4767 | -0.3778  | -0.4932 |
| 3  | LS0616                       | -0.8970   | -0.3885 | -0.4253  | -0.5703 |
| 4  | LS0159                       | -1.1179   | -0.6544 | -0.5298  | -0.7674 |
| 5  | LS0162                       | -0.6988   | -0.5336 | -0.0943  | -0.4422 |
| 6  | LS0185                       | -0.6350   | -0.4998 | -0.1008  | -0.4119 |
| 7  | LS0612                       | -1.2127   | -1.2197 | -1.2122  | -1.2127 |
| 8  | LS0081                       | -0.8875   | -0.8623 | -0.4046  | -0.7181 |
| 9  | LS0752                       | -1.0311   | -1.0316 | -0.6381  | -0.6876 |
| 10 | LS0759                       | -0.7772   | -0.7728 | -0.7785  | -0.7762 |
| 11 | LS0688                       | -1.0795   | -0.9007 | -0.4983  | -0.3599 |
| 12 | U070                         | -1.4864   | -0.7061 | -0.2483  | -0.4955 |
| 13 | U273                         | -1.8522   | -1.3727 | -1.3413  | -1.5221 |
| 14 | U278                         | -0.6743   | -0.4831 | -0.4098  | -0.5224 |
| 15 | U017                         | -0.3296   | -0.2470 | -0.2276  | -0.2681 |
Table 2. Average simulation data from generation 1,000 times each repetition for the energy function of Gene IDs

| No | Energy Function of Gene IDs | Repetition | Average |
|----|-----------------------------|------------|---------|
|    |                             | U1        | U2      | U3      |         |
| 1  | LS0132                      | -0.5616   | -0.5307 | -0.4781 | -0.5235 |
| 2  | LS0341                      | -1.1792   | -1.1134 | -1.0687 | -1.1204 |
| 3  | U001                        | -0.8628   | -0.5305 | 0.0158  | -0.4592 |
| 4  | U325                        | -1.4493   | -1.3215 | -0.6623 | -0.9236 |
| 5  | U385                        | -0.8709   | -0.4621 | -0.3058 | -0.5463 |
| 6  | U447                        | -0.7124   | -0.6218 | -0.0870 | -0.4737 |
| 7  | CA0426                      | -1.0557   | -0.8208 | -0.0495 | -0.3519 |
| 8  | CA0534                      | -0.7974   | -0.7974 | -0.6152 | -0.7367 |
| 9  | CA0702                      | -0.9262   | -0.8219 | -0.4988 | -0.7490 |
| 10 | CA0810                      | -0.8932   | -0.6565 | -0.4592 | -0.6696 |
| 11 | CA1120                      | -0.7348   | -0.7129 | 0.3356  | -0.3707 |

Table 3. Result of Kolmogorov-Smirnov's goodness of fit test between Chickpea data and simulation data for defence function Gen IDs

| No | Defence Function of Gene IDs | n | Data Scenarios Built according to Chickpea Data Patterns | P-value Average |
|----|------------------------------|---|-------------------------------------------------------|----------------|
| 1  | LS0024                       | 3 | Two Components Normal Mixture Indication               | 0.996          |
| 2  | LS0035                       | 3 | Unimodal Normal Indication                             | 0.996          |
| 3  | LS0616                       | 3 | Two Components Normal Mixture Indication               | 0.518          |
| 4  | LS0159                       | 3 | Two Components Normal Mixture Indication               | 0.996          |
| 5  | LS0162                       | 3 | Two Components Normal Mixture Indication               | 0.996          |
| 6  | LS0185                       | 3 | Two Components Normal Mixture Indication               | 0.996          |
| 7  | LS0612                       | 3 | Unimodal Normal Indication                             | 0.518          |
| 8  | LS0081                       | 3 | Two Components Normal Mixture Indication               | 0.996          |
| 9  | LS0752                       | 3 | Two Components Normal Mixture Indication               | 0.518          |
| 10 | LS0759                       | 3 | Unimodal Normal Indication                             | 0.518          |
| 11 | LS0688                       | 3 | Two Components Normal Mixture Indication               | 0.996          |
| 12 | U070                         | 3 | Three Components Normal Mixture Indication             | 0.996          |
| 13 | U273                         | 3 | Two Components Normal Mixture Indication               | 0.996          |
| 14 | U278                         | 3 | Unimodal Normal Indication                             | 0.996          |
| 15 | U017                         | 3 | Unimodal Normal Indication                             | 0.996          |
Table 4. Result of Kolmogorov-Smirnov's goodness of fit test between Chickpea data and simulation data for energy function Gen IDs

| No | Energy Function of Gene IDs | n | Data Scenarios Built according to Chickpea Data Patterns | P-value Average |
|----|-----------------------------|---|--------------------------------------------------------|-----------------|
| 1  | LS0132                      | 3 | Unimodal Normal Indication                              | 0.996           |
| 2  | LS0341                      | 3 | Unimodal Normal Indication                              | 0.996           |
| 3  | U001                        | 3 | Three Components Normal Mixture Indication              | 0.996           |
| 4  | U325                        | 3 | Two Components Normal Mixture Indication                | 0.996           |
| 5  | U385                        | 3 | Two Components Normal Mixture Indication                | 0.996           |
| 6  | U447                        | 3 | Two Components Normal Mixture Indication                | 0.996           |
| 7  | CA0426                      | 3 | Two Components Normal Mixture Indication                | 0.996           |
| 8  | CA0534                      | 3 | Two Components Normal Mixture Indication                | 0.518           |
| 9  | CA0702                      | 3 | Two Components Normal Mixture Indication                | 0.996           |
| 10 | CA0810                      | 3 | Three Components Normal Mixture Indication              | 0.996           |
| 11 | CA1120                      | 3 | Two Components Normal Mixture Indication                | 0.996           |

5. Conclusion

In this study, the data distribution approach used is a unimodal normal distribution with a mean and standard deviation values according to the original data. The simulation data generation scenario algorithm has been successfully carried out with the same results as the original data with a tolerance value of 0.0999. Based on the results of the goodness of fit test Kolmogorov-Smirnov (KS) shows that all simulation data generated can be in accordance with the original data pattern where the data distribution is indicated to have a normal mixture distribution in some gene IDs in Chickpea plants in Indonesia. The goodness of fit test of Kolmogorov-Smirnov (KS) uses an $\alpha$ value of 5%.

Acknowledgement

This paper is part of research Grants the DPP/SPP and Doctoral Program Activities University of Brawijaya in 2018. We would like the first thanks to University of Brawijaya which have financed this research and the second thanks to Harijati for allowing use of research data and thank to anonymous reviewer to this paper.

References

[1] Baggerly, K. A., Coombes, K. R., Hess, K. R., Stivers, D. N., Abruzzo, L. V. and W., Z. 2001. Identifying Differentially Expressed Genes in cDNA Microarray Experiments. Journal Computational Biology. 8: 639–659.
[2] Knudsen, S. 2004. A Guide to Analysis of DNA Microarray Data. Second Edition. John Wiley & Sons, Inc. (Canada: New Jersey).
[3] Gosh, J. K., Delampady, M. and Samanta, T. 2006. An Introduction to Bayesian Analysis Theory and Method. (New York: Springer).
[4] Mengersen, K. 2009. Module 1 Bayesian Analysis, Short Course on Bayesian Modeling. Department of Statistics. (Surabaya: ITS).
[5] Iriawan, N. 2012. Modeling and Analysis of Data-Driven. (in Indonesian) Vol.1 (Surabaya: ITS).

[6] Harijati, N. 2007. A Study of the Resistance of Chickpea (Cicer Arietinum) to Ascochyta Rabiei and the Effect of Age of Plant Tissue on Disease Development. Ph.D. Thesis. (Australia: La Trobe University).

[7] Astuti, A. B., Iriawan, N., Irhamah and Kuswanto, H. 2015. An Algorithm for Determining the Number of Mixture Components on the Bayesian Mixture Model Averaging for Microarray Data. Journal of Mathematics and Statistics, 11 (2). pp. 45-51, DOI: 10.3844/jmssp.2015.45.51.

[8] Astuti, A. B., Iriawan, N., Irhamah and Kuswanto, H. 2017. Bayesian Mixture Model Averaging (BMMA) for Identifying the Differences Genes Expression of Chickpea (Cicer Arietinum) Plant Tissue, Communications in Statistics-Theory and Methods. 46 (21). pp. 1-18, DOI: http://dx.doi.org/10.1080/0361092016.1239112.

[9] Law, A. M. and Kelton, W. D. 1991. Simulation Modeling and Analysis. Second Edition. McGraw-Hill, Inc. New York.

[10] Box, G. E. P. and Tiao. 1973. Bayesian Inference in Statistical Analysis. (Boston MA: Addison-Wesley Massachusetts).

[11] Zellner, A. 1971. An Introduction to Bayesian Inference in Econometrics. (New York: John Wiley).

[12] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. 1995. Bayesian Data Analysis. (London: Chapman & Hall).

[13] Iriawan, N. 2003. Teaching Modules of Simulation Technique (in Indonesian). (Surabaya: ITS).

[14] Congdon, P. 2006. Bayesian Statistical Modelling. 2nd. (USA: John Wiley & Sons).

[15] Casella, G. B. and Roger, L. 2001. Statistical Inference. 2nd ed. Duxbury. ISBN 0-534-24312-6.

[16] Wlodzimierz, B. 1995. The Normal Distribution: Characterizations with Applications. (Springer-Verlag). ISBN 0-387-97990-5.

[17] Sheskin, D. J. 2007. Handbook of Parametric and Nonparametric Statistical Procedures, 4th edition. Chapman & Hall/CRC.