Distribution Testing On The Average Room Occupancy Rate Of Hotels By Province In 2013-2017

Khawarita Siregar¹, Aulia Ishak², Marcelina Esterlita³

¹,²,³Department of Industrial Engineering, Faculty of Engineering, Universitas Sumatera Utara, Medan 20155, Indonesia

E-mail: marcelinaesterlitabarus@gmail.com khawarita@usu.ac.id aulia.ishak@gmail.com

Abstract. Data on Average Room Occupancy Rate of Hotels by Province in 2013-2017 obtained from the Central Statistics Agency. The Average Data of Hotel Room Occupancy Rate by Province in 2013-2017 is continuous data. This data will be tested using 4 types of continuous distribution data, normal, exponential, lognormal, and gamma distribution. This data processing uses Minitab software to determine the opportunity value of BKA and BKB. This distribution test uses a hypothesis where if there is no data in accordance with the distribution used, further data processing will be carried out to determine what type of distribution is in accordance with the continuous data using EasyFit Software. Meanwhile, if the data matches more than 1 distribution, the coefficient of determination testing will be done to find out which distribution is the closest to the distribution test results.

1. Introduction

Data is a collection of information or values obtained from observations (observations) of an object, data can be numbers and can also be symbols or characteristics. Several kinds of data, among others; population data and sample data, observation data, primary data, and secondary data.[1]

Statistics broadly means a science that studies how to collect, process / group, present and analyze data and how to draw conclusions in general based on the results of research that is not comprehensive.[2]

Statistics is essentially concerned with the presentation and interpretation of events that are opportunities that occur in a planned investigation or scientific research. Opportunity theory is a branch of mathematics concerned with probability analysis and random phenomena. The opportunity distribution is defined by an opportunity function, denoted by p (x) or f (x), which shows the probability of each random variable value. The main object of opportunity theory is random variables, stochastic processes, and events are non-deterministic mathematical abstractions of events or measurable quantities that can be single events or develop over time in apparently random modes. Important characteristics of a number of data can be identified immediately by grouping the data into several classes, and then counting the number of observations that enter each class. This arrangement, in tabular form, is called frequency distribution[3].

Before a particular probability distribution model is applied to fit real world data, it is essential to confirm whether the given probability distribution satisfies the underlying requirements of its characterization. Thus, characterization of a probability distribution plays an important role in
statistics and mathematical sciences. A probability distribution can be characterized through various methods, see, for example, Ahsanullah, Kibria, and Shakil, Huang and Su, Nair and Sudheesh, Nanda, Gupta and Ahsanullah, and Su and Huang, among others. In recent years, there has been a great interest in the characterizations of probability distributions by truncated moments. For example, the development of the general theory of the characterizations of probability distributions by truncated moment began with the work of Galambos and Kotz. Further development on the characterizations of probability distributions by truncated moments continued with the contributions of many authors and researchers, among them Kotz and Shanbhag, Glänzel, and Glänzel, Telcs, and Schubert, are notable. However, most of these characterizations are based on a simple proportionality between two different moments truncated from the left at the same point. It appears from the literature that there has been little attention to characterizing a continuous distribution using truncated moments. As pointed out by Glänzel these characterizations may also serve as a basis for parameter estimation. In this paper, some new characterizations of continuous distributions by truncated moment have been established [4].

Testing of continuous distribution can be done using data from the Central Statistics Agency (BPS), which is the Average Occupancy Rate of Hotel Rooms by Province in 2013-2017.

2. Literature Review

2.1. Normal Distribution

The most important continuous opportunity distribution in all fields of statistics is the normal distribution. The graph on this distribution called the normal curve is a bell-shaped curve like Figure 1. The normal distribution is often called the Gauss distribution in honor of Gauss, who also managed to get his equation from the study of errors in repeated measurements of the same object:

![Figure 1. Normal distribution curve](image)

A continuous random variable $X$ that has a bell-shaped distribution is called a normal random variable. The mathematical equation for the probability distribution of this normal random variable depends on two parameters $\mu$ and $\sigma$, which are the mean and the standard deviation. Therefore symbolize the value of the concentration function for this $X$ with $n(x; \mu, \sigma)$.

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad \text{for } -\infty < x < \infty$$

Whereas in this case $\pi = 3.14159 \ldots$ and $e = 2.71828 \ldots$

If the values of $\mu$ and $\sigma$ are known, then the normal curve is certain. For example, if $\mu = 50$ and $\sigma = 5$, then the ordinate $n(x; 50, 5)$ can easily be calculated for various values of $x$, and then the curve can be drawn.

From observations, we obtain the properties of normal curves as follows:

- The mode is the point on the horizontal axis that makes the function reach its maximum occur at $x = \mu$.
- The curve is close to a vertical line through the middle value $\mu$.
- This curve approaches the horizontal axis asymptotically in both directions if we follow the middle value.
2.2. Exponential Distribution

Exponential distribution is a special case of gamma distribution with form factor $\alpha = 1$ and scale factor $\beta = 1 / \alpha$. This distribution is widely used as a model in engineering and science. Substituting these values can be seen that the continuous random variable $X$ has an exponential distribution with the parameter $\lambda$ where $\lambda > 0$, then the probability density function of $x$ is:

$$f_{e}(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{other} \end{cases}$$

(2)

Figure 2. The exponential distributed probability density function

2.3. Lognormal Distribution

Lognormal distribution is a theoretical distribution that is widely used in engineering, especially as a model for various types of material properties.

A non-negative continuous variable $X$ is said to have a lognormal distribution if $\ln (X)$ has a normal distribution. The probability density function of a random variable that satisfies the lognormal distribution if $\ln (X)$ is normally distributed with parameters $\mu$ and $\sigma$ are:

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln(x) - \mu)^2}{2\sigma^2}} \quad x \geq 0$$

(3)

Figure 3. Lognormal distribution probability density function

2.4. Gamma Distribution

The Gamma Distribution is an alternative model that is widely used for probability experiments whose results show a form of distribution that has a significant variation in the size of the slope.

The gamma function is defined by:

$$\Gamma(\alpha) = \int_{0}^{\infty} x^{\alpha-1} e^{-x} \, dx$$

(4)
A continuous variable X is said to have a gamma distribution, with the shape parameter $\alpha$ and the scale parameter $\beta$, where $\alpha > 0$ and $\beta > 0$ the probability density function of x is:

$$f_{G}(x,\alpha,\beta)=\frac{1}{\beta^{\alpha}}\frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)} \quad x>0$$

(5)

2.5. Frequency Distribution
The recorded information that is collected in accordance with the original, both in the form of calculations and measurements, is called raw data. The raw data varies so that it is difficult to interpret, interpret and conclude. Therefore, the raw data needs to be presented in forms that are easily responded to, such as tabulations and diagrams. Another alternative that can be used to present data so that it is easier to respond is to compile the raw data encountered in the form of frequency and leaf-leaf distribution. The frequency distribution consists of frequency tables, frequency histograms, frequency polygons and ogives.

According to H.A. Sturges, the formula for determining the number of classes is as follows:

$$k = 1 + 3.3 \log n$$

(6)

The class interval or class width is the same for each class. The more number of classes means the smaller the class interval and vice versa.

To determine the class size (interval length) a formula is used [7]:

$$C=\frac{X_{m}-X_{1}}{k}$$

(7)

2.6. Distribution Testing
Statistical hypotheses are statements or conjectures about one or more populations. True or false population will never be known with certainty, unless the entire population is examined. Of course in most situations this is not possible. Therefore a random sample can be drawn from the population and use the information contained in that example to decide whether the hypothesis is most likely true or false.

The hypothesis formulated in the hope that it will be rejected will lead to the use of the term null hypothesis which is denoted by $H_0$. The rejection of $H_0$ will result in the acceptance of an alternative hypothesis denoted by $H_1$. The null hypothesis regarding a population parameter must be pronounced in such a way as to state with certainty a value for that parameter, while the alternative hypothesis allows several possible values. So if $H_0$ states the null hypothesis that $p = 0.5$ for a binom population, then the alternative hypothesis $H_1$ can be $p > 0.5$, $p < -0.5$ or $p \neq 0.5$ [8].

2.7. P-Value
P Value is the magnitude of the opportunity to make a mistake if you decide to reject $H_0$. In general the p-value is compared with a certain significance level $\alpha$, usually 0.05 or 5%. The real level $\alpha$ is interpreted as an opportunity to make a mistake to conclude that $H_0$ is wrong, when in fact the statement $H_0$ is correct [9].

3. Research Methodology
Secondary data is data obtained from the results of existing research or experiments. Secondary data obtained from continuous distribution taken from the Central Statistics Agency (BPS), which is about the average Occupancy Rate of Hotel Rooms by Province in 2013-2017. Continuous distribution testing is performed using chi-square testing. Chi-square testing is used to test whether the observed frequency deviates significantly from the expected frequency distribution or not. If the chi-square test is accepted, then proceed with the P-Value. However, if it is rejected, it is not necessary to determine the P-Value. The data that has been obtained is processed by testing the data according to the type of distribution. Data with a discrete distribution were tested using a single Chi-Square test and the P value test. Data with continuous distributions were tested using the Chi Square test group, such as normal distribution, gamma distribution, exponential distribution, lognormal distribution.
4. Result
The steps for testing the continuous distribution of the Central Statistics Agency (BPS) data on the average hotel room occupancy rate by province (percent) in 2013-2017 are as follows:

- Sorting Data

| No | Data | No | Data | No | Data | No | Data | No | Data |
|----|------|----|------|----|------|----|------|----|------|
| 1. | 31,40| 29. | 43,10| 57. | 47,60| 85. | 50,00| 113. | 54,20| 141. | 58,35|
| 2. | 34,40| 30. | 43,40| 58. | 47,91| 86. | 50,10| 114. | 54,30| 142. | 58,60|
| 3. | 35,80| 31. | 44,10| 59. | 48,00| 87. | 50,39| 115. | 54,30| 143. | 59,03|
| 4. | 37,50| 32. | 44,10| 60. | 48,20| 88. | 50,40| 116. | 54,47| 144. | 59,06|
| 5. | 37,50| 33. | 44,60| 61. | 48,20| 89. | 50,60| 117. | 54,67| 145. | 59,10|
| 6. | 37,60| 34. | 44,70| 62. | 48,20| 90. | 50,60| 118. | 54,80| 146. | 59,20|
| 7. | 38,40| 35. | 45,10| 63. | 48,40| 91. | 50,70| 119. | 55,03| 147. | 59,30|
| 8. | 38,41| 36. | 45,30| 64. | 48,50| 92. | 50,80| 120. | 55,10| 148. | 59,40|
| 9. | 38,70| 37. | 45,70| 65. | 48,57| 93. | 51,00| 121. | 55,22| 149. | 59,60|
| 10. | 39,20| 38. | 45,75| 66. | 48,58| 94. | 51,10| 122. | 55,60| 150. | 59,70|
| 11. | 39,20| 39. | 45,90| 67. | 48,60| 95. | 51,10| 123. | 55,70| 151. | 59,90|
| 12. | 39,37| 40. | 46,00| 68. | 48,70| 96. | 51,60| 124. | 55,80| 152. | 60,16|
| 13. | 39,40| 41. | 46,10| 69. | 48,80| 97. | 51,80| 125. | 56,08| 153. | 60,50|
| 14. | 39,50| 42. | 46,10| 70. | 48,80| 98. | 52,30| 126. | 56,10| 154. | 60,56|
| 15. | 40,00| 43. | 46,10| 71. | 48,80| 99. | 52,60| 127. | 56,10| 155. | 60,80|
| 16. | 40,20| 44. | 46,19| 72. | 49,17| 100. | 52,70| 128. | 56,50| 156. | 60,90|
| 17. | 40,20| 45. | 46,20| 73. | 49,20| 101. | 52,70| 129. | 56,54| 157. | 61,10|
| 18. | 40,40| 46. | 46,20| 74. | 49,20| 102. | 52,80| 130. | 56,68| 158. | 61,20|
| 19. | 40,60| 47. | 46,40| 75. | 49,20| 103. | 52,84| 131. | 56,70| 159. | 61,70|
| 20. | 40,80| 48. | 46,60| 76. | 49,30| 104. | 52,90| 132. | 56,80| 160. | 62,50|
| 21. | 41,30| 49. | 46,60| 77. | 49,40| 105. | 53,20| 133. | 57,10| 161. | 62,62|
| 22. | 41,40| 50. | 46,60| 78. | 49,40| 106. | 53,40| 134. | 57,10| 162. | 64,24|
| 23. | 41,70| 51. | 46,80| 79. | 49,40| 107. | 53,70| 135. | 57,30| 163. | 67,66|
| 24. | 41,80| 52. | 46,90| 80. | 49,40| 108. | 53,80| 136. | 57,30| 164. | 69,80|
| 25. | 41,80| 53. | 46,90| 81. | 49,50| 109. | 53,90| 137. | 57,70| 165. | 71,10|
| 26. | 42,00| 54. | 47,24| 82. | 49,80| 110. | 53,90| 138. | 57,74| | |
| 27. | 42,80| 55. | 47,43| 83. | 49,90| 111. | 54,00| 139. | 57,90| | |
| 28. | 42,90| 56. | 47,50| 84. | 49,90| 112. | 54,10| 140. | 58,30| | |

- Calculating the size of the data range
- Count the number of classes
- Look for class intervals
- Look for the bottom and top intervals
- Find the middle value of the interval (t)
- The Oi value is obtained from the amount of data that is between the lower and upper intervals.
- Looking for relative frequency
- Find the cumulative probability f (t)
- Look for R (t), i.e. opportunities do not occur
- Finding the parameter values for each type of distribution using the Minitab Software 15. The steps are as follows:
- Input data that has been sorted into Minitab 15
Click Stat> Quality Tools> Individual Distribution Identification.
Selected data to find the parameter value.
Click OK, the parameter values will appear as shown in the image below.

Figure 4. Results of parameter values for continuous distribution testing

Test the data of each distribution according to parameters
Furthermore, distribution testing is done using 4 types of continuous distribution. Where the authors use the example of a normal distribution to show the steps of distribution testing are as follows.

Formulate Hypothesis
\( H_0 \) : Normal Distributed Data
\( H_1 \) : Data is not normally distributed

Assuming \( \alpha = 0.05 \)
Calculate the probability (P) of each class using the Minitab Software 15. The steps to calculate the probability using the Minitab 15 Software are as follows:

1. Input the LCL and UCL values
2. Click Calc> Probability Distribution> Normal
3. Input the mean value of 50.42642 and standard deviation of 7.23482. In the Input column, fill in the location of the data source (LCL). In Optional Storage, it is filled with the output, which is pLCL. (The step is repeated to look for pUCL)
4. Click OK. Then the output will be displayed on the worksheet

| Distribution      | Location | Shape | Scale | Threshold |
|-------------------|----------|-------|-------|-----------|
| Normal*           | 50.42642 | 7.23482 |       |           |
| Box-Cox Transformation* | 50.42642 | 7.23482 |       |           |
| Lognormal*        | 3.91002  | 0.14661 |       |           |
| 3-Parameter Lognormal | 6.22051   | 0.01453 | -452.75639 | |
| Exponential       |          |       | 50.42642 |           |
| 2-Parameter Exponential |       | 19.14243 | 31.28399 |   |
| Weibull           | 7.83656  | 5.85191 |       |           |
| 3-Parameter Weibull | 3.47797   | 25.24772 | 27.71108 | |
| Smallest Extreme Value | 54.03935   | 7.23501 |       |           |
| Largest Extreme Value | 46.82871   | 6.98101 |       |           |
| Gamma             | 47.80244  | 1.05689 |       |           |
| 3-Parameter Gamma  | 107.90322 | 0.69939 | -25.05623 | |
| Logistic          | 50.43384  | 4.16499 |       |           |
| Loglogistic       | 3.81568   | 0.08375 |       |           |
| 3-Parameter Loglogistic | 8.99425 | 0.00052 | -5006.19501 | |

Table 2. PLCL and pUCL values for normal distribution calculations

| pLCLN | pUCLN |
|-------|-------|
| 0.004270 | 0.023961 |
| 0.023961 | 0.092384 |
| 0.092384 | 0.250039 |
| 0.250039 | 0.491017 |
| 0.491017 | 0.735433 |
| 0.735433 | 0.899934 |
| 0.899934 | 0.973383 |
| 0.973383 | 0.995129 |
| 0.995129 | 0.999395 |

Conclusions
\[ X^2_{\text{calculate}} = 10.8142 \]
\[ V = \text{number of class parameter (Mean, Standard Deviation and Total of Frequency)} = 6 - 3 = 3 \]
\[ X^2_{\text{table}} = 7.815 \text{ so that } X^2_{\text{count}} = 10.8142 > X^2_{\text{table}} = 7.815. \] Then \( H_0 \) is rejected so it can be concluded that the data is not in Normal distribution.

The results of data processing that have been carried out using the Chi-Square test and P-Value test on continuous distribution data are as follows:

| No | Distribution   | \( X^2_{\text{count}} \) | \( X^2_{\text{table}} \) | \( P\text{-Value} \) | Information  |
|----|----------------|---------------------------|--------------------------|---------------------------|--------------|
| 1  | Normal         | 10.8142                   | 7.8150                   | -                         | Rejected     |
| 2  | Exponential    | 174.6558                  | 9.4880                   | -                         | Rejected     |
| 3  | Lognormal      | 11.0774                   | 7.8150                   | -                         | Rejected     |
| 4  | Gamma          | 11.5637                   | 9.4880                   | -                         | Rejected     |

Evaluation on the test results of continuous distribution data distribution shows that the Hotel Room Occupancy Rate Data by Province 2013-2017 has Chi-Squared (2P) distribution, which is indicated by the results of EasyFit software.

**Figure 5. Grafik chi-squared (2P)**

5. Conclusion
Data for continuous distribution testing were obtained from Hotel Room Occupancy Rate Data by Province 2013-2017 totaling 165 data. Based on the results of continuous data processing, the results show that the continuous data is Chi-Squared (2P) distribution.

References

1. Nasruddin M et al 2008 *Statistika Edisi Revisi* (Medan: USU PRESS)
2. Sukoco, Agus dan Santiraningrum Soebandhi 2012 *Statistik dan Statistika*
3. Ronald E. Walpole 1993 *Pengantar Statistika Edisi ke-3* (Jakarta: PT Gramedia Pustaka Utama)
4. Ahsanullah. 2016 *Characterizations of Continuous Distributions by Truncated Moment* (USA: Digital Commons)
5. Walpole R E 1982 *Pengantar Statistika III* (Jakarta: Gramedia Pustaka Utama).
6. Harinaldi 2005 *Prinsip-Prinsip Statistik Untuk Teknik dan Sains* (Jakarta: Penerbit Erlangga).
7. Ribety Aprian 2009 *Distribusi Triangular dan Sifat-sifatnya*
8. Ronald E Walpole 1992 *Pengantar Statistika* (Jakarta: Gramedia Pustaka Utama)
9. Deny Kurniawan *Linear Regression*, [http://ineddeni.wordpress.com](http://ineddeni.wordpress.com), View on 1 September 2020