EXPLORING FIRST YEAR UNIVERSITY STUDENTS' STATISTICAL LITERACY: A CASE ON DESCRIBING AND VISUALIZING DATA

Ezra Putranda Setiawan, Heru Sukoco

Universitas Negeri Yogyakarta, Jl. Colombo 1 Yogyakarta, Indonesia
Email: ezra.ps@uny.ac.id

Abstract

Statistical literacy, which is the ability to use statistics in daily life, is an essential skill for facing society 5.0. This study aims to explore first-year university students' ability to properly use simple descriptive statistics and data visualization. Qualitative data were collected using a set of questions from 39 undergraduate students. Many students were able to calculate various descriptive statistics, but some of them were still unable to determine suitable statistics to describe the data clearly. Related to data visualization, many students failed to provide a meaningful chart that effectively shows the difference between two groups of data. Students with higher statistical literacy tend to use comparison or variability reasoning to determine the usage of descriptive statistics, and use data-based reason in visualizing the data. Improvement in statistical teaching – both in the university and the secondary school – is needed so that the students can use descriptive statistics and data visualization correctly.

Keywords: Statistics Education, Data Visualization, Descriptive Statistics, Qualitative

Abstrak

Literasi Statistika, kemampuan untuk menggunakan statistika dalam kehidupan sehari-hari, merupakan salah satu kemampuan penting untuk menghadapi masyarakat 5.0. Penelitian ini bertujuan mengeksplorasi kemampuan mahasiswa tahun pertama dalam mempergunakan statistik deskriptif sederhana dan visualisasi data yang telah mereka pelajari pada jenjang sekolah menengah. Data kualitatif dikumpulkan dari 39 mahasiswa menggunakan beberapa soal, dan diperoleh hasil sebagai berikut. Mahasiswa mampu menghitung bermacam statistik deskriptif, namun sebagian diantaranya tidak mampu menentukan statistik yang tepat untuk menggambarkan data secara jelas. Berkaitan dengan visualisasi data, sejumlah mahasiswa tidak mampu meng-gambar diagram yang bermakna untuk menunjukkan perbedaan antara dua kelompok data. Siswa dengan literasi statistika relatif tinggi cenderung menggunakan alasan terkait perbandingan dan keragaman untuk menentukan penggunaan statistik deskriptif, serta alasan berdasar-data untuk dalam menyajikan visualisasi data. Diperlukan peningkatan kualitas pengajaran statistika, baik pada jenjang pendidikan tinggi maupun sekolah menengah, agar mahasiswa dapat mempergunakan statistik deskriptif maupun visualisasi data dengan tepat.

Kata kunci: Pendidikan Statistika, Visualisasi Data, Statistik Deskriptif, Kualitatif

The 4.0 industrial revolution has occurred around the world, as indicated by the rapid growth of big data, the Internet of Things (IoT), and artificial intelligence. Gandomi and Haider (2015) stated that big data is characterized by several properties namely the data volume, velocity, variety, veracity, and value. These types of data can be obtained from various sources such as the usage of applications, social media, and the internet of things (IoT) (Yaqoob et al., 2019). As the volume of data grows at a rate of 50% per annum (Waal-Montgomery, 2016), understanding data is an essential skill for any data-driven society. Analysis of data using statistics is needed to filter massively available information and separate them from opinions (Widjajanti et al., 2017). These situations increase the need for statistical literacy competence (Frost, 2013).

What is statistical literacy? The definition of statistical literacy varies across literatures (Sharma, 2017). In the beginning, statistical literacy is defined as the ability to "... comprehend text and the meaning and implications of the statistical information in it, in the context of the topic to which (it) pertains" (Rumsey, 2002). It is also defined as "trans-numerative thinking" where people will be capable of making sense of and use a different representation of data to make sense of the situation among them (Chick et al., 2005). In a broader view, statistical literacy consists of statistical understanding required in modern democracies as well as peoples' dual role of statistical producer and consumer (Gould, 2017). Statistical literacy is also related to statistical reasoning, which can be defined as the way people reason with statistical ideas and make sense of statistical information (Ben-Zvi & Garfield, 2004).

To describe someone's statistical literacy, some frameworks have been proposed, for example by Wild and Pfannkuch (1999), Gal (2004), Watson and Callingham (2003), and Sharma (2017). These frameworks can be used to classify students' statistical literacy into several level or categories. Based on these frameworks, a high level of statistical literacy can be represented by the ability to critically question engagement with context, use proportional reasoning, appreciate the need for uncertainty, as well as understand the purpose of data, data analysis, and data representation.

In general, testing students' statistical literacy can be done at any level. However, measuring their statistical literacy on their first-year university study has several benefits. First, statistical literacy should become the goal for the Introductory Statistics course (Rumsey, 2002). Therefore, students' statistical literacy can be used as a diagnostic tool before learning the introductory statistics course at the university. Proper ability to use statistics, as well as the result of the university statistical course (if any) should enable students to work with data and do meaningful research for their thesis. Secondly, statistics is widely known as an awkward course in the university and sometimes causes statistical anxiety (Hedges, 2017). One of the factors affecting statistical anxiety is the students' previous knowledge of statistics (Sutarso, 1992). Understanding students' previous knowledge might help the instructor to adjust the statistics course and prevent statistical anxiety. Third, this result may represent how students have learned statistics in secondary school. It is widely known that statistical literacy competence is a part of the school curriculum in many countries (Watson, 2003; Garfield & DelMas, 2010), such as the USA (Weiland, 2017), Brazil (Campos et al., 2011), and Indonesia (Setiawan, 2019). Therefore, any information about first-year university students' statistical literacy might be insightful for the improvement of statistical teaching at the secondary school level. Lastly, the result can help the statistics educator to ensure that students' statistical literacy competence is sufficient for their life in their workplace and in their society.

As listed in Ziegler and Garfield (2017), various instruments had been proposed to evaluate students' statistical literacy by measuring many aspects on statistical literacy. Most of them was an objective test with focus on a broad competence of statistical literacy rather than a specific one. Using some of these instruments, many studies have been done to evaluate the statistical literacy of university

students. Yotongyos et al. (2015) presented that students from a university in Thailand have a moderate level of overall statistical literacy. Kim et al. (2019) did a test and observed the lesson plan to measure pre-service mathematics teachers' statistical literacy in Korea. In Pakistan, a survey revealed that BS students have a low level of statistical literacy (Hassan et al., 2020). However, there are several limitations on such studies that should be anticipated. For example, studies by Khaerunnisa and Pamungkas (2017), Takaria and Talakua (2018), as well as Jatisunda et al. (2020) give emphasis on students' ability to calculate statistics or follow a prescribed statistical procedure. Although proposing to describe statistical literacy, these studies did not measure students' critical reasoning which is the indicator of statistical literacy. These studies also measure wide aspects of statistical procedures, from data tabulation up to the statistical hypothesis testing, and analyze them as a whole as the statistical literacy competence. Hence, it is difficult to identify the ability that is not yet mastered by the students.

Following Gould (2017), knowing how to analyze data and create a basic representation of data is a part of the minimum level of statistical literacy. These abilities are essential to understand what the data have to say. Therefore, students must be able to do these procedures correctly before learning more advanced topics such as hypothesis testing, regression analysis, etc. For educational purposes, higher-order thinking skills (HOTS) should be encouraged so that the student can use the proper statistics and data visualization. Moreover, they must be able to give critical justifications on the usage of statistics and data visualization.

This study aims to present undergraduate students' statistical literacy in terms of their ability to use descriptive statistics and visualize data appropriately. We identify whether they can select and calculate correct descriptive statistics which can represent the data. Similarly, we examine how the students visualize the data using their way, without any guidance on which types of diagrams should be used. Simply speaking, this study focuses on how students can produce meaningful statistics and data visualization based on the data.

METHOD

Approach and Subject

To extensively describe students' abilities in descriptive statistics and data visualization, we used a qualitative approach. Following Creswell (2014), this type of study examines a natural situation and is suitable to describe the actual result from the subjects. We provided no treatment or manipulation to the respondents before and during the study.

The subjects of this study were 39 students in the first semester of the undergraduate program in statistics at a public university in Indonesia. Therefore, we expected that they would have some interest in statistics and want to learn more about it.

The profile of the respondents is as follows. Most of them were 17-18 years old, and 9 (23.1 %) were male. One student had graduated from the pharmacy stream of a vocational school (SMK), whereas the rest had graduated from the general or Islamic high school (SMA/MA) in Indonesia.

Data Collection

The data were collected using a test for students. In preparing the test, we assumed that students had already known several descriptive statistics and data visualization since these concepts were studied in primary and secondary school. Although some students had learned statistical inference in the twelfth grade of high school (Setiawan, 2020), we did not explore this topic further since it was taught only in the mathematics and natural science stream.

Formulation of the questions was inspired by a list of questions for assessing statistical education presented by Garfield and Ben-Zvi (2007) as well as Sharma (2017), with a focus on using descriptive statistics and visualizing data. The questions given to the subjects are presented in Figure 1.

You are familiar with descriptive statistics (eg. mean, mode, standard deviation, etc.). Suppose that we have the data of age (in years) of people with disease X from two different villages.
Village A: 10, 12, 14, 35, 56, 58, 60
Village B: 32, 34, 35, 36, 38

a) A researcher reported that the mean and median age of peoples with disease X in village A are the same as people in village B. Do you agree with that statement? Explain.
b) Suppose you are asked to give other descriptive statistics to explain the report given by the researcher mentioned in previous question, what statistics will you give? Calculate it and give your reason.
c) Someone asks you to present the above data in form of a chart/diagram. What type of chart you will use? Draw the chart and give your reason.

Figure 1. Translation of questions used in this study

As shown in Figure 1, these questions have a context, namely the ages of the patients, which means that the data must be positive. The dataset consists of one variable with two categories or groups. Students with a higher level of statistical literacy would be aware of the presence of two groups and be able to show the correct comparison between them.

The first question ensures that the respondents can calculate the mean and median from raw-ungrouped data. Since this is a closed question, students' answers can be classified only into two groups, namely the correct and incorrect answers. The correct answer for the median and mean of these two groups is 35, which implies that they are equal.

The second question moves to the proper usage of descriptive statistics, while the last question checks the ability to visualize the data in a suitable form. In answering these questions, students were allowed to use a calculator but not open any textbooks or references. Different from the test arranged in Garfield and Ben-Zvi (2007) or Jatisunda et al. (2020), we did not give any specific descriptive statistics nor chart type. Therefore, these questions can identify how the respondents understand the proper usage of descriptive statistics and chart types. Validity of the above questions was examined by consulting an expert in mathematics education.

Data Analysis

Following Mayring (2000), this study used two approaches to analyze the data. The deductive and inductive approaches were used for classifying students' answers and their reasons, respectively.

Regarding the level of students' statistical literacy, we found that the framework by Jones et al. (2000) or Watson and Callingham (2003) was somewhat abstract and difficult to follow. On the other hand, Sharma (2017) introduced four stages of students' statistical literacy, with informal/idiosyncratic as the lowest level, followed by consistent non-critical, early critical, and advanced critical. A description for each level was available and applicable to our problems. However, since it was rather difficult to separate the answers into four categories, we proposed a similar framework with three different levels of statistical literacy, as shown in Table 1.

Table 1. Framework for classifying students answer in question (b) and (c)

Level	Answer of question (b)	Answer of question (c)
Low	Student provide non-sense statistics or repeat calculation of statistics that already used in (a).	Student provide incorrect chart or use wrong data to create the chart.
Middle	Student calculate descriptive statistics other than mean/median correctly, but failed to show the difference between the two groups *or*	Student create a chart but unable to show the difference between the two groups *or*
	Any descriptive statistics other than mean/median were calculated incorrectly.	Student failed to use a proper scale on the chart.
High	Student calculate descriptive statistics other than mean/median correctly and correctly presents the difference between the two groups.	Student create a chart using proper scale and clearly show the difference between the two groups.

In reference to Sharma (2017), the low level on Table 1 corresponds to the informal level, in which students provide random or inappropriate explanations. The middle level on Table 1 represents the consistent-non-critical level since students in this level are able to use simple statistics and graphs. Since the high-level indicates that students are able to present the difference between the two groups, it can be classified into the early or advanced critical on Sharma (2017) model. The samples of students' answers from each level are presented in the next section.

The application of the framework on Table 1 is as follows. First, we checked whether each student's calculation of descriptive statistics in questions (a) and (b) was correct. These statistics used in (b) and charts used in (c) were then classified using the above framework. The classifying and coding process were done by two researchers independently, with 94.9% inter-coder agreement.

For the classification of students' reasoning, this study wants to present the original reason given by the students. Therefore, we did the data analysis procedure using inductive category development as follows. First, we list the reasons given by the students and group them to produce the initial coding. Later, a revision was done on the initial coding to produce the final coding criteria. Obtained result from

the final coding then presented as the result of this study.

RESULTS AND DISCUSSION

Students' Ability on Using Descriptive Statistics

In general, any descriptive statistics can be calculated based on quantitative or numerical data manually or using an electronic calculator. Following the Indonesian mathematics curriculum, students learn several descriptive statistics from primary school up to high school. In primary school, they study how to calculate the mean, mode, and median, whereas, in secondary school, they explore the quartile(s) and range. In high school, students learned the absolute deviation, variance, and standard deviation. They also calculate each descriptive statistic studied before for the grouped data. Consequently, when starting their study at the undergraduate level, students are familiar and should be able to calculate and use various descriptive statistics.

From the answers to the first question, we find that almost all respondents can calculate the mean and the median. They can show that both the mean and median of the patients' age from village A are equal to the mean and median of patients' age from village B. As seen in Figure 2, the calculation of these statistics is quite simple.

untuk Nilai Rata-Rata Data Desa A. $\bar{X} = \dfrac{10 + 12 + 14 + 35 + 56 + 58 + 60}{7} = 35$ untuk Nilai rata-rata dari Desa B $\bar{X} = \dfrac{32 + 34 + 35 + 56 + 38}{5} = 35$ Nilai Median (Q_2) dari Desa A × B = 35 Penjelasan; untuk Mencari Nilai Median dengan Mudah, urutkan data dari yang terkecil. (Berlaku untuk data yg kecil) 10, 12, 14, ㉟, 56, 58, 60 = 35 32, 34, ㉟, 36, 38. = 55.	*For the average of the village A data ....* *For the average of the village B data ...* *The median from village A and B ...* *Explanations: to find the median, sort the data from the smallest*

Figure 2. Sample correct answer for the first question

If the students recognize the difference between the data from these two villages, they should note that the second question asks them to give statistics that represent these differences. Based on this idea, any statistics that the calculation based on the data from each village yields a different value is a correct answer.
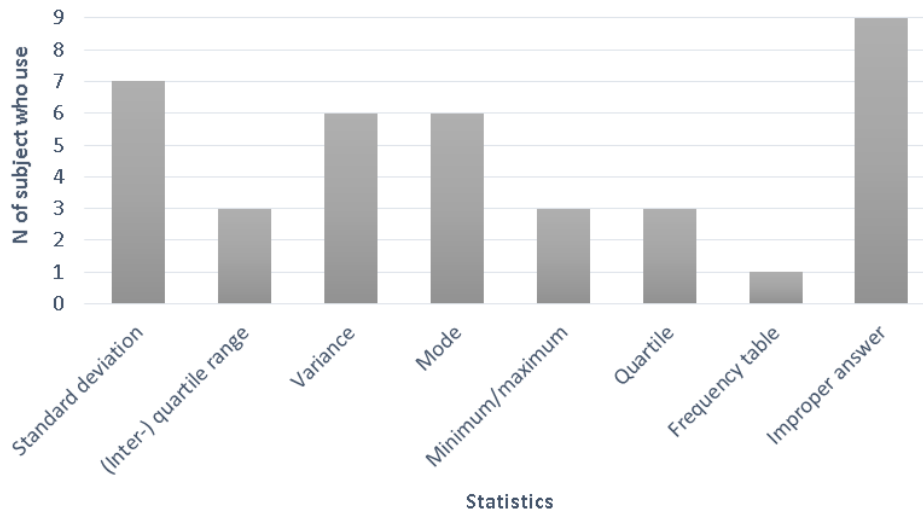
Figure 3. Descriptive statistics given by the respondent to answer the second question

Figure 3 shows that there are seven descriptive statistics given by the respondents to answer the second question, which was dominated by dispersion measures (i.e. standard deviation and interquartile range). Since all the data are different, the value of any dispersion measure between these two villages is different. Similarly, the first quartile, third quartile, minimum, and maximum of data from these two villages are not equal. Based on Table 1, more than half of the students belong to the high literacy group since they show the difference between these two groups of data using correct statistics.

> *"Berdasarkan perhitungan dapat kita ketahui simpangan baku usia penderita penyakit X di desa A dan B tidak sama.* (From our calculation, we know that the standard deviation of age of patients with disease X in the village A and B are not equal)" (Student #9)

> *"Karena menunjukkan tingkatan keberagaman dari data tersebut.* (Because it presents the degree of variability of the given data)" (Student #30)

> *"Dengan adanya Q1, Q3 dan IQR maka akan dapat mendukung informasi dalam sajian data.* (By the presence of Q1, Q3, and IQR will support information in the data presentation)" (Student #25)

> *"Berdasarkan data yang diberikan, kita dapat menghitung simpangan baku dari kedua desa tersebut.* (Based on the given data, we can calculate the standard devation from these two villages)" (Student #34)

It can be seen that the reason given by students from the high group were related to various aspects, namely the difference between the two groups (e.g. Student #9), the variability of the data without mentioning the difference (e.g. Student #30), the goal of giving information based on the data (e.g. Student #25), and the possibility of calculation (e.g. Student #34).

Seven students in the medium level of statistical literacy can use other descriptive statistics but fail to show the difference between these groups. This group is dominated by students that (incorrectly) calculate the mode of the data, while the data for each village have no mode. The samples of reasons

given by students in this group were as follows.

"*Rata-rata, median, dan modusnya sama.* (The mean, median, and mode are equal)" (Student #18)

"*Karena nilai yang sering muncul hanyalah 35. 35 muncul paling banyak 2 kali, yang lain hanyalah sekali.* (Because the most frequent data is 35. 35 was appeared two times, while the others are only once)" (Student #21)

"*Karena dengan banyaknya modus dapat mengetahui nilai yang sering keluar dan bisa memperjelas informasi.* (Because by using mode we are able to know the most occurred value and can get clearer information)" (Student #3).

Both Student #18 and Student #21 incorrectly calculated the mode, which caused them stated that the mean, median, and mode were the same. On the other hand, Student #3 did not calculate the mode so that he/she might not realize that there was no mode in the data. Another type of student in this group provided a wrong calculation of variance (i.e. did not take the square of difference) resulting in zero variance and did not provide other statistics. As a consequence, this student claimed that the variances between these two villages were the same.

In the third group, namely the lowest statistical literacy, we found several students that repeated the calculation of mean or median or transforming the data into a table. These students might not be aware of several descriptive statistics mentioned in the questions, as presented by Student #28. Further classification of the improper answer is presented in Table 2.

"*Memberikan rata-rata merupakan informasi yang tepat dan pasti karena sudah diketahui berapa rata-ratanya.* (Present the mean is a correct and certain information because its value has been known)" (Student #28, present the mean which already calculated).

Table 2. Incorrect answer of question (b) represent low statistical literacy

Types of improper answer	Sample answer
Non-sense statistics: calculating the mean and present them as a percentage.	
Assuming presence of another information (*jenis kelamin* = sex) that did not given nor asked in the question.	

Doing hypothesis testing.	Statistik yang digunakan adalah perbedaan antara dua rata-rata tersebut $\Rightarrow$ $\bar{X}_A - \bar{X}_B$ ... ... $\bar{X}_A - \bar{X}_B = 35 - 35 \Rightarrow \bar{X}_A - \bar{X}_B = 0$ ... ... Dalam hal ini, yang ditanyakan adalah apakah $\bar{X}_A = \bar{X}_B$ dan apakah med A = med B? Dalam hal ini kita akan menguji dua rata-rata, sehingga statistik yang digunakan adalah perbedaan rata-rata antara desa A dan desa B. Hal ini dapat dituliskan dalam hipotesis: $H_0 : \bar{X}_A - \bar{X}_B = 0$ jika setelah diuji ternyata $\bar{X}_A - \bar{X}_B = 0$ maka ini menunjukkan bahwa $H_a : \bar{X}_A - \bar{X}_B \neq 0$ tidak ada perbedaan $\bar{X}_A$ dg $\bar{X}_B$, dengan kata lain $\bar{X}_A = \bar{X}_B$ • Selain itu, data ini merupakan data kuantitatif

From Table 2, a notable result is that several students were unable to get the information from the dataset. Surprisingly, they asked for more information or add some information that was not presented nor asked by the question, with some reason as follows.

> "*Statistik yang perlu ditambahkan seperti: jumlah penduduk desa A dan desa B; jumlah penduduk desa A dan desa B yang sehat* (Added statistics should be: the number of population in village A and village B; number of healthy people in village A and village B)" (Student #30)

> "*Dengan adanya jumlah penderita di tiap umur, maka kita dapat mencari nilai dan membandingkannya* (If the number of patients in each level of age was available, we can find the value and compare them)" (Student #33)

The reason given by Student #33 shows that she did not understand that the presented data was the overall raw data instead of arranged data in the frequency distribution. Student #10, who wrote a hypothesis testing procedure to answer this question, might think that hypothesis testing was part of descriptive statistics instead of inferential statistics.

Based on these results, we can infer that most students know and are able to calculate various descriptive statistics. Meanwhile, some of them may have low competence in determining suitable statistics to describe and compare the raw data, or may be unaware that not all descriptive statistics (i.e. mode) can be used in any dataset.

Students Ability on Using Data Visualization

As presented in Figure 1, the data in the question is about the ages of people with a specific disease, which is on an interval or ratio scale. Theoretically, the suitable data visualization might be bar charts, dot plot, histograms, box (-and whisker-) plot, or stem-and-leaf plot.

Of 39 participants of this study, only 37 participants answered this question. Most of the students (40%) use a bar chart to display the data, which seems reasonable since this type of data visualization is taught from the primary school level up to the higher secondary school level. The box plot and the stem-and-leaf plot are used by one student each, whereas the dot plot is created by four students. Most of the students create two separate charts, or one chart for each village, instead of combining this information into one chart. As a consequence, the same chart type (e.g. bar chart) may look very different: one presents the difference between the two villages clearly, whereas the other contains

various mistakes and difficult to understand.

Following the classification on Table 1, students' answers that exhibit a high level of statistical literacy are displayed in Table 3. These charts clearly show the difference between the patients' ages from these two villages and use a correct scale in the axes. Some of the students in this group present all data in one chart, whereas the others use two separate charts. When the latter is used, statistical-literate students must create the same scale for each chart so that the reader can compare the data between these two groups easily.

Table 3. Sample proper chart types with clear/correct graphing

Figure	Explanations
	A boxplot clearly show the same median and the different variability of the data between these two groups.
	A stem-and-leaf plot was good for presenting the difference of variability between the two villages.
	A dot plot comparing the patients age, clearly show the variability of age between these two villages. The central tendency can be simply guessed from the display.
	A bar chart made by grouping the patients' age into five classes. The difference of patients' age among the two groups can be clearly seen.

In the group of students with a medium level of statistical literacy, we find two kinds of answers as follows. Despite presenting the data in a suitable chart, several students did not give much attention to the scale on the axis. As presented in Figure 4, they only put the value of the data below the axis without seeing the difference between them. We can read this chart, but the comparison between those two groups would be difficult to observe. Compared to the charts in Table 2, these charts did not represent the difference of variability between these two groups.
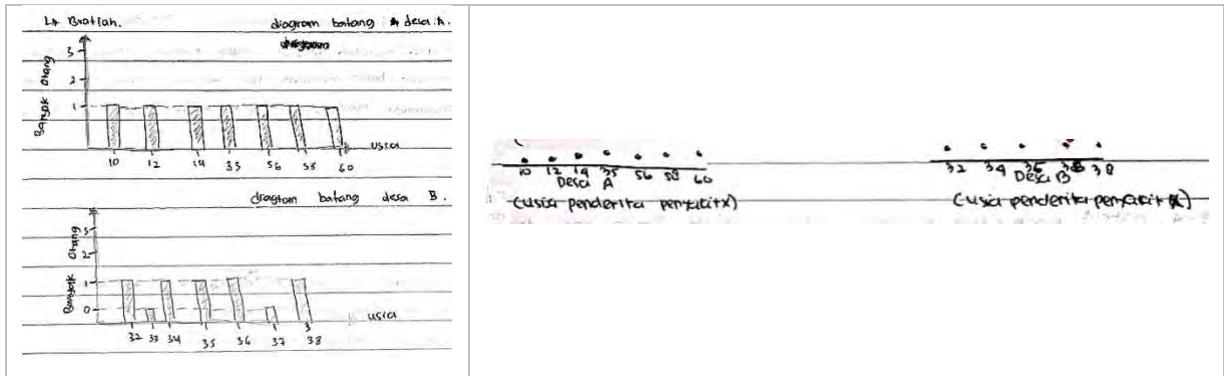
Figure 4. Correct data visualization in wrong scale. Note the equal distance between each bar or each dot, despite the difference of the level.

Secondly, we found that several students use improper chart types to display the data, as presented in Table 4. Similar to Figure 4, these charts were unable to show the different variability of the data between these two villages. The correct parts of charts in Table 4 were only the variable and groupings of data. Instead of making the reader directly understand the data, these chart types might cause the reader to feel confused. Difficulties in creating and reading would arise when the number of the subject becomes larger.

Table 4. Sample improper chart representing middle level of statistical literacy

Figure	Explanations
	Only convert the raw data into bars; no explanations on the horizontal axis. Although the difference between the two villages is represented by the colour, this chart is unacceptable.
	Since the data consists of only one variable, there was no reason to made two dimensional plot.
	A pie chart might represent the grouped data, but it is very difficult to compare the patients' age between these two groups.

Figure	Explanations
	Although the content is correct, Venn diagram is used in set theory and not proposed to display numerical data. The variable also not mentioned in the diagram.

Students with a low level of statistical literacy created several charts that did not represent the original data. These charts might use irrelevant data or unimportant variable. Such types of chart, which is displayed in Table 5, represent the lowest competence in data visualization among all students who participated in this study.

Table 5. Sample improper chart representing low statistical literacy on data visualization

Types of mistakes	Sample answer
Failed to identify the proper variable	
Wrongly present the descriptive statistics (which were same) instead of the original data.	
Add more information, namely the patients' sex that were not available in the original data.	

Why do students choose a chart type to represent the dataset? Our study finds that seven respondents give no reason for their data display types. However, various reasons given by the students can be classified into four categories, as shown in Table 6.

Among these types, more than half of the participants wrote people-based reason, while the rarest was the purpose-based reason. How are students' reasons related to data visualization? Almost all students that draw a correct diagram (performing a high level of statistical literacy) wrote data-based reason, which were sometimes combined with other types. In contrast, most students with incorrect

diagrams were only able to give a people-based reason or no reason at all.

Table 6. Classification of student reason when visualizing the data

Types of reason	Explanation	Sample answer
Data-based reason	Mention data properties, eg. number of variables, groups, types, etc.	"*Karena data yang diberikan merupakan data tunggal dan terdiri dari dua kelompok A dan B* (Because the data are single-valued and consist of two groups A and B)" (Student #16)
Chart-based reason*	Related to the chart types or chart usage.	"*Informasi dari data kuantitatif dapat terangkum dalam diagram ini.* (Information from quantitative data can be presented in this diagram)" (Student #10)
Purpose-based reason	Mention the purpose, ie. what are need to show or present from the data.	"*Dapat menampilkan hubungan jumlah penderita dengan umurnya.* (Can represent the relations between the number of patients and their age) (Student #31).
People-based reason	Related to the people who make and/or will read the diagram.	"*Agar mudah terbaca sehingga dapat dengan mudah menentukan simpulan mana yang mudah diambil.* (Will be easier to read, so that the conclusion can be taken easier)" (Student #13)

*In this study, all chart-based reason were combined to the data-based reason.

Descriptive statistics is the procedure used to organize and describe the characteristics or factors of a given sample to understand it (Fisher & Marshal, 2009). Data visualization can help the human eye see things that are difficult to understand in large datasets (Rodríguez et al., 2015). The learning of these topics frequently precedes the discussion on inferential statistics procedures such as hypothesis testing, estimation, and data analysis. Despite its simplicity, improper use of descriptive statistics or data visualization may lead to serious problems in understanding the data. A small book entitled 'How to Lie with Statistics' (Huff, 1954) and a textbook entitled 'Statistics, Concept and Controversies' (Moore & Notz, 2009) present various problems related to the wrong use of descriptive statistics and data visualization. For example, using an improper scale on a bar chart may cause wrong interpretation of the data. Usage of various descriptive statistics such as mean and median may yield very different results especially when the data contains one or more outliers.

In this study, we present the statistical literacy of undergraduate students in terms of their ability to use descriptive statistics and visualize data. The use of ill-structured essay questions yields some benefit over the multiple-choice question test like the ARTIST test (Garfield et al., 2002). First, various types of students' answers can be found and classified to measure their ability to use descriptive statistics and visualize data. Compared to Sharma (2017), this study not only measures students' ability to read and interpret the statistics or chart but also measures their ability to use proper statistics as well as meaningful charts. We can say that students need to apply higher-order thinking skills (HOTS) to create an understandable representation of the data. Secondly, these questions can identify their' ability to

work with more than one group of data that represent their level of statistical literacy. Lastly, the reasons given by the students might be used to identify their statistical reasoning ability, which needs to be confirmed using other tests (Sabbag et al., 2018).

Respondents of this study were first-year students from an undergraduate program in statistics, whom we expected to have more awareness (and maybe more interest) on statistics. As expected, in the test, almost all students could easily show that the mean and median between the two villages were equal. We also find that most of them have a medium to a high level of statistical literacy in using proper descriptive statistics, as represented by their ability to provide other statistics that represent the difference between two groups of data. Similarly, there is only a small number of subjects that exhibit a high level of statistical literacy on visualizing data. By examining various types of charts, we find that students are unfamiliar with histograms, boxplots, stem-and-leaf plots, and dot plots, which are more suitable for continuous data such as age. Many subjects failed to identify that a pie chart, line chart, and even a Venn diagram, are not suitable to display this type of data.

The low ability on statistical descriptive and/or data visualization indicate a problem on statistical education. As mentioned by Ismail and Chan (2015), many studies show that there is misconception about the usage of descriptive statistics on students from various level. In addition, our study show that these problems also presented on data visualization. Since the problems did not suggest any type of data visualization, students must use their knowledge to decide which type that can be used. These approaches were not used by previous statistical literacy studies on Indonesian undergraduate student (Khaerunnisa & Pamungkas, 2017; Jatisunda et al., 2020; Tiro et al., 2020). As a consequence, none of these studies' present students' ability on visualizing data properly.

This study also explores various students' reasoning on using descriptive statistics and data visualization. Various reasons can be found on the high-level of statistical literacy regarding the use of descriptive statistics, namely (1) comparison between the groups, (2) variability measurement, (3) additional information, and (4) possibility of calculation. Among the middle-level and low-level groups of statistical literacy, most of the reasons were based on calculation, and none of them were related to the comparison of variability. Based on these results, we can say that students with a higher level of statistical literacy would be more likely to obtain the idea of comparing the groups of data and measuring the variability. This result confirms the frameworks from Jones et al. (2000) that place the ability to make a comparison on the high level of statistical reasoning. Similarly, understanding the variability or variation is somewhat complex and difficult, so not all students and even teachers can understand this concept well (Sánchez et al., 2011). Regarding the data visualization, we find that students' reasons for creating them can be divided into four groups, as presented in Table 6. Data-based reason becomes the most frequent reason used by students with a higher level of statistical literacy, i.e. those who can create correct and meaningful visualization.

This study provides more insight on understanding how students develop their reasoning when summarizing data (using descriptive statistics) and visualizing data with minimal guidance. It can be

used to identify how students will act when facing a real situation related to data, namely when there is nobody that asks them to make a specific chart or calculate specific statistics. Therefore, it can be seen as an alternative to the framework on statistical reasoning by Chan and Ismail (2013) as well as Chan et al. (2016) which seems to be more rigid.

Even though this study only accounts for several undergraduate students in statistics, the result might be applied for any undergraduate students especially those that come from the mathematics and natural science stream in secondary school. Further studies with a qualitative approach should be done to profile another concept related to statistical literacy and its reasoning. Suitable concepts for the study may include the idea of sampling, design of experiments, statistical inference, and many more. To ensure generalizability, more respondents from various undergraduate programs can be chosen to participate in a similar study.

Implications on Teaching Statistics

This study raises a question: How can we increase the statistical literacy of undergraduate students, especially in the term of using descriptive statistics and data visualization? In Indonesia, it has been known that some of these concepts were introduced in primary and secondary schools (Setiawan, 2019; Funny et al., 2019). Therefore, an improvement on these two topics should be carried out on the university level as well as on the secondary school level.

At first, statistical learning in schools, which is dominated by computational aspects of statistics instead of conceptual understanding (Tiro, 2018), should be synchronized (Ridgway et al., 2011). Statistical literacy should become an important part of the multi-literacy model used in developing primary and secondary school curricula (Abidin, 2017; Nurgiyantoro et al., 2020). Data – and statistics – should become used in various subjects outside mathematics. As an example, in Geography students can learn how to represent spatial datasets, while in Economy student may visualize and identify the pattern on time series data. This approach may help students to obtain the conceptual understanding of the data beside the mathematical procedures of calculation.

We should realize that sufficient ability in statistical literacy could not be developed by statistical teaching that focused on gathering statistical knowledge, learning facts and formulas, and obeying standard procedures (Schield, 2004). As a consequence, more innovation on the strategies used for teaching statistics is needed. One of them, for example, is the guided discovery learning (Hariyanti & Wutsqa, 2020). The use of various modern technology (Suhermi & Widjajanti, 2020) should help the students increase their statistical literacy and statistical reasoning.

Usage of real datasets in the statistics classroom also encouraged in which open data can be used (Ridgway, 2016; Rivera et al., 2019). This approach will help students to face the emergence of big data, data science, and data analytics.

Moving into the data visualization, it is noted that teaching strategy using various modern tools has been developed, for example, by Nolan and Perret (2016) or Gelman and Nolan (2017). Following

Wolfe (2015), several textbooks on communication courses can help us to find completed guidelines for determining the visualization types. Numerous literatures on student difficulties related to data visualization (Boels et al., 2019; Dewi et al., 2020), misconceptions (Zaidan et al., 2012; Chan & Ismail, 2013; Ismail & Chan, 2015; Yusuf et al., 2017), as well as learning obstacles (Sotos et al., 2007) can be used as references on improving the course of data visualization.

Drawing data visualization by hand might be irrelevant for undergraduate students. When a software is used, the focus on teaching data visualization needs to be put on the usage of proper diagram instead of the steps to produce it. Recent types of data visualization such as *heatmap* and *violin plot* can be introduced. On the other hand, since drawing charts still become a competence for students in primary and/or secondary school (Setiawan, 2019; 2021), this suggestion is relevant for (mathematics) teacher in these levels. Similar study can be carried out to identify whether mathematics teacher and/or pre-service teacher are able to teach the correct usage of data visualization.

CONCLUSION

This study analyzed the statistical literacy of first-year students from the undergraduate program on statistics, with a focus on their ability to use descriptive statistics and to visualize the data. Half of the students participated in this study exhibit high level of literacy on using descriptive statistics, but with middle level of literacy on visualizing the data.

Related to the usage of descriptive statistics, almost all students are able to calculate the mean and the mode, which are central measure. We find that students with a higher level of statistical literacy become able to use comparison and variability reasons when choosing descriptive statistics that lead to the usage of a dispersion measure. On the contrary, students with lower level of statistical literacy are unable to realize the different variation between the groups of data.

Students with higher level of statistical literacy can visualize the data in the way so that the difference between these two groups are clear. They also more likely to give data-based reason, which mentions data properties such as the number of groups, types, etc. Sometimes these reasons are combined with chart-based reason or purpose-based reason. In contrast, students with low level of statistical literacy only able to give people-based reason such as 'easier to read' or 'easier to understand'. Further study with a similar approach should be done to profile the other components of statistical literacy and its reasoning.

ACKNOWLEDGMENTS

REFERENCES

Abidin, Y. (2017). Developing literacy learning model based on multi literacy, integrated, and differentiated concept at primary school. *Jurnal Cakrawala Pendidikan*, *36*(2), 156-166. https://doi.org/10.21831/cp.v36i2.13283

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In: Ben-Zvi D., Garfield J. (Eds.) *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*. Springer, Dordrecht. https://doi.org/10.1007/1-4020-2278-6_1

Boels, L., Bakker, A., Van Dooren, W., & Drijvers, P. (2019). Conceptual difficulties when interpreting histograms: a review. *Educational Research Review*, *28*, 100291. https://doi.org/10.1016/j.edurev.2019.100291

Campos, T. M. M., Cazorla I. M., & Kataoka V. Y. (2011) Statistics school curricula in Brazil. In: Batanero C., Burrill G., Reading C. (eds) *Teaching statistics in school mathematics-challenges for teaching and teacher education*. New ICMI Study Series, vol 14. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-1131-0_1

Chan, S. W., & Ismail, Z. (2013). Assessing misconceptions in reasoning about variability among high school students. *Procedia-Social and Behavioral Sciences*, *93*, 1478-1483. https://doi.org/10.1016/j.sbspro.2013.10.067

Chan, S. W., Ismail, Z., & Sumintono, B. (2016). A framework for assessing high school students' statistical reasoning. *PLOS ONE*, *11*(11), e0163846. https://doi.org/10.1371/journal.pone.0163846

Chick, H. L., Pfannkuch, M., & Watson, J. M. (2005). Transnumerative thinking: finding and telling stories within data. *Curriculum Matters*, *1*, 87-109.

Creswell, J. W. (2014). *A concise introduction to mixed methods research*. London: SAGE Publications.

Dewi, D. K., Khodijah, S. S., & Zanthy, L. S. (2020). Analisis kesulitan matematik siswa SMP pada materi statistika [Analysis of the mathematical difficulties of junior high school students on statistical material]. *Jurnal Cendekia: Jurnal Pendidikan Matematika*, *4*(1), 1-7. https://doi.org/10. 31004/cendekia.v4i1.148

Fisher, M. J., & Marshall, A. P. (2009). Understanding descriptive statistics. *Australian Critical Care, 22*(2), 93–97. https://doi:10.1016/j.aucc.2008.11.003

Frost, J. (2013) Why statistics is important. *The World of Statistics*. Retrieved from http://www.worldofstatistics.org/2013/03/04/why-statistics-is-important/

Funny, R. A., Ghofur, M. A., Oktiningrum, W., & Nuraini, N. L. S. (2019). Reflective thinking skills of engineering students in learning statistics. *Journal on Mathematics Education*, *10*(3), 445-458. https://doi.org/10.22342/jme.10.3.9446.445-458

Gal, I. (2004). Statistical literacy. In D. Ben-Zvi, & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 47-78). Dordrecht: Springer. https://doi.org/10.1007/1-4020-2278-6_3

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review*, *75*(3), 372-396. https://doi.org /10.1111/j.1751-5823.2007.00029.x

Garfield, J., & DelMas, R. (2010). A web site that provides resources for assessing students' statistical literacy, reasoning and thinking. *Teaching Statistics*, *32*(1), 2-7. https://doi.org/10.1111/j.1467-9639.2009.00373.x

Garfield, J., Hogg, B., Schau, C., & Whittinghill, D. (2002). First courses in statistical science: The status of educational reform efforts. *Journal of Statistics Education*, *10*(2), 1-14. https://doi.org/10.1080/10691898.2002.11910665

Gelman, A., & Nolan, D. (2017). *Teaching statistics: A bag of tricks*. Oxford: Oxford University Press.

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal*, *16*(1), 22-25. https://iase-web.org/documents/SERJ/SERJ16(1)_Gould.pdf

Hariyanti, F., & Wutsqa, D.U. (2020). Pengembangan perangkat pembelajaran statistika dan peluang untuk mengembangkan statistical literacy siswa SMP [Developing statistical learning tools and probability to develop statistical literacy for junior high school students]. *Jurnal Riset Pendidikan Matematika*, *7*(1), 46-58. https://doi.org/10.21831/jrpm.v7i1.14997

Hassan, A., Ghaffar, A., & Zaman, A. (2020). An investigative study on university students' statistical literacy in Pakistan. *Sir Syed Journal of Education & Social Research*, *3*(1), 159-164. https://doi.org/10.36902/sjesr-vol3-iss1-2020(159-164)

Hedges, S. (2017). Statistics student performance and anxiety: Comparisons in course delivery and student characteristics. *Statistics Education Research Journal*, *16*(1), 320-336. http://iase-web.org/documents/SERJ/SERJ16(1)_Hedges.pdf

Huff, D. (1954). *How to Lie with Statistics*. New York: Norton.

Ismail, Z., & Chan, S. W. (2015). Malaysian students' misconceptions about measures of central tendency: An error analysis. *AIP Conference Proceedings*, *1643*(1), 93-100. https://doi.org/10.1063/1.4907430

Jatisunda, M. G., Nahdi, D. S., & Suciawati, V. (2020). Literasi statistika mahasiswa administrasi public [Statistical literacy of public administration students]. *SJME (Supremum Journal of Mathematics Education), 4*(2), 134-146. https://doi.org/10. 35706/sjme.v4i2.3488

Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B., & Putt, I. J. (2000). Mathematical thinking and learning a framework for characterizing children's statistical thinking. *Mathematical Thinking and Learning*, *2*(4), 269–307. https://doi.org/10.1207/S15327833MTL0204

Khaerunnisa, E., & Pamungkas, A. S. (2017). Profil kemampuan literasi statistis mahasiswa jurusan Pendidikan Matematika Universitas Sultan Ageng Tirtayasa [The undergraduate mathematics education students' profile of Statistical literacy ability at Universitas Sultan Ageng Tirtayasa]. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *6*(2), 246-255. http://dx.doi.org/10.24127/ajpm.v6i2.970

Kim, D. E., Kang, P. L., & Lee, M. H. (2019). Analysis on preservice mathematics teachers' statistical literacy in lesson plan. *East Asian Mathematical Journal*, *35*(4), 429-449. http://dx.doi.org/10.7858/eamj.2019.036

Mayring, P. (2000). Qualitative Content Analysis. *Forum: Qualitative Social Research*, *1*(2), Art. 20. https://doi.org/10.17169/fqs-1.2.1089

Moore, D. S., & Notz, W. I. (2009). *Statistics: Concepts and Controversies.* New York: Freeman and Company.

Nolan, D., & Perrett, J. (2016). Teaching and learning data visualization: Ideas and assignments. *The American Statistician*, *70*(3), 260-269. https://doi.org/10.1080/00031305.2015.1123651

Nurgiyantoro, B., Lestyarini, B., & Rahayu, D. H. (2020). Mapping junior high school students' functional literacy competence. *Jurnal Cakrawala Pendidikan, 39*(3), 560-572. https://doi.org/10.21831/cp.v39i3.34061

Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical Review*, *84*(3), 528-549. https://doi.org/10.1111/insr.12110

Ridgway, J., Nicholson, J., & McCusker, S. (2011). Developing statistical literacy in students and teachers. In *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education* (pp. 311-322). Dordrecht: Springer. https://doi.org/10.1007/978-94-007-1131-0_30

Rivera, R., Marazzi, M., & Torres-Saavedra, P. A. (2019). Incorporating open data into introductory courses in statistics. *Journal of Statistics Education*, *27*(3), 198-207. https://doi.org/10.1080/10691898.2019.1669506.

Rodríguez, M. T., Nunes, S., & Devezas, T. (2015). Telling stories with data visualization. In *Proceedings of the 2015 Workshop on Narrative & Hypertext* (pp. 7-11).

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal of Statistics Education*, *10*(3), 1-12. https://doi.org/10.1080/10691898.2002.11910678

Sánchez, E., da Silva, C. B., & Coutinho, C. (2011) Teachers' Understanding of Variation. In: Batanero C., Burrill G., Reading C. (eds) *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education*. New ICMI Study Series, vol 14. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-1131-0_22

Schield, M. (2004). Statistical literacy curriculum design. *IASE Curriculum Design Roundtable*. Retrieved from *http://www.StatLit.org/pdf/2004SchieldIASE.pdf*

Sabbag, A., Garfield, J., & Zieffler, A. (2018). Assessing statistical literacy and statistical reasoning. *Statistics Education Research Journal*, *17*(2), 141-160. https://iase-web.org/documents/SERJ/SERJ17(2)_Sabbag.pdf

Setiawan, E. P. (2019). Analisis muatan literasi statistika dalam buku teks matematika Kurikulum 2013 [Statistical literacy content analysis in 2013 Curriculum mathematics textbooks]. *Pythagoras: Jurnal Pendidikan Matematika*, *14*(2), 163-177. https://doi.org/10.21831/pg. v14i2.28558

Setiawan, E. P. (2020). Introducing statistical inference to senior high school students: A textbook analysis. *Journal of Physics: Conference Series*, *1663*(1), 012014. https://doi.org/10.1088/1742-6596/1663/1/012014

Setiawan, E.P. (2021). Literasi Statistika dalam Kurikulum Matematika Sekolah Dasar (SD) 2004-2020: Tinjauan Historis dan Pengembangannya [Statistical Literacy in Elementary School Mathematics Curriculum 2004-2020: Historical Review and Its Development]. *Jurnal Pendidikan dan Kebudayaan, 6*(1), 1-20. https://doi.org/10.24832/jpnk.v6i1.1915

Sharma, S. (2017). Definitions and models of statistical literacy: a literature review. *Open Review of Educational Research*, *4*(1), 118-133. https://doi.org/10.1080/23265507.2017.1354313

Sotos, A. E. C., Vanhoof, S., den Noortgate, V.W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, *2*(2), 98-113. https://doi.org/10.1016/j.edurev.2007.04.001

Suhermi, S. & Widjajanti, D. B. (2020). What are the roles of technology in improving student statistical literacy?. *Journal of Physics: Conference Series, 1581*(1), 012067. https://doi.org/10.1088/1742-6596/1581/1/012067

Sutarso, T. (1992). Some variables in relation to students' anxiety in learning statistics. Paper presented at the *Annual Meeting of the Mid-South Educational Research Association*, Knoxville 1992, November 11-13.

Takaria, J., & Talakua, M. (2018). The ability of statistical literacy student teacher candidate in terms of prior-ability on mathematics. *Jurnal Kependidikan: Penelitian Inovasi Pembelajaran*, *2*(2), 395-408. https://doi.org/10.21831/jk.v2i2.18768

Tiro, M. A. (2018). National Movement for Statistical Literacy in Indonesia: An Idea. *Journal of Physics: Conference Series*, *1028*(1), 012216. https://doi.org/10.1088/1742-6596/1028/1/012216

Tiro, M. A., Ruliana, R., & Aswi, A. (2020). Eksplorasi literasi statistika deskriptif mahasiswa program studi Statistika Universitas Negeri Makassar dalam suasana pembelajaran daring akibat darurat Covid-19 [Exploration of descriptive statistical literacy for Statistics undergraduate students at Universitas Negeri Makassar in an online learning atmosphere due to the Covid-19 situation]. In *Seminar Nasional Variansi (Venue Artikulasi-Riset, Inovasi, Resonansi-Teori, dan Aplikasi Statistika)* (pp. 149-159). Makassar: Program Studi Statistika, FMIPA Universitas Negeri Makassar. https://ojs.unm.ac.id/variansistatistika/article/view/19497/pdf

Waal-Montgomery, M. D. (2016). *World's data volume to grow 40% per year & 50 times by 2020* Aureus. Retrieved from https://e27.co/worlds-data-volume-togrow-40-per-year-50-times-by-2020-aureus-20150115-2/.

Watson, J., & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal, 2*(2), 3–46.

Watson, J. M. (2003). Statistical literacy at the school level: What should students know and do. *The Bulletin of The International Statistical Institute, Berlim*, *54*, 1-4. https://iase-web.org/documents/papers/isi54/3516.pdf

Weiland, T. (2017). Problematizing statistical literacy: An intersection of critical and statistical literacies. *Educational Studies in Mathematics*, *96*(1), 33-47. https://doi.org/10.1007/s10649-017-9764-5

Widjajanti, D. B., Listyani, E., Susanti, M., & Setyaningrum, W. (2017). Examining Prospective Teachers' Mathematical Communication Skills in Statistics. In *1st Annual International Conference on Mathematics, Science, and Education (ICoMSE)*. Atlantis Press. https://doi.org/10.2991/icomse-17.2018.41

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223-248. https://doi.org/10.1111/j.1751-5823.1999.tb00442.x

Wolfe, J. (2015). Teaching students to focus on the data in data visualization. *Journal of Business and Technical Communication*, *29*(3), 344-359. https://doi.org/10.1177/1050651915573944

Yaqoob, I., Hashem, I. A. T., Ahmed, A., Kazmi, S. A., & Hong, C. S. (2019). Internet of things forensics: Recent advances, taxonomy, requirements, and open challenges. *Future Generation Computer Systems*, *92*, 265-275. https://doi.org/10.1016/j.future.2018.09.058

Yotongyos, M., Traiwichitkhun, D., & Kaemkate, W. (2015). Undergraduate Students' Statistical Literacy: A Survey Study. *Procedia-Social and Behavioral Sciences*, *191*, 2731-2734. https://doi. org/10.1016/j.sbspro.2015.04.328

Yusuf, Y., Titat, N., & Yuliawati, T. (2017). Analisis hambatan belajar (learning obstacle) siswa SMP pada materi statistika [Analysis of learning obstacles for junior high school students on statistical material]. *Aksioma: Jurnal Matematika dan Pendidikan Matematika*, *8*(1), 76-86. https://doi.org/10.26877/aks.v8i1.1509

Zaidan, A., Ismail, Z., Yusof, Y. M., & Kashefi, H. (2012). Misconceptions in descriptive statistics among postgraduates in social sciences. *Procedia-Social and Behavioral Sciences*, *46*, 3535-3540. https://doi.org/10.1016/j.sbspro.2012.06.100

Ziegler, L. & Garfield, J. (2017). Developing a statistical literacy assessment for the modern introductory statistics course. *Statistics Education Research Journal. 17*(2), 161-178. http://iase-web.org/documents/SERJ/SERJ17(2)_Ziegler.pdf