Draft date: 2019.3.29

Proper Data Description

<Material provided by>
AMED “International Journals Project”

Contents

Introduction

In conducting research involving human subjects, knowing the characteristics of the people from whom research data has been collected will be the key for ascertaining the types of people to whom the research results will be applicable.

Most research papers describe the "background of research subjects" at the beginning of the Results section. There summarized are the information on the research subjects who provided data, such as the average age and the gender ratio. Such organization of data is referred to as "data description" or "data summation," and the summarized data is called "descriptive (or summary) statistics." A mistake in this data description could mislead readers. Please do not waste the valuable data you have collected. A proper data description is the first step in statistical analysis.

In this module, you will learn proper ways to describe data.

Learning Objectives

- To understand the types of data.
- To learn the difference between the mean and median, and how to use them.
- To learn the difference between the standard deviation and standard error, and how to use them.
- To understand the meaning and features of the 95% confidence interval.

Types of Data

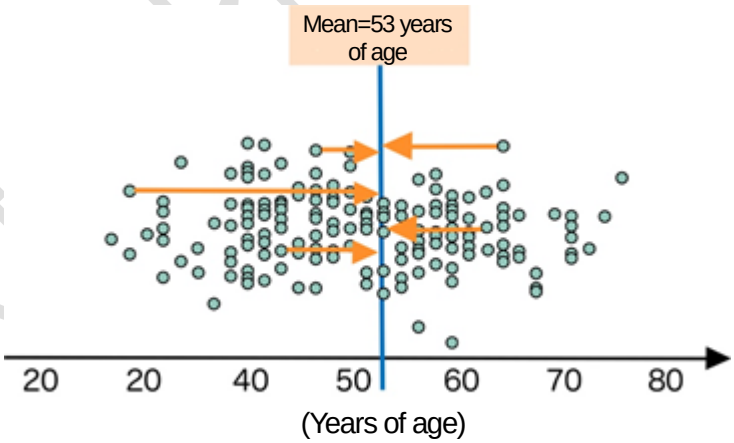
How to describe data depends on the types of data. Therefore, it is important to pay attention to the type of data you intend to describe. There are roughly two types of data: *categorical data*, which is sorted by category such as men and women; and *continuous data*, which consists of continuous values such as age, body weight, and blood pressure. Categorical data is described by way of *frequency* and *proportion*. For example, if 30 persons out of 50 persons are male, the frequency and proportion of males are 30 and 60%, respectively. Continuous data is described by way of a representative value which is a measure of central tendency (e.g., the *mean* or *median*), and variability which is measured by the standard deviation and interquartile range (IQR). Generally, variability is measured by the standard deviation when the mean is used while it is measured by the interquartile range when the median is used.

Mean and Standard Deviation (SD)

Mean – a Measure of Central Tendency of Data

A mean is a value obtained by dividing the sum of all values by the number of values in the data set. For example, let us suppose five research subjects, who are aged 10, 20, 30, 40, and 50, respectively. The mean of the ages of these five subjects is 30 years of age, as obtained by dividing the sum of ages of all subjects (150) by the number of subjects (5). Such value that indicates the central position of the observed values is referred to as a "measure of central tendency." However, a measure of central tendency alone is not enough to clearly demonstrate the distribution pattern of the observed values (the ages of the five subjects in this example). In order to make the data description clearer, it is necessary to indicate the variability of data, in addition to the measure of central tendency.

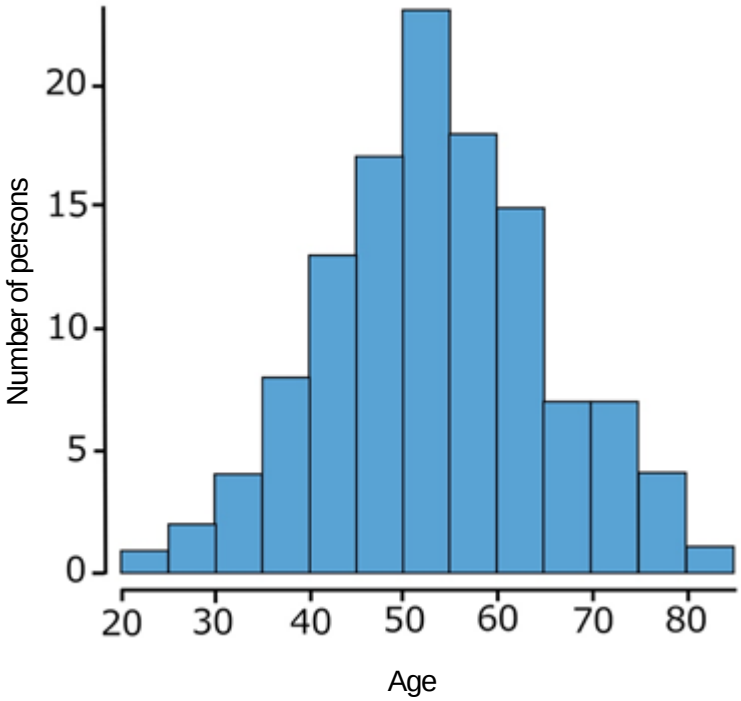
Standard Deviation – a Measure of Variability of Data



In the example mentioned above, the mean of the ages of the subjects is 30 years, but persons who are older or younger than this age are contained in this data set. To understand the data correctly, it is necessary to ascertain how much the data varies, with the mean at its center. A common indicator of data variability is the standard deviation (SD), which can be described as a concept meaning the "average distance between each observed value and the mean." The above graph shows the ages of 100 persons. The distance between the observed value of 63 years of age and the mean of 53 years of age is 10 years. The standard deviation is obtained by computing such distances for all persons and taking the square root of the average(*Note) of the square sum of these distances. The shorter the distance between each value and the mean, the smaller the standard deviation, meaning a low variability. Conversely, a large standard deviation means that the values vary widely.

(*Note) Strictly speaking, the value obtained by dividing the square sum of (sample average – each observed value) by (n-1) is used. Further details are beyond the scope of this module and therefore omitted.

The following figure shows the age data of 100 persons in a graph called histogram.

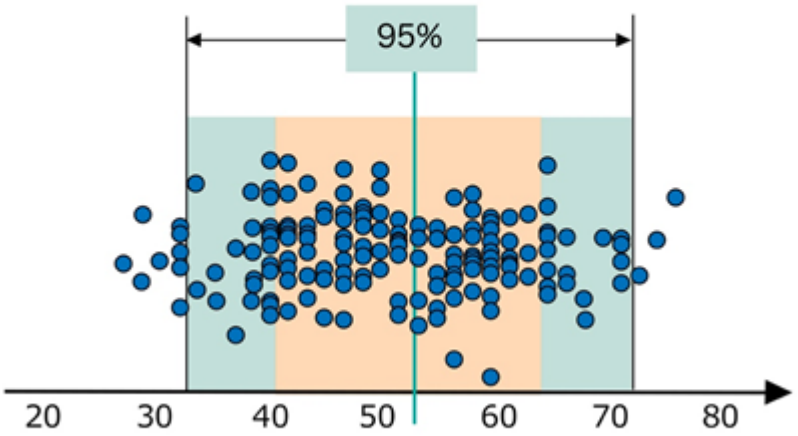


This graph illustrates the number of persons in each category when the subject persons are categorized into groups set at the interval of five years of age, e.g., 1 person in the 20-25 group, 2 persons in the 25-30 group, 3 persons in the 30-35 group and the like. It shows that the number of persons in the 50-55 group, which includes the average age of 53, is the largest, and the number of persons decreases almost symmetrically as the data points move away from the center. When the data is distributed almost symmetrically around the mean, it is assumed that the data follows a *normal distribution*.

When it is assumed that the data follows a normal distribution, it can be statistically thought that the data values of approximately 95% of all research subjects are distributed within the range from the mean minus the standard deviation to the mean plus the standard deviation (53 ±11 ×2), that is, between 31 and 75 years of age.

The ages of 95% of all subjects are distributed within the range of "the mean ± 2SD".

$53 \pm 2 \times 11 = (31.75)$

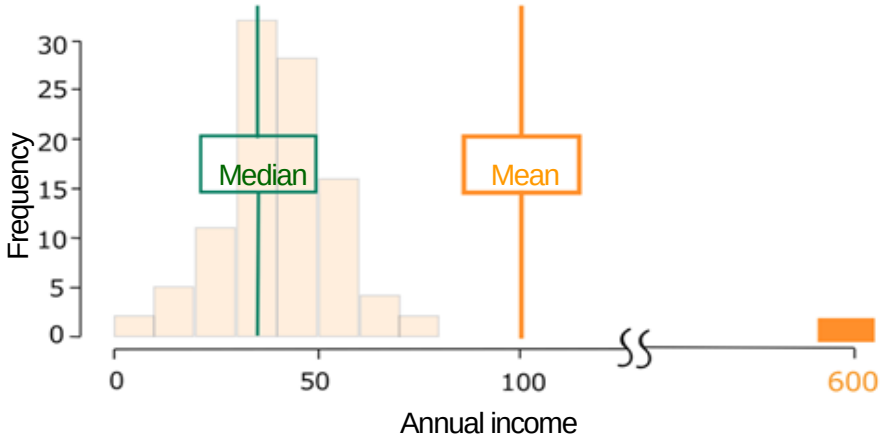


Median and Interquartile Range

Let us consider data that does not follow a normal distribution.

Problems with the Mean, and the Median

Suppose you conducted a street survey and collected annual income data from 101 persons. For the first to 100th persons, the annual income was within an ordinary range between 1 million yen and 7 million yen. But, as the 101st person, you met a professional baseball



player who earns 600 million yen a year. As a result, the mean or the central value of the data set for the 101 persons comes near to 10 million yen. Then, can you say that this mean value is really a measure of central tendency of the whole data set?

The figure above is a histogram of this data set. The annual income of the subjects other than the baseball player is within the range between 1 million yen and 7 million yen. Therefore, these 100 subjects earn only 7 million yen per person at maximum. It would be a wrong interpretation if you thought that the subjects in this group earn 10 million yen on average.

As this example suggests, the mean (10 million yen) is calculated under the influence of an extreme value (600 million yen), although the actual central value of the data set is lower than that. In such a case, the mean is not appropriate as a measure of central tendency of the data set.

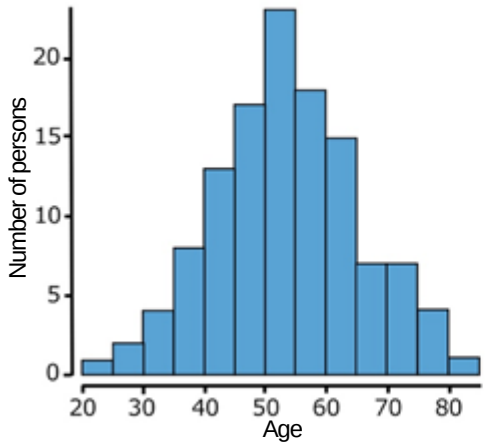
When a value that has an extreme deviation from the mean exists (such value is called an "outlier") or the data does not follow a normal distribution, the median is used as a measure of central tendency instead of the mean. In the annual income data mentioned above, the median is 4 million yen, which corresponds to the annual income earned by the person in the middle of the persons arranged in the order of the amount of annual income. If the median is used, the annual income earned by the 51st person who is in the middle of the 101 persons comes at the center of this data set even when one of these persons has an extremely large value of 600 million yen. Thus, it is possible to avoid a value affected by an outlier, such as the mean of 10 million yen, from being selected as a measure of central tendency of the data set.

Problems with the Standard Deviation, and the Interquartile Range

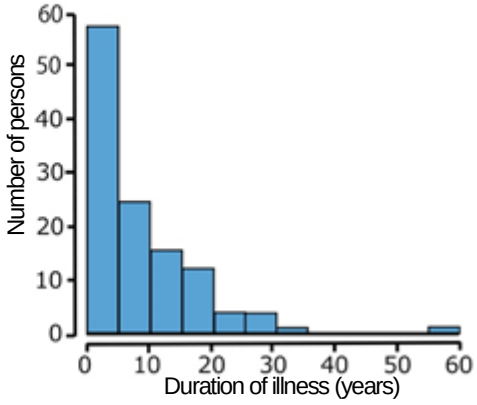
When the median is used as a measure of central tendency, the data variability is expressed with the interquartile range, which refers to the range between the 25th percentile point (first quartile point) and the 75th percentile point (third quartile point). Therefore, the interquartile range contains 50% of the values in the data set. The 50th percentile point (second quartile point) corresponds to the median. In the example mentioned above, the median annual income is 4 million yen, and the interquartile range is between 3 million yen and 6 million yen, which suggests that half of the subjects earn annual income of between 3 million yen and 6 million yen.

Normal Distribution and Medical Data

Following a normal distribution



Not following a normal distribution



As briefly mentioned earlier in this module, a normal distribution refers to a symmetrical, bell-shaped distribution (the graph on the above left). If the data sample shows that the number of data values around the mean is the largest and it gradually becomes smaller evenly as the values deviate away from the mean to the right and left, it is assumed that the variables in the population from which the sample is extracted follow a normal distribution. For example, the age data is assumed to follow a normal distribution, but this does not always apply to other categories of data. It is often the case that the data used in medical research cannot be assumed to follow a normal distribution (as is shown in the graph on the above right), and the use of the means and standard deviation may lead to misunderstanding.

Which Should Be Chosen, the Mean (Standard Deviation) or the Median (Interquartile Range)?

Earlier in this module, you learned that the mean and standard deviation are commonly used to describe data if the data can be assumed to follow a normal distribution, whereas the median and interquartile range are more desirable if a normal distribution cannot be assumed. Then, should the choice between the mean and the median be made after confirming whether the data can be assumed to follow a normal distribution? While some research papers use both the mean and the median depending on the type of data, it has become popular to describe data only by the median and interquartile range regardless of the type of data, because the mean and the median are nearly equal when the data can be assumed to follow a normal distribution.

Standard Error (SE) and Confidence Interval (CI)

Standard Error

As explained above, the standard deviation expresses the variability of data. This section explains standard error.

Let us suppose a research project to verify the effect of an antihypertensive drug. At Institute A, this drug was given to 100 subjects, and the blood pressure dropped by 20 mmHg on average. Then, can you say that this drug has an effect of lowering the blood pressure by 20 mmHg for the same type of people as the research subjects? The fall in blood pressure observed in this research is only a result derived from the data at hand, and you cannot be sure if the same effect can be achieved for all people around the world without giving them the same drug and collecting the data on their blood pressure. The process of analyzing the data at hand and inferring the effect of an antihypertensive drug based on the data of all people around the world is called "statistical inference." In this example, the effect of the antihypertensive drug of "lowering the blood pressure by 20 mmHg" has been inferred from the data at hand. If a similar research project is carried out at another institute, is it possible to obtain a similar result? For example, if the same drug is given to 100 subjects at Institute B, Institute C, Institute D… and so on, and the blood pressure data is collected at 100 institutes, it is unlikely that the same effect as that observed at Institute A, i.e., lowering the blood pressure by 20 mmHg on average, can be achieved at all these institutes. There could be variability with the estimate value of the average blood pressure-lowering effect if the research is conducted repeatedly. It is nearly impossible and unrealistic to conduct the same research at 100 institutes, but it is possible to theoretically calculate how much the estimate value (the variation in blood pressure) would vary if the same research is conducted many times. Such theoretical variability of the estimate value is called "standard error." A standard error can be calculated based on mathematical grounds, by dividing the standard deviation, which demonstrates the variability of data, by the square root of the number of research subjects.

Standard error = Standard deviation /√ Number of research subjects

For example, when the mean of the blood pressure data collected from 100 persons is 80 mmHg, and the standard deviation is 10 mmHg, then the standard error is calculated as 10/√100=1. As this formula indicates, the number of research subjects is taken as the denominator, hence the more research subjects involved, the smaller the standard error becomes.

Standard Error and True Value

As you learned earlier in this module, a standard deviation is a concept meaning the average distance from the mean. A standard error indicates the variability of the mean with regard to the data sets obtained in an unlimited number of similar research projects, each conducted with the same number of subjects. The variability expressed by a standard error signifies the average distance between each mean and the *true value*. The "true value" in the example above means the effect of the antihypertensive drug on all people around the world – a value that only God knows.

This explanation may be too abstract and difficult to understand. Try to imagine that you were God. Nobody but you know that this antihypertensive drug has an effect of lowering the blood pressure by 15 mmHg. When you look down at the human world, the researchers at Institute A say that the drug can lower the blood pressure by 20 mmHg, and you may think, "Oh, they have come close." When you turn to Institute B, the researchers have concluded that the drug can reduce the blood pressure by 30 mmHg, then you may think how ridiculous they are. Only you know the true value of the drug's antihypertensive effect and how far the values estimated by the researchers are away from it. You may also be able to draw a distribution chart with the true value positioned at the center and describe the true data (true value ± standard error). However, researchers cannot do this because they know nothing of the true value. Therefore, researchers give up an attempt of expressing a true mean that they can never know, and instead use the concept of confidence interval (CI) rather than the mean of the data they have collected.

Confidence Interval

The range expressed with the estimate value (e.g., the mean) and the standard error derived from data is called the "confidence interval." If the estimate value can be deemed (assumed) to follow a normal distribution, the range calculated by the formula "estimate value ±2× standard error" (*Note) is referred to as the "95% confidence interval." The confidence interval represents the precision of the value estimated from the data at hand. The wider the confidence interval, the less precise the estimate is, and vice versa.

(*Note) Strictly speaking, the formula is: "estimate value ±1.96× standard error."

For example, if the average blood pressure among 100 subjects is 80 mmHg, and the standard deviation is 30 mmHg, the standard error is $30/\sqrt{100}=3$. If the 95% confidence interval is calculated with the average blood pressure of 80 mmHg at its center, the lower limit is $80-2\times3=74$ and the higher limit is $80+2\times3=86$, and therefore the interval is estimated as [74, 86]. As indicated in the formula of the

standard error, the number of research subjects is taken as the denominator, hence the more research subjects involved, the smaller the standard error becomes. With the same average and standard deviation, if the number of research subjects is 10,000, the confidence interval is estimated as [79.4, 80.6] (the lower limit: 80-2×(30/100)=79.4; the higher limit: 80+2×(30/100)=80.6). Thus, the more research subjects involved, the more precise the estimate becomes.

This module was prepared by "Ethics Education Program on the Research Reliability Standards of International Medical Journals" (or "AMED International Journals Project") with funding support from "Research and Development of Material/Program for Research Integrity Education" of the Japan Agency for Medical Research and Development. Please visit the APRIN website for the names of the experts who prepared and/or reviewed this module.

Below are the checkpoints from the five international journals that are related to the content of this module (these descriptions are partially paraphrased to help you understand them).

[1] Nature (http://image.sciencenet.cn/olddata/kexue.com.cn/upload/blog/file/2010/12/20101282125135575 01.pdf; visited on 2019.01.24)

[2] New England Journal of Medicine (http://www.nejm.org/page/author-center/manuscript-submission#electronic; visited on 2019.01.24)

[3] Science (http://www.sciencemag.org/authors/science-editorial-policies; visited on 2019.01.24)

[4] The EMBO Journal (http://emboj.embopress.org/authorguide#statisticalanalysis; visited on 2019.01.24)

[5] JAMA (http://jamanetwork.com/journals/jama/pages/instructions-for-authors; visited on 2019.01.24)

[1] Nature

- The sample size for each data set used should be given.
- A clearly labeled measure of central tendency (e.g. mean or median) should be given.
- A clearly labelled measure of variability (e.g., standard deviation or interquartile range) should be given.
- All numbers following a ± sign should be identified as standard errors (s.e.m.) or standard deviations (s.d.).

[2] New England Journal of Medicine

- Measures of uncertainty of data, such as confidence intervals, should be used consistently from the beginning to the end. The same applies to figures that illustrate the results.

[3] Science

- Data pre-processing steps such as transformations, re-coding, re-scaling, normalization, truncation, and handling of below-detectable level readings and outliers should be fully described; any removal or modification of data values must be fully acknowledged and justified.
- Descriptive statistics should be presented for variables that are integral to subsequent analyses and interpretation of the study findings. The number of sampled units (N), mean, median and other items upon which each reported statistic is based must be stated.
- For continuous variables that are approximately normally distributed, mean and SD are suitable measures for center and dispersion, respectively. For continuous variables with asymmetrical distributions, median and range (or interquartile range) are preferred to mean and SD. All measures of central tendency or dispersion that are used should be identified.
- For very small samples sizes (e.g., N < 20), presentation of all data values in tabular format is desirable unless presentation would violate restrictions for privacy or confidentiality for human subjects. Units

should be supplied for all measurements.

- Point estimates of population parameters (e.g., mean, correlation coefficient, slope) or comparative measures (e.g., mean difference, odds ratio, hazard ratio) should be accompanied by a measure of uncertainty such as a standard error or a confidence interval.

[4] The EMBO Journal

- Descriptive statistics should include a clearly labelled measure of center (such as the mean or the median), and a clearly labelled measure of variability (such as standard deviation or range). Ranges are more appropriate than standard deviations for small data sets. Standard error or confidence interval is appropriate to compare data to a control.

[5] JAMA

- In the “Result” section, when possible, present numerical results with appropriate indicators of uncertainty, such as confidence intervals.
- It is generally not necessary to provide a detailed description of the methods used to generate summary statistics, but the tests should be briefly noted in the Methods section.
- In the reporting of results, when possible, quantify findings (e.g., frequency, rates) and present them with appropriate indicators of measurement error or uncertainty, such as confidence intervals.
- Use means and standard deviations (SDs) for normally distributed data and medians and ranges or interquartile ranges (IQRs) for data that are not normally distributed.