QBS
Vol. 36, No. 1, pp. 1-6 (2017)
https://doi.org/10.22283/qbs.2017.36.1.1
pISSN 2288-1344 eISSN 2508-7185 **Review Article**

Quantitative Bio-Science
Official Journal of the Institute of Natural Science at Keimyung University
Available Online at https://qbs.kmu.ac.kr:442/

P-value and Reduction to Absurdity

Tae Yoon Kim[1], Mi-Kyung Choi[2], Hee Soo Lee[3,*]

[1]Department of Statistics, Keimyung University, Daegu 42601, Korea
[2]Department of Food Science and Nutrition, Keimyung University, Daegu 42601, Korea
[3]Departments of Business Administration, Sejong University, Seoul 05006, Korea
(Received April 13, 2017; Revised May 9, 2017; Accepted May 10, 2017)

ABSTRACT

Statistical hypothesis testing is a procedure that verifies a given hypothesis through statistical analysis, using collected data. Hypothesis testing employs the p-value to draw conclusions about the underlying population via sampled data. The p-value, a calculated probability from sampled data, is an important tool in determining whether to reject the null hypothesis. However, most users seem to have difficulties understanding the logic behind it (e.g., [1]). Hence, this paper aims to help users understand and use the p-value adequately by elaborating on its logical meaning.

Key words : P-value, Power, Reduction to absurdity

1. Reduction to Absurdity

The logic of using a p-value for hypothesis testing is closely related to the "reduction to absurdity," a mathematical proof technique. This technique assumes that an existing fact is true and creates a critical contradiction to this assumption to conclude that the new finding is indeed true, such as when it is applied to prove that "$\sqrt{2}$ is an irrational number." In the process of proving this, we assume the starting hypothesis (null hypothesis) "$\sqrt{2}$ is a rational number expressed as an irreducible fraction," and find a critical contradiction to the null. Since a critical contradiction is found from assuming the null hypothesis (i.e., $\sqrt{2}$ cannot be expressed as an irreducible fraction as assumed initially), the null is not supported. Therefore, we are able to prove that a new hypothesis (alternative hypothesis) "$\sqrt{2}$ is an irrational number" is true. The logic of the reduction to absurdity in this problem shows the null hypothesis as an existing hypothesis and the alternative hypothesis as a new hypothesis to be proved.

* Correspondence should be addressed to Dr. Hee Soo Lee, Department of Business Administration, Sejong University, Seoul 05006, Korea. Tel: +82-2-3408-3177, Fax: +82-2-3408-4310, E-mail: heesoo@sejong.ac.kr

2. Practical Episode and Uncertainty

Let us consider a simple but practical episode to explain how hypotheses testing based on the logic of the reduction to absurdity works. Consider a couple who have been dating for about a year. One day, Sam calls Kate, his girlfriend. However, Kate has neither answered his calls nor returned them for four weeks. Consequently, it is only natural that Sam wants to know if Kate desires to break up with him. In this couple episode, the null and alternative hypotheses are as follows:

H_0: Kate is still interested in Sam (existing hypothesis).
H_1: Kate wants to break up with Sam (new hypothesis to prove).

Sam applies the following logic of the reduction to absurdity to test the above hypotheses.

"Let us assume Kate is still interested in me (null hypothesis). Then, it is a contradiction for her having lost contact with me for the past four weeks. Therefore, she wants to break up with me (alternative hypothesis)."

In this episode, the difficulty in applying the logic of the reduction to absurdity is that it is uncertain whether the observed result of Kate having lost contact with Sam for four

weeks is a critical contradiction to the null hypothesis that she is still interested in him. For instance, if the couple have exchanged phone calls every day for the last one year, it is almost impossible for Kate to have lost contact with Sam for a month. On the other hand, if the couple did not call each other frequently (i.e., assume they called each other once every three weeks), then it is somewhat likely for Kate to have lost contact with Sam for a month. In the former case, based on the logic of the reduction to absurdity, Sam can safely conclude that Kate wants to break up with him. In the latter case, it is difficult for him to conclude that Kate wants to break up with him because it is hard to establish a convincing contradiction to the null hypothesis. Here, judgment should be made via probability of obtaining the observed result (Kate having lost contact with Sam for four weeks), calculated from the couple's phone call frequency during the last one year, when Kate had been interested in Sam (under the null hypothesis).

3. P-value for Practical Episode

A dictionary definition of the p-value is "the probability of obtaining a result equal to or 'more extreme' than what was actually observed (*event A*), when the null hypothesis is true." A "more extreme" result in this definition supports the alternative hypothesis established by the user testing hypothesis. In the above couple episode, *event A* refers to Kate losing contact with Sam for four weeks (actual observed result) or more than four weeks (more extreme result) because the alternative hypothesis is "Kate wants to break up with Sam." Therefore, no contact for "more than" four weeks supports the alternative hypothesis and *event A* is the case of Kate losing contact with Sam for equal to or 'more than' four weeks. The p-value in this episode is defined as the following probability:

$$\text{p-value} = \text{P}(\text{Kate lose contact for equal to or more than four weeks} \mid \text{Kate is still interested in Sam}).$$

Note that under the null hypothesis that Kate is still interested in Sam, the calculated p-value should decrease as the period during which Kate loses contact with Sam increases. This is logical because, if the observed result of Kate having lost contact with Sam for four weeks is judged to be a critical contradiction to the null hypothesis, no contact for "more than" four weeks should be also judged to be contradictory to the null hypothesis.

If we look at the p-value according to the above logic, a small p-value indicates a small chance that Kate, if interested in Sam, loses contact with him for equal to or more than four weeks. In this case, Sam had better reject the null hypothesis, because Kate is very likely to have broken up the relationship. With a larger p-value, Sam might conclude that it is probable for Kate, who is still interested in Sam, to have lost contact with him for equal to or more than four weeks. As such, he might adhere to the null hypothesis since Kate still might consider him her boyfriend. For illustration purpose with a specific p-value, let us say that the calculated p-value is 5%. This means that the probability that Kate lost contact with Sam for equal to or more than four weeks, while still being interested in him, is 5%. If Sam considers the 5% probability small (i.e., it is very unlikely for Kate to have having lost contact with him for equal to or more than four weeks), he naturally concludes that Kate wants to break up with him. This suggests it is essential to set a predetermined criterion that allows us to determine "the level of p-value smallness" when the user tests a hypothesis via a calculated p-value.

4. P-value Duplicity

There are two possible cases for Kate losing contact with Sam, one being her intention to leave him. The other refers to understandable circumstances occurring, such as her being ill or losing her phone. If the former applies, then Sam is correct in rejecting the null hypothesis due to the small p-value of 5% or the unlikelihood of the event. In the latter case, the probability that Kate lost contact with him for a month due to other understandable reasons despite still being interested in him is also 5%. Therefore, Sam has to tolerate this 5% probability of making an error if he considers the 5% p-value small and rejects the null hypothesis, concluding that Kate wants to break up with him. Thus, the p-value of 5% also indicates the error probability. As such, the p-value can be seen with duplicity (i.e., unlikelihood or committing an error).

Duplicity leads to two possible ways for determining the level of a small p-value. From the perspective of unlikelihood, the level is called *significance level usually denoted by* α because it determines whether the unlikelihood due to the observed data is significant enough to contradict or reject the null hypothesis. From the perspective of committing an error, the level of the small p-value is considered as the maximum

allowable probability of making a type I error, which rejects a "true" null hypothesis. In this sense, the level is also called *test size* α.

5. Power and P-value

As previously discussed, the p-value is a probability calculated under the null hypothesis. The probability of *event A* (a result equal to or "more extreme" than actually observed) can also be calculated under the alternative hypothesis, which is usually called the *power* of test. In the couple episode, the probability of *event A* (Kate losing contact with Sam for equal to or more than four weeks) under the assumption that Kate wants to break up with Sam (alternative hypothesis) is the power. Sam should reject the null hypothesis and conclude that Kate wants to break up with him if the power is high. This *power* is generally expressed as $1-\beta$, where β indicates the probability of rejecting a true alternative hypothesis (type II error). Here, β is the probability that Sam incorrectly concludes that Kate is still interested in him:

power = P (Kate has lost contact with Sam for equal to or more than four weeks | Kate wants to break up with Sam) = $1-\beta$.

Logically, the *power* is the probability of making a correct decision that rejects the null hypothesis when the alternative hypothesis is true. In principle, *it is possible to test a hypothesis using the power, but it is not recommended because event A's distribution is difficult to be found under the alternative hypothesis (new hypothesis)*. Here, it is difficult to calculate the probability under the alternative hypothesis (she wants to break up with him) that Kate has lost contact with Sam for equal to or more than four weeks due to lack of previous data under the alternative hypothesis.

6. P-value Calculation for Coin Tossing

The coin tossing is an example where it is easy to calculate the probability of *event A* under the null and alternative hypotheses. This clarifies the practice of the p-value and power. Consider an experiment where we toss a coin five times and observe the number of heads. If the coin is fair, the probability of obtaining heads or tails is 1/2. If you observe four heads out of five tosses, the question is whether it can be concluded the coin is fair. The null hypothesis for testing whether the coin is fair using *observed data (i.e., four heads out of five tosses)* is established as follows:

H_0: The probability of heads from tossing the coin is 1/2 ($P(H) = 1/2$).

With the experiment results at hand, a reasonable alternative hypothesis would be:

H_1: The probability of heads from tossing the coin is above 1/2 ($P(H) > 1/2$).

Indeed, the result of four heads out of five tosses makes the above alternative hypothesis reasonable because the probability of obtaining heads seems to be greater than 1/2. However, other types of alternative hypotheses can be established. If the user strongly believes before tossing the coin that the probability of heads is below 1/2, to test whether it is true, then he or she can state the alternative hypothesis as follows:

H_1: The probability of heads from tossing the coin is below 1/2 ($P(H) < 1/2$).

If the user does not care whether the probability of obtaining heads is greater or smaller than 1/2, but is only concerned with whether the coin is fair or not, then he or she states the alternative hypothesis as follows:

H_1: The probability of heads from tossing the coin is not equal to 1/2 ($P(H) \neq 1/2$).

As such, the alternative hypothesis depends on what the user wants to verify. How do the above three alternative hypotheses with the same null hypothesis affect hypotheses testing? The p-value and power are determined by the alternative hypothesis, resulting in different conclusions. Let B be the number of heads from tossing a fair coin five times. Fig. 1 shows the probability mass function of B under the null hypothesis.

Fig. 1. Results from the following calculations:

$P(B=0 \mid P(H)=1/2) = 1/32$, $P(B=1 \mid P(H)=1/2) = 5/32$,

$P(B=2 \mid P(H)=1/2) = 10/32$, $P(B=3 \mid P(H)=1/2) = 10/32$,

$P(B=4 \mid P(H)=1/2) = 5/32$, $P(B=5 \mid P(H)=1/2) = 1/32$.

Let us consider the first alternative hypothesis, H_1: $P(H) > 1/2$, and calculate the p-value. In this case, *event A* (a result equal to or "more extreme" than what was actual observations) refers to the results of four or five heads out of five

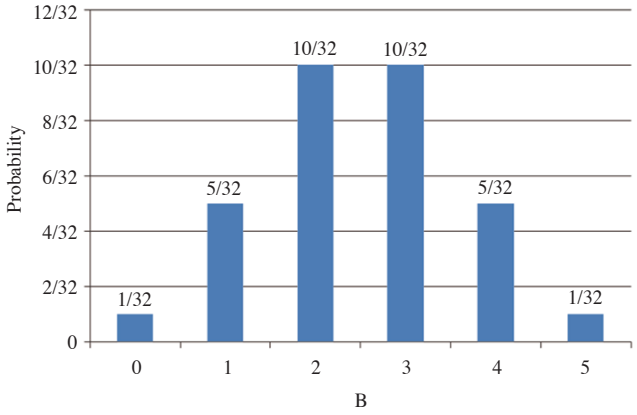

Fig. 1. A probability mass function of B under the null hypothesis (P(H) = 1/2).

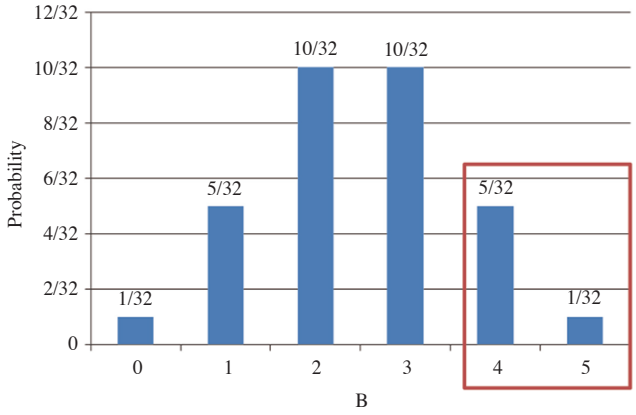

Fig. 2. *Event A* for p-value under H_1: P(H) > 1/2 and H_1: P(H) ≠ 1/2: A red box contains *event A*, needed to calculate p-value corresponding to alternative hypothesis H_1: P(H) > 1/2 and an *empirically defined p-value* p_0 corresponding to alternative hypothesis H_1: P(H) ≠ 1/2.

tosses, because the extreme results supporting the alternative hypothesis that the probability of heads is above 1/2 are the cases with more number of heads. Therefore, the p-value is calculated as follows:

$$\text{p-value} = P(B \geq 4 \mid P(H) = 1/2) = 6/32.$$

At a significance level of 5%, the user does not reject the null hypothesis, since the p-value is above 5% (6/32 > 0.05), and concludes the coin is fair.

Let us consider the second alternative hypothesis, H_1: P(H) < 1/2, and calculate the p-value. In this case, *event A* (a result equal to or "more extreme" than actual observations) refers to the results of four heads or less out of five tosses, because the extreme results supporting the alternative hypothesis that the probability of heads is below 1/2 are the cases with fewer number of heads. Therefore, the p-value is calculated as follows:

$$\text{p-value} = P(B \leq 4 \mid P(H) = 1/2) = 31/32.$$

At a significance level of 5%, the user does not reject the null hypothesis, as the p-value is greater than 5% (31/32 > 0.05), and concludes that the coin is fair.

The last alternative hypothesis, H_1: P(H) ≠ 1/2 makes it difficult to determine the extreme results supporting H_1 because it does not specify whether the coin is biased towards heads or tails. For this type of alternative hypothesis, the extreme results used for calculating the p-value being empirically defined from the observed results and the given null hypothesis, an *empirically defined p-value* p_0 is calculated for extreme results. Then, the p-value is calculated by doubling this *empirically defined p-value* p_0 (i.e., p-value = $2 \times p_0$). In the example, the extreme result for calculating p_0 is $B \geq 4$, since the observed result ($B = 4$) is greater than the expected value of B (E(B) = 2.5) under the null hypothesis, H_0: P(H) = 1/2. Therefore, the p-value corresponding to this alternative hypothesis is calculated as follows:

$$\text{p-value} = 2 \times p_0 = 2 \times P\left(B \geq 4 \mid P(H) = \frac{1}{2}\right) = 2 \times \frac{6}{32} = 12/32.$$

At a significance level of 5%, the user does not reject the null hypothesis because the p-value is above 5% (12/32 > 0.05) and concludes that the coin is fair. Figs. 2 and 3 respectively show *event A* inside the red box corresponding to alternative hypotheses H_1: P(H) > 1/2 and H_1: P(H) ≠ 1/2, and H_1: P(H) < 1/2.

7. Power Calculation for Coin Tossing

Let us calculate the power of this experiment. As previously discussed, the power is a probability of *event A* (a result equal to or "more extreme" than actual observations), as defined from the p-value calculation under the alternative hypothesis. As such, a specific case under the alternative hypothesis has to be assumed to calculate it. For example, under the first alternative hypothesis H_1: P(H) > 1/2, we assume P(H) = 2/3. Fig. 4 shows the probability mass function of B when tossing a coin with P(H) = 2/3 five times. The red box in Fig. 4 contains *event A* needed to calculate the power.

In Fig. 4, the distribution of B with P(H) = 2/3, which is needed to calculate the power, results from the following cal-

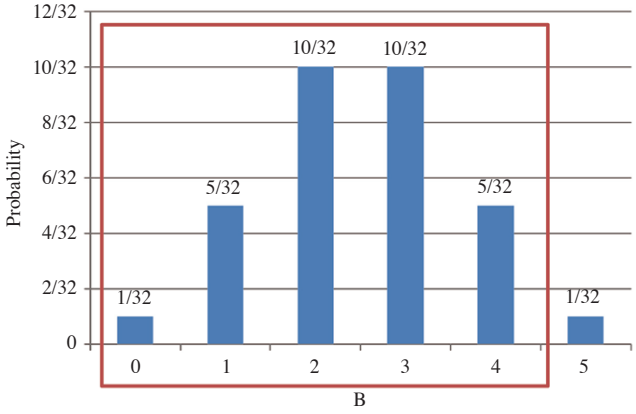

Fig. 3. *Event A* for p-value under H_1: P(H) < 1/2: A red box contains *event A*, needed to calculate p-value corresponding to the alternative hypothesis H_1: P(H) < 1/2.

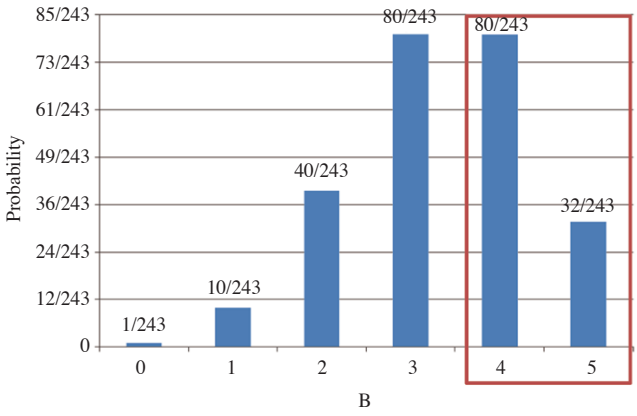

Fig. 4. A probability mass function of B with P(H) = 2/3: A red box contains the *event A* needed to calculate power under the alternative hypothesis H_1: P(H) = 2/3 (P(H) > 1/2).

culations:

$P(B=0 \mid P(H)=2/3)=1/243$, $P(B=1 \mid P(H)=2/3)=10/243$,

$P(B=2 \mid P(H)=2/3)=40/243$, $P(B=3 \mid P(H)=2/3)=80/243$,

$P(B=4 \mid P(H)=2/3)=80/243$, $P(B=5 \mid P(H)=2/3)=32/243$.

Therefore, the power of event A is $P(B \geq 4 \mid P(H)=2/3)=112/243$.

Under the second alternative hypothesis H_1: P(H) < 1/2, we set P(H) = 1/4. Fig. 5 shows the probability mass function of B when tossing a coin with P(H) = 1/4 five times. The red box in Fig. 5 contains *event A* needed to calculate the power.

In Fig. 5, the distribution of B with P(H) = 1/4, which is needed to calculate the power, results from the following calculations:

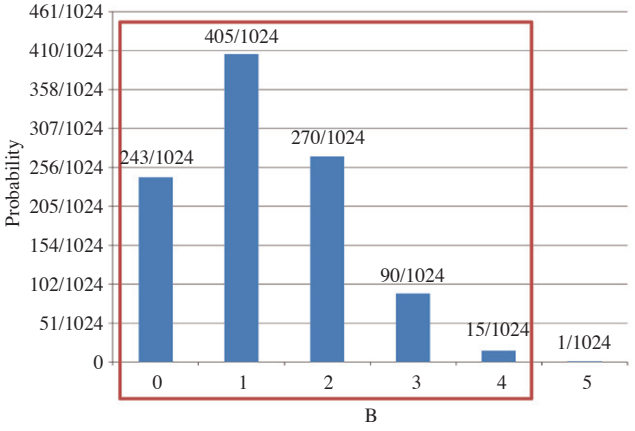

Fig. 5. A probability mass function of B with P(H) = 1/4: A red box contains the *event A* needed to calculate power under the alternative hypothesis H_1: P(H) = 1/4 (P(H) < 1/2).

$P(B=0 \mid P(H)=1/4)=243/1024$, $P(B=1 \mid P(H)=1/4)=405/1024$,

$P(B=2 \mid P(H)=1/4)=270/1024$, $P(B=3 \mid P(H)=1/4)=90/1024$,

$P(B=4 \mid P(H)=1/4)=15/1024$, $P(B=5 \mid P(H)=1/4)=1/1024$.

Therefore, the power of event A is $P(B \leq 4 \mid P(H)=1/4)=1023/1024$, which is extremely high.

8. P-value with Real Data

Let us consider results from real data analysis. Since most statistical analysis programs provide a p-value under the two-sided hypothesis (or $2 \times p_0$) in their algorithms, the user must be careful about using correctly the calculated p-value. In finance, individual stock systematic risk is estimated by a market model that follows linear regression with the independent variable of market portfolio return (R_M) and dependent variable of individual stock return (R_i):

$$R_{it}=\alpha+\beta R_{Mt}+\varepsilon_t,$$

where ε_t is an independent error and β represents changes in individual stock return when the market portfolio return changes by one unit, thus indicating an effect of the market portfolio return on the individual stock return. The hypotheses to test whether β is zero are as follows:

H_0: The effect of the market portfolio return on individual stock return is 0 ($\beta=0$).

Table 1. Market model estimation results

Parameter	Estimate	Standard error	t-value	p-value
Intercept (α)	−0.0253	0.0229	−1.1078	0.2702
R_M (β)	0.2008	0.0869	2.3106	0.0226

H_1: The effect of the market portfolio return on individual stock return is different from 0 ($\beta \neq 0$).

The results from a statistical software are summarized in Table 1.

Then the p-value (or $2 \times p_0$-value) for the estimated β (0.2008) is 0.0226 with H_1: $\beta \neq 0$, 0.0113 with H_1: $\beta > 0$, and 0.9887 with H_1: $\beta < 0$. At a significance level of 5%, we reject the null hypothesis as the p-value is below 5% when H_1: $\beta > 0$ and H_1: $\beta \neq 0$. In other words, the estimated β (0.2008) critically contradicts the null hypothesis H_0: $\beta = 0$. However, if H_1: $\beta < 0$ the user could not reject the null hypothesis. Therefore, we conclude there is a significant (positive) effect of the market portfolio return of a given individual stock return.

Another example is from existing literature in the field of medical nutrition. The study investigates the effect of running nutrition education programs on borderline hypertensive adult patients ([2]). Table 2 reports results from the analysis of changes in anthropometric characteristics and blood pressure after an eight-week education program. The study then conducts paired t-tests to test significant differences in anthropometric characteristics and blood pressure between the baseline (before education) and after eight weeks (after education). The following null and alternative hypotheses are established:

H_0: The mean difference in the measurements of related variables before and after education is 0 ($\mu_D = 0$).

H_1: The mean difference in the measurements of related variables before and after education is different from 0 ($\mu_D \neq 0$).

In the above, μ_D is the mean of $D = X_A - X_B$, where X_A is measurement of the related variable after education and X_B before education for a given patient. The p-value for testing the weight difference shows that the observed weight difference contradicts the null hypothesis at a 5% significance level, concluding to a significant difference in weight between the baseline and after eight weeks. The observed differences in body mass index (BMI), percentage of body fat, and systolic blood pressure (SBP) also contradict the null hypothesis at a 0.1% significance level. Therefore, we conclude that there are

Table 2. Changes of anthropometric characteristics and blood pressure after the education program for 42 adults [2]

Variables	Baseline	Eight weeks	Difference[2)]
Weight (kg)	59.40 ± 7.24[1)]	59.00 ± 7.21	−0.39 ± 1.02*
BMI (kg/m^2)	24.23 ± 2.24	23.93 ± 2.18	−0.30 ± 0.47***
Percent body fat (%)	32.61 ± 5.88	31.57 ± 5.43	−1.04 ± 1.63***
SBP (mmHg)	133.67 ± 14.50	127.36 ± 12.89	−6.31 ± 11.81***
DBP (mmHg)	78.19 ± 10.27	77.10 ± 8.87	−1.10 ± 8.37

[1)]Mean ± SD. [2)]eight weeks-baseline. *$p < 0.05$, ***$p < 0.001$.

significant differences in these variables between the baseline and after eight weeks and that education is significantly effective. Conversely, the p-value for testing the diastolic blood pressure (DBP) difference is above 5% and the null hypothesis is not rejected. Therefore, we conclude that there is no significant difference in DBP before and after education.

However, Table 2 has some technical difficulties in providing correct information about the test results. The last column employs a range of p-values to report significant differences for each measurement variable. Since the alternative hypothesis is two-sided and the differences in the form of mean ± standard deviation could be positive or negative, it is rather confusing to conclude whether education program is effective (i.e., it yields decreases in the measurements of the four variables except DBP). These confusions could have been avoided by specifying negative one-sided hypotheses and providing the specific p-values.

9. Concluding Remarks

The p-value, a probability calculated from sampled data, is an important tool that determines whether to reject the null hypothesis. An adequate understanding of the p-value by elaborating on its logical meaning helps users utilize p-value correctly and draw logical conclusions from their data. As discussed, it is highly recommended that the user specify the alternative hypothesis more cautiously for taking advantage of the p-value for data analysis.

References

1. Nuzzo R. Statistical errors. Nature 2014;506:150-152.
2. Jung EJ, Son SM, Kwon JS. The effect of sodium reduction education program of a public health center on the blood pressure, blood biochemical profile and sodium intake of hypertensive adults. Korean J Community Nutr 2012;17:752-771.