COMPARING STRUCTURES OF STATISTICAL HYPOTHESIS TESTING WITH PROOF BY CONTRADICTION: IN TERMS OF ARGUMENT

Hiroki Otani
Graduate School of Education, Hiroshima University, Japan
(Currently, Japan Women's University, Japan)

Abstract

Modern society requires us to become statistically literate. The right and proper understanding and interpretation of statistical hypothesis testing is essential to statistical literacy, however it is frequently confused with mathematical proof by contradiction. This study explores how statistical hypothesis testing can be distinguished from mathematical proof by contradiction, that is, what is specifically different between them. To achieve the purpose in this study, their logical structures are compared using the analytical framework: argument. Consequently, four differences are found. In statistical hypothesis testing, unlike mathematical proof by contradiction, (a) its premise is mostly non-mathematical statement and not invariable, (b) contradiction in the strict sense of the word does not arise, (c) its conclusion/claim cannot always be supported and defended by its premise, and (d) defending its conclusion/claim is necessary. The analogical approach with proof by contradiction will work effectively when hypothesis testing is taught and learned, but using that approach alone hypothesis testing has the risk of being assimilated into proof by contradiction. To understand the essence of statistical hypothesis testing properly, it is necessary to compare intentionally hypothesis testing with proof by contradiction and characterize the former as not the same as the latter.

Key words: Statistical hypothesis testing, Mathematical proof by contradiction, Statistical literacy, Toulmin's argument model, Analogical and comparative approach

INTRODUCTION

Contemporary society is filled with statistical data. Data are processed using statistical methods and are embedded in society as statistical information. Our daily lives are closely related to statistical information, and we heavily consume large amounts of such information irrespective of us being aware of it. This aspect of modern society requires us to become fully statistically literate. Gal (2004) describes statistical literacy as follows: "*statistical literacy* refers broadly to two interrelated components, primarily (a) people's ability to *interpret and critically evaluate* statistical information, data-related arguments, or stochastic phenomena, which they may encounter in diverse contexts, and when relevant (b) their ability to *discuss or communicate* their reactions to such statistical information, such as their understanding of the meaning of the information, their opinions about the implications of this information, or their concerns regarding the acceptability of given conclusions" (p.49). We as consumers of statistical information must have these characteristics. We tend to accept whatever statistical information we encounter without even questioning it and can thus be deceived, if we are not statistically literate. Modern society requires school education to develop statistical literacy.

In many countries, statistics is taught and learned within mathematics subject (Shaughnessy, 2007). However, many researchers have argued that statistics should not be taught and learned as part of mathematics (Cobb & Moore, 1997; Garfield & Ben-Zvi, 2008; Shaughnessy, 2007). The mathematics referred to here is not confined to pure academic mathematics. According to Wild & Pfannkuch (1999), even applied mathematics, which has considerable contact with real contexts, is incompatible with statistics because it ignores the variation of data or the variability in social or natural phenomena. In other words, not to treat statistics as mathematics is, therefore, not to ignore the variation inherent in data, and then to consider, predict, explain, or control it. Statistical literacy can only be developed by teaching and learning statistics *as statistics*.

However, statistics appears to be frequently taught as a domain within school mathematics and is deterministically learned as if it were mathematics (Garfield & Ben-Zvi, 2008; Burrill & Biehler, 2011). One of the statistical topics that tends to be taught and learned as deterministic mathematics is hypothesis testing. Vallecillos (1996) has analyzed the arguments which 436 college students used to support their opinions chosen as the answers to the items proposed as follows: “A statistical test of hypotheses correctly carried out establishes the truth of one of the two hypotheses, either the null or the alternative one. T/F. Support your answer” (p.51). As the result, it was found that only 6% of the total students understood hypothesis testing correctly and most students regarded it as a probabilistic proof of the hypothesis (Falk & Greenbarm, 1995) or a mathematical demonstration of the veracity of the hypothesis. Castro-Sotos et al. (2007) reviewed statistics education research concerning misconception about the nature of hypothesis testing, and organized them into two misconceptions: test as a mathematical (logical) proof and as a probabilistic proof of one of the hypotheses. These misconceptions result from identifying the logical structure of statistical hypothesis testing with that of mathematical proof by contradiction which is similar to it. Many students tend to fall into the illusion that the analogical approach with mathematical proof by contradiction is applicable to hypothesis testing in spite of the fact that the analogy does not actually work for hypothesis testing (Falk & Greenbarm, 1995).

Hypothesis testing can indeed appear to be superficially similar to proof by contradiction (Batanero, 2000; Batanero & Díaz, 2006; Castro-Sotos et al., 2007; Falk & Greenbarm, 1995; Liu & Thompson, 2005), because both include negation in a premise whose truth is not doubted and make some kind of indirect claim. Given that hypothesis testing is taught within school mathematics, it can easily be assumed that it has a structure similar to proof by contradiction in order to teach and learn the logic of hypothesis testing. As is often said, learning is to connect the known to the unknown, so in most cases the analogical approach works effectively. Considering the difficulty of the logic and the procedure inherent in hypothesis testing itself (Garfield & Ben-Zvi, 2008), it can be rather appropriate to utilize proof by contradiction to learn hypothesis testing. On the Internet, such topics can be found everywhere.

Identifying them, however, is harmful to scientific development, but can also disturb our daily lives in today’s highly information-oriented society. Understanding statistical hypothesis testing rightly and properly is essential to statistical literacy, because it underlies the statistical information omnipresent around us. The right and proper understanding and interpretation of it can make us not to be misled by statistical information omnipresent around us and therefore we can develop statistical literacy. Modern society requires us to distinguish statistical hypothesis testing from mathematical proof by contradiction. After learning statistical hypothesis testing using an analogical approach essential to learning, it must be distinguished from mathematical proof by contradiction, and the difference between them must be clarified in order not to identify them. What is specifically different between similar two?

The purpose of this article is to show how the logic of mathematical proof by contradiction can be distinguished from statistical hypothesis testing by comparing structures of them. Toulmin's argument (2003) is used to answer this question. Recently, there have been studies that characterize statistical inference and mathematical proof from the viewpoint of argument or argumentation (e.g., Antonini & Mariotti, 2008; Ben-Zvi, 2006; Garfield & Ben-Zvi, 2008). In Ben-Zvi (2006) and Garfield & Ben-Zvi (2008) it has been shown that the argument logic can be used to explain hypothesis testing, and Antonini & Mariotti (2008) has attempted to treat proof by contradiction as an indirect argument. Argument is a broader concept than statistical hypothesis testing and mathematical proof by contradiction; hence, it is useful for organizing and comparing their structures.

ANALYTICAL FRAMEWORK: ARGUMENT

According to Toulmin (2003), argument is a wide-range verbal act to establish and justify particular kinds of claims and conclusions. In his book, "The Uses of Argument," he analyzed elements of arguments in detail and envisioned applied logic, including the field-dependent and informal arguments that we encounter in daily life, as well as abstract and formal arguments in formal logic. Because of the broader definition of argument, the association of mathematical argument with mathematical proof has often been discussed in the context of the mathematics education research (e.g., Antonini & Mariotti, 2008).

Judging by Toulmin's characterization of argument (2003), arguments can be described in terms of six elements, including three basic elements ("Data," "Warrant," and "Claim/Conclusion") and three additional elements ("Qualifier," "Rebuttal," and "Backing"), that characterize arguments in more detail. Formal logic has only three basic elements (D, W, and C), but practical informal arguments cannot be described, using only those elements. It is considered that Toulmin's achievement renders practical informal arguments describable by adding the three additional but essential elements (Q, R, and B). Arguments can be shown geometrically as follows.

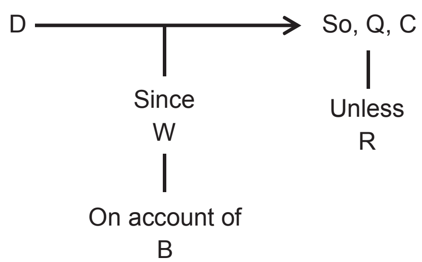

Figure 1. Layout of arguments (Toulmin, 2003, p.97)

First, there is a Claim that we want to establish. To do that, we must answer questions that might arise that criticize the Claim, for example "What do you base the claim on?" Data are the facts we present as the answer and that "we appeal to as a foundation for the claim" (Toulmin, 2003, p.90). This factual information is also evidence provided to support the Claim. Only by providing Data we can convince some critics of the Claim; however, we might still be asked again, "how did you get the claim?" In that case, to "show that, taking these data as a starting point, the step to the original claim or conclusion is an appropriate and legitimate one" (Toulmin, 2003, p.91), we present the Warrants, hypothetical statements that bridge the gap between the Data and the Claim. Warrant explains the general legitimacy and soundness of the reason why the Claim is supported by the Data. Toulmin clearly distinguishes between the facts appealing as Data and Warrants authorizing the step in the argument, stating that "data are appealed to explicitly, warrants implicitly" (p.91).

However, it is not sufficient to characterize arguments only in terms of these elements because the degree of force that the Warrant provides for the steps from the Data to the Claim is implicit. To explicitly state it, we must add two elements, a Qualifier indicating "the strength conferred by the warrant on this step" and Rebuttal conditions that indicate "circumstances in which the general authority of the warrant would have to be set aside" (Toulmin, 2003, p.94). Furthermore, in the case where some critics are not yet satisfied with the presentation of these elements, to make the Warrant acceptable to them, it is necessary to add a Backing of the Warrants that assures critics that the Warrants possess authority or currency (Toulmin, 2003, pp.95-96). Given the work by Antonini & Mariotti (2008), Backing may correspond to meta-theory. We can defend an argument using the three additional elements, even if there is a leap in the step from the Data to the Claim. A concrete example of arguments characterized by these six elements is as follows.

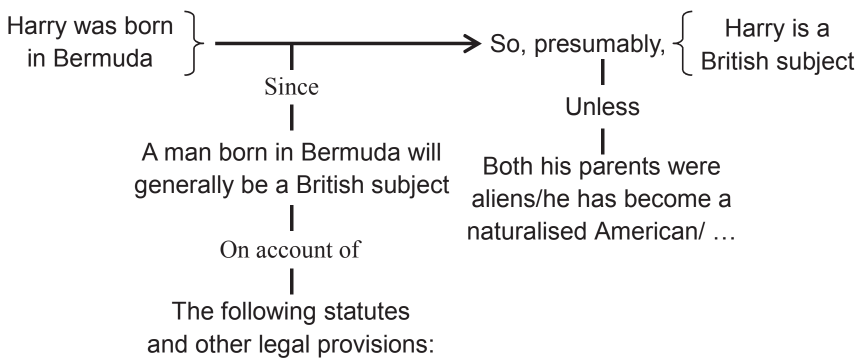

Figure 2. Concrete example of the layout of arguments (Toulmin, 2003, p.97)

From the viewpoint of argument structure provided by Toulmin (2003), both the logic of statistical hypothesis testing and mathematical proof by contradiction are Warrant. Because they entitle one to transition to Claim from Data appealed to as a foundation for the claim. However, the contents of the elements cannot be the same. In the following, Hypothesis testing and proof by contradiction are characterized from the viewpoint of argument. Proof by contradiction is considered first, with the aim of characterizing hypothesis testing by comparison with it.

LOGIC OF MATHEMATICAL PROOF BY CONTRADICTION

Mathematical proof can be classified into two types. First is direct proof, which claims that a statement Q is true based on a premise P that is supposed to be true. This follows the proper and valid logical form of modus ponens. On the other hand, we can claim indirectly that a conclusion Q is true. Mathematical proof by contradiction is one of the latter type. It is often used when we cannot derive, or it is difficult for us to derive, that a conclusion Q is true from a premise P directly.

In proof by contradiction, a formula equivalent to it is used in order to claim that a conclusion Q is true. Instead of deriving directly that a conclusion Q is the truth from two premises P and $P \rightarrow Q$ that are supposed to be true, we claim indirectly that a conclusion Q is the truth using the fact that a contradiction is derived when we suppose that both the premise P and the negation of the conclusion $\neg Q$ are true (Antonini & Mariotti, 2008). The contradiction to be derived might be the negated premise $\neg P$ against premise P that is supposed to be true, the conclusion Q against premise $\neg Q$ that is supposed to be false, or the conjunction of a statement R and its negation $\neg R$, which is different from P or Q. Note that R can include P or Q. Deriving a contradiction implies the truth of conclusion Q. Mathematical proof by contradiction, an indirect proof, is represented as follows: $P \wedge \neg Q \Rightarrow R \wedge \neg R$ (Antonini & Mariotti, 2008). This proof method is proper and formally valid. As long as the premise P is true, the conclusion Q is necessarily and certainly true. We do not state that conclusion Q "is probably true." Uncertainly and probability are not involved.

Mathematical proof by contradiction is, therefore, the method of claiming indirectly that conclusion Q is the truth using the fact that supposing the premise P and the negation of the conclusion $\neg Q$ to be the truth leads to a contradiction, and it is often used when it is difficult to claim directly that the conclusion Q is the truth (Antonini & Mariotti, 2008). Even if we can prove a statement directly using modus ponens, we can rewrite it and prove it indirectly, if we choose. For simplicity, let us consider an exemplary proof as follows. That is, we prove the conclusion Q "x+1 is an irrational number" being the truth using the promise P "x is an irrational number."

1. Negation of the conclusion Q is "x+1 is a rational number ($\neg Q$)."
2. The difference between rational numbers is a rational number.
3. x, which is a difference between "rational number x+1" and 1, is a rational number.
4. This contradicts the premise P.
5. Therefore, x+1 is an irrational number.

In this way, we claim that the conclusion Q is true in that $P \wedge \neg Q$ contradicts.

In this argument, the Claim is "Q: x+1 is an irrational number." The Data, which is the fact we appeal to as a foundation for this Claim, is that contradiction arise from supposing that both the premise "P: x is an irrational number" and the negation of the conclusion "$\neg Q$: x+1 is a rational number" are true. We can claim by appealing to this fact that they are contradictory because the Warrant is a reasoning form of proof by contradiction. The Claim is derived *necessarily* and *certainly* from the Data of contradiction and the Warrant of proof by contradiction. However, it is important to consider that proof by contradiction is based on the logical theory of inference rules such as law of excluded middle, double negative elimination, and so on.

These are the Backing which lends authority to the Warrant of proof by contradiction and which establishes the acceptability of it. In view of intuitionistic logical theory, for example, these Backings are not accepted, so the Claim is refuted. This argument is valid unless it is not based on the classical logical theory, which is one of the Rebuttal. Generalizing this example, therefore, we can characterize mathematical proof by contradiction as follows in terms of argument (Figure 3).

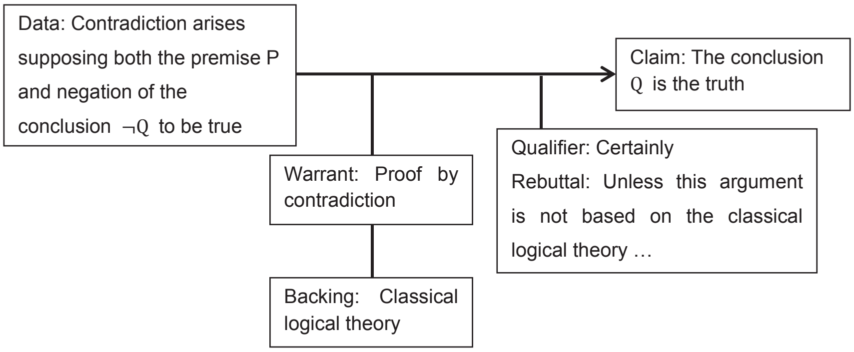

Figure 3. Argument of proof by contradiction

LOGIC OF STATISTICAL HYPOTHESIS TESTING

Statistical hypothesis testing is the technique that "aims to state the evidence in a sample against a previously defined (null) hypothesis, minimizing certain risks" (Castro Sotos et al., 2007, p.103). This technique is often used when the data of interest cannot be repeatedly observed many times. For example, if we obtain only one head in tossing the same coin ten times, we will surely test the coin and claim that the coin is an unfair one based on the observed data. At such a time, the claim that the coin is an unfair one cannot be directly established and supported even if we appeal to this observed data. Rather, the appeal can establish and support the opposite hypothesis that the coin is a fair one. Therefore, we attempt to establish and support the original hypothesis that we really want by posing another opposite hypothesis to contradict the hypothesis we want to accept and then rejecting it.

In hypothesis testing, two types of hypotheses are dealt with in this manner. One is the hypothesis we really want to support, which is referred to as the alternative hypothesis, and the other is the hypothesis we want to oppose and reject, which is referred to as the null hypothesis, in order to support the alternative hypothesis. We attempt to support the alternative hypothesis by means of tentatively supposing that the null hypothesis is true and then rejecting it based on observed data. To put it concretely, supposing the coin is fair, it is difficult for the (null) hypothesis to be supported by observed data, which is the fact that we obtain only one head in tossing the same coin ten times, with the result that based on observed data, the (null) hypothesis is rejected and another (alternative) hypothesis is accepted and supported. Supporting the alternative hypothesis by rejecting the null hypothesis based on observed data is the logic of hypothesis testing.

What we need think of here is a criterion for rejecting the null hypothesis. How likely is it that observed data are observed under the tentatively supposed null hypothesis? The standard for estimating this likelihood is defined in advance and is usually 0.05 or 0.01, the value of which is referred to as the significance α. Under the tentatively supposed null hypothesis, the fact that the total probability that observed data and the more extreme case are observed (p-value) is less than the level of significance α means that it is difficult for the hypothesis to be supported in terms of observed data. By this procedure, the null hypothesis is rejected, and the alternative hypothesis is accepted. In other words, the alternative hypothesis is accepted and supported taking into account at most the possibility of causing an error, and α is the probability of rejecting the null hypothesis by mistake, when it should actually be accepted and supported.

Based on observed data, hypothesis testing is conducted against a null hypothesis previously defined and supports an alternative hypothesis, hence it is a rule for decision making or a procedure for obtaining the support for the hypothesis, but its rule or procedure does not prove that a hypothesis is true mathematically or probabilistically (Castro Sotos et al., 2007; Falk & Greenbarm, 1995; Vallecillos, 1996). Using the example mentioned above, the null hypothesis is "the coin is fair," the alternative hypothesis is "the coin is unfair," the observed data are "obtaining only one head in tossing the same coin ten times," and the level of significance α is 0.05. The claim that we want to support here is "the coin is unfair." To establish this claim, supposing the null hypothesis that the coin is fair, we calculate the prior probability of obtaining only one head in tossing the same coin ten times and more extreme cases under this hypothesis, with the result that the total probability (p-value) is approximately 0.021, and hence it turns out that the value is less than the level of significance α. This fact makes it difficult to support the null hypothesis based on this observed data, and we must reject the null hypothesis and accept the alternative hypothesis with the probability of an error of approximately 0.021. In conclusion, this observed data establish and support the alternative hypothesis positively.

To compare it with mathematical proof by contradiction, I try to generalize this specific example. Although a hypothesis is not a statement, for comparison it is rewritten into the form of a statement: P is observed data, Q is the alternative hypothesis, and ¬Q is the null hypothesis. The logic of statistical hypothesis testing is as follows.

1. The fact that supposing ¬Q to be true contradicts P is used to support Q based on P.
2. However, P∧¬Q does not contradict in the strict sense of the word, hence the fact that the calculated value supposing both P and ¬Q to be true is less than the level of significance α is regarded as merely similar to contradiction.
3. When P∧¬Q is a state similar to contradiction, Q is supported based on P with an α error.

The Claim that we want to establish is the alternative hypothesis Q. This claim is supported by appealing to the observed data, but in more detail, the fact that the p-value calculated from the observed data under the null hypothesis is smaller than the significance level α. Note that the statistical data to support hypotheses are written as "observed data" in this paper to distinguish it from Data as an argument component in Toulmin's model. We can support the Claim by appealing to the fact that the p-value is smaller than the significance level α because the logic of statistical hypothesis testing is used as the Warrant. The Backing of this warrant is probability distribution, sampling distribution, or significance level, for example.

Arguments involving hypothesis testing are proper and formally valid reasoning as proof by contradiction, but the information of the Claim is not embedded in the Data. The Claim is not derived necessarily from the Data and the Warrant. The very possibility that the Claim is erroneous is entailed necessarily. In contrast to proof by contradiction, the step from the Data to the Claim is neither necessary nor certain, but only *tentative* and *provisional*. Articulating the Rebuttal is clearly, therefore, particularly essential in supporting and defending tentative arguments (Toulmin, 2003).

Showing the condition of the Rebuttal, which represents the case and the possibility that the Claim through hypothesis testing is an error, is not only required for defending an alternative hypothesis, but is also indispensable for arguments of hypothesis testing. Given that the supported hypothesis is based on a particular sample of observed data, other hypotheses might be supported, if other varied samples are used for testing hypotheses. In addition, there can be different hypotheses that are more valid, reasonable, and leading than the supported hypothesis for now. The Claim that is accepted and supported through the Warrant of hypothesis testing is always derived with the possibility of an error; hence it is defended by exhibiting the condition of Rebuttal. In statistical information using hypothesis testing, there must inevitably be the condition of Rebuttal, but it is often tacit and invisible. Therefore, it is important to keep in mind that even statistical information using hypothesis testing has the condition of Rebuttal and tentative and provisional claims.

Using Toulmin's argument (2003), this step is characterized as follows (Figure 4).

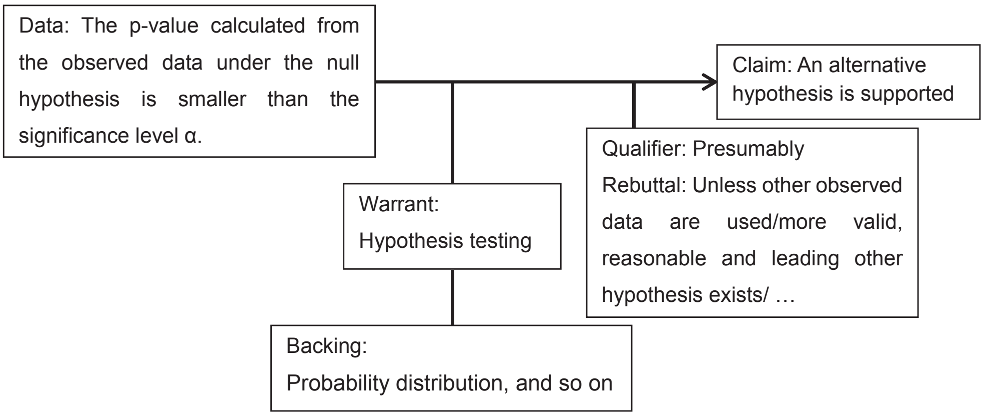

Figure 4. Argument of hypothesis testing

DISCCUSION

Mathematical proof by contradiction and statistical hypothesis testing are characterized as above, each using Toulmin's argument (2003) to organize and compare their structures. As a result, they are indeed found to have the same structure, argument-wise, but the contents of the elements are different. I discuss the differences in the following.

In Data, "supposing both P and ¬Q, contradiction arises" is used in mathematical proof by contradiction, and "supposing both P and ¬Q, the p-value is less than the level of significance α" is used in hypothesis testing. This provides the basis for establishing and supporting the claim that the contradiction (a state similar to contradiction) arises supposing both the premise P and the negation of the conclusion ¬Q. The following two differences, however, are recognized in both.

(1) Premise P in the case of proof by contradiction that is certain statement is *mathematical statement*, which is *invariable*, and therefore it is always supposed to be true. On the other hand, observed data P in the case of hypothesis testing that is certain sample is *non-mathematical statement*, which is *variable* owing to it being a sample, and nevertheless it is supposed to be true in the reasoning process, because it is supposed to represent the population correctly.

(2) $P \wedge \neg Q$ in the case of proof by contradiction is *complete and strict contradiction* within classical logical theory, while in the case of hypothesis testing it is not contradiction in the strict sense of the word, where the fact that the p-value is less than the level of significance α is regarded merely as *a state similar to contradiction*. The criterion that can regard it as contradiction is not invariable but depends on the sample size and the value of a significance level α, hence contradiction in hypothesis testing is distinct from that in proof by contradiction.

In hypothesis testing, it is supposed that observed data that comprise a particular sample are correct, and this is not invariable. Samples are nearly always different for every sampling. With the variability or uncertainty inherent in samples and sampling, a sample might not fully represent the characteristics of the population from which it is drawn. According to the analogy with proof by contradiction, however, observed data are considered to be representative and invariable. An analogical approach would contribute to the development of an incorrect perspective that statistical inference is deterministic, lacking a proper conception of sample that balances sample representativeness with variability (Rubin et al., 1991). Recognizing that observed data P in case of hypothesis testing is mostly *non-mathematical statement* seems to be essential to distinguish it from deterministic proof by contradiction.

There is also a difference in contradiction. The "contradiction" in hypothesis testing is not the same as the complete and strict contradiction in proof by contradiction, and is only regarded as a state similar to contradiction. Calculated based on observed data, the p-value depends on the sample. As the sample varies, the p-value can also vary. It is impossible to contradict completely. According to the analogy to proof by contradiction, nevertheless, when a p-value provided based on particular observed data is less than the level of significance α, this state can be regarded as deterministic, complete, and a strict contradiction. In addition to a desirable conception of sample, the recognition that *strict contradiction does not exist in hypothesis testing* and that there is merely a state similar to contradiction is indispensable to distinguish the two.

The following two differences are also recognized regarding the Claim.

(3) Conclusion Q in proof by contradiction is *necessarily and certainly* true to the extent that premise P is supposed to be true. On the other hand, in case of hypothesis testing, even if premise P, which describes observed data, is supposed to be true, conclusion Q is not always thereby supported and always remains *tentative and provisional*. The supported conclusion Q is always entailed with the possibility of an error. The Data always supports the Claim in proof by contradiction, while in hypothesis testing the Data only supports the Claim tentatively.

(4) Conclusion Q is derived conclusively in proof by contradiction; therefore, a Rebuttal, Qualifier and

Backing to support and defend the Claim are *unnecessary*. In contrast, conclusion Q in hypothesis testing is derived only tentatively, and therefore, *it is necessary to state them clearly and explicitly*.

The conclusion in proof by contradiction, which is a form of mathematical proof, is *necessarily* true and must not be *unnecessarily* defended. However, the conclusion in statistical hypothesis testing is *tentatively* true, and therefore, the claim to be supported must be *necessarily* defended. Seen from proof by contradiction, defending the claim is generally *redundant*, but seen from hypothesis testing, defending the claim by referring explicitly to a Rebuttal, Qualifier and Backing is *essential*. Unlike proof by contradiction, even supported hypotheses can involve an error. A conclusion derived according to a proper and formally valid reasoning process can also involve an error. These do not exist in mathematical proof using deductive reasoning. It is indispensable to recognize these characteristics of hypothesis testing in order to make it possible to recognize the tentativeness of claims and the need to defend arguments in hypothesis testing.

Comparing hypothesis testing with proof by contradiction is important in recognizing such differences concerning the Data and the Claim and then distinguishing them. The analogical approach with proof by contradiction, which is essential in teaching and learning hypothesis testing, is one to pay attention to their similarity; hence using that approach alone hypothesis testing has the risk of being assimilated into proof by contradiction. This appears to be a cause of the misconception concerning the nature of hypothesis testing (Castro Sotos et al., 2007; Falk & Greenbarm, 1995; Vallecillos, 1996). Their logical structures must be compared and differentiated after learning hypothesis testing by the analogical approach, in order to not assimilate it into proof by contradiction. In statistical hypothesis testing, unlike mathematical proof by contradiction, the supposed premise P (observed data) is not invariable and non-mathematical statement, contradiction from $P \wedge \neg Q$ in the strict sense of the word does not arise there, its Claim or the conclusion Q cannot always be supported and defended by observed data, and defending its conclusion Q with a Rebuttal, Qualifier and Backing is necessary. To understand the essence of statistical hypothesis testing, not only is it important to use the analogical approach with proof by contradiction, but also, more than that, it is necessary to compare hypothesis testing with proof by contradiction and characterize the former as not the same as the latter. In order not to teach and learn statistics as mathematics, and to develop statistical literacy well, efforts to highlight the differences of both tending to become tacit such as comparing intentionally their logical structures is necessary.

CONCLUSIONS

In this study, I shed light on four differences between the logic of statistical hypothesis testing and mathematical proof by contradiction by capturing the essence of them from the viewpoint of the elements of Toulmin's argument (2003). These differences are essential to recognizing hypothesis testing as such. The analogical approach with proof by contradiction, which is essential in teaching and learning hypothesis testing, lacks the necessary differentiation. It can be expected to follow that using the approach alone constructs the misconception concerning identifying them by mistake. In order not to teach and learn hypothesis testing as proof by contradiction, after the analogical approach it is necessary that both of them be compared and that the logic of hypothesis testing be characterized by contrast with proof by contradiction. Comparing and characterizing can contribute to proper understanding of both proof by contradiction and hypothesis testing.

The right and proper idea of hypothesis testing is necessary to become statistically literate in a society filled with data, statistics, and statistical information (Gal, 2004). In many countries worldwide, more often than not this subject is introduced in the last year of secondary education. However, currently in Japan, hypothesis testing is not dealt with within secondary education. Many citizens do not know about hypothesis testing or understand incorrectly it as mathematical proof by contradiction. Improving statistics education in Japan is an urgent problem in developing fully statistically literate people, and it is also a future task to concretely consider the suggestions offered in this study.

REFFERENCES

Antonini, S. & Mariotti, M. A. (2008). Indirect proof: What is specific to this way of proving? *ZDM Mathematics Education*, *40*(3), 401-412.

Batanero, C. (2000). Controversies around significance tests. *Mathematical Thinking and Learning*, *2*(1-2), 75-98.

Batanero, C. & Díaz, C. (2006). Methodological and didactical controversies around statistical inference. *Actes du 36iémes Journées de la Societé Française de Statistique*. Paris: Societé Françaisede Statistique.

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In Rossman, A. J. & Chance, B. L. (Eds.), *Proceedings of the Seventh International Conference on the Teaching of Statistics (ICOTS-7)*. Voorburg, The Netherlands: International Statistical Institute. Retrieved from https://iase-web.org/documents/papers/icots7/2D1_BENZ.pdf

Burrill, G., & Biehler, R. (2011). Fundamental statistical ideas in the school curriculum and in training teachers. In Batanero, C. et al. (Eds.), *Teaching statistics in school mathematics-Challenges for teaching and teacher education: A joint ICMI/IASE Study* (pp.57-69). Dordrecht: Springer.

Castro Sotos, A. E., Vanhoof, S., van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*, *2*, 98-113.

Cobb, G. W., & Moore, D. S. (1997). Mathematics, statistics, and teaching. *The American Mathematical Monthly*, *104*(9), 801-823.

Falk, R., & Greenbaum, C. W. (1995) Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology, 5*(1), 75-98.

Gal, I. (2004). Meanings, components, responsibilities. In Ben-Zvi, D., & Garfield, J. (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp.47-78). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Garfield, J. & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Springer Publishers.

Liu, Y., & Thompson, P. W. (2005). Teachers' understanding of hypothesis testing. In Wilson, S. (Ed.), *Proceedings of the Twenty-seventh Annual Meeting of the International Group for the Psychology of Mathematics Education*, Roanoke, VA. Vicksburg, VA: Virginia Tech. Retrieved from http://pat-thompson.net/PDFversions/2005PMENA%20HypTest.pdf

Pedemonte, P. (2007). How can the relationship between argumentation and proof be analysed?. *Educational Studies* *in Mathematics*, *66*, 23-41.

Rubin, A., Bruce, B. & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In Vere-Jones, D. (Ed.), *Proceedings of the Third International Conference on Teaching Statistics* (pp.314-319). Voorburg, The Netherlands: ISI Publications in Statistical Education.

Shaughnessy, M. (2007). Research on statistics learning and reasoning. In Lester, F. K.(Ed.), *Second handbook of research on mathematics teaching and learning* (vol.2, pp.957-1009). Charlotte, NC: Information Age Publisher.

Toulmin, S. (2003). *The use of argument*. Updated Edition. Cambridge: Cambridge University Press.

Vallecillos, A. (1996). Students' conceptions of the logic of hypothesis testing. *Hiroshima Journal of Mathematics Education*, *4*, 43-61.

Wild, C. J. & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223-265.

Hiroki Otani
Japan Women's University
1-1-1, Nishi-ikuta, Tama-ku, Kawasaki,
Kanagawa Prefecture, Japan
E-mail: otanih@fc.jwu.ac.jp