교육학박사학위논문

Assessment of Informal Statistical Inference

비형식적 통계적 추리의 평가

2015년 2월

서울대학교 대학원

수학교육과

박 민 선

Assessment of Informal Statistical Inference

비형식적 통계적 추리의 평가

지도교수 이 경 화

이 논문을 교육학박사 학위논문으로 제출함

2014 년 10 월

서울대학교 대학원

수학교육과

박 민 선

박민선의 교육학박사 학위논문을 인준함

2015 년 1 월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

위 원 _____ (인)

Assessment of Informal Statistical Inference

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Education
to the faculty of the Graduate School of
Seoul National University

by

Minsun Park

February 2015

Supervising Professor: Kyeong-Hwa Lee

Approved by Dissertation Committee:

Chair	Dr. Sang Kwon Chung	_____
Vice Chair	Dr. Jin Kon Hong	_____
Committee	Dr. Seokil Kwon	_____
Committee	Dr. Bomi Shin	_____
Committee	Dr. Kyeong-Hwa Lee	_____

ABSTRACT

Assessment of Informal Statistical Inference

Minsun Park
Department of Mathematics Education
The Graduate School
Seoul National University

In recent studies in statistics education, the teaching and learning of informal statistical inference have been emphasized as a precedence stage of formal statistical inference, which is taught at the tertiary level. Subsequently, there is an abundance of research identifying the meaning of informal statistical inference and exploring ways of improving students' informal statistical inference abilities. As research on how to instruct informal statistical inference grows, how to assess students' informal statistical inference abilities should be discussed. There has been a need for assessment methods that reflect and align with the characteristics of statistics, but substantive discussion about such assessments are still lacking. Thus, this study aims to clarify the nature of informal statistical inference and to propose appropriate assessment methods that reflect its nature.

Through an epistemological analysis of statistical inference, it is found that a statistical inference consists of two thinking components, abduction and induction. Statistical inference as induction can regulate its inherent characteristic according to

how it deals with the uncertainty. To address the dilemmas posed by uncertainty, statistical inference uses the quantification of uncertainty using probability and applies modus tollens. Statistical inference as abduction is introduced to denote the importance of generating the simplest and most likely explanation of a hypothesis based on the characteristics and patterns of the sample and the context. Both induction and abduction serve as components of thinking in regard to statistical inference and need to be recognized as separate stages.

Through a didactical review of research on informal statistical inference, the treatment of essential concepts and thinking in informal statistical inference were examined. The essential concepts include descriptive statistics as expectation and variation, sample and population, the size of a sample, and sampling distribution, and the essential thinking includes abduction and induction. In informal statistical inference, abduction and induction are carried out as construction of argumentation and verification of argumentation, respectively. In particular, to address the uncertainty in verification of argumentation, students can use probabilistic representations, draw a conclusion by recognizing the importance of repeated sampling, and attempt to validate the argumentation by establishing norms for dealing with uncertainty during the communication. The characteristics of informal statistical inference include that it is an informal argumentation using natural language, that it is based on context, and that it occurs within an interaction. Therefore, the realization of informal statistical inference demands a situation of teaching and learning processes in which communication occurs based on argumentation using verbal language.

Due to the nature of informal statistical inference, assessments must occur in parallel with the teaching and learning process. For this reason, the meaning of integration of instruction and assessment was examined and several assessment models, such as the general assessment triangle model and assessment models based on interaction, were analyzed. As a result, an assessment model for informal statistical inference was developed. The assessment model includes the integration of instruction and assessment as a universal set and interaction between a teacher and students as two intersecting sets. The procedure of assessment is represented in the intersection, which consists of a teacher's providing tasks, students' initial responses, a teacher's interpretation based on an assessment element, a teacher's feedback, and students' final responses. The characteristics of proper assessment tasks, the assessment elements, and the proper method for providing feedback for assessing informal statistical inference are described. Finally, the pedagogical implication of the study is discussed, and future research based on the assessment model developed in the study is suggested.

Keywords: informal statistical inference, assessment model, induction, abduction, argumentation, integration of instruction and assessment

Student Number: 2010-30404

Table of Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

Statistical inference is common in daily life. A lot of information can be observed from the media about results of surveys and experiments on behavior. During the election season, people can learn about the results of public opinion polls and exit polls almost every day. When this occurs, inferences are induced about the population. Gal (2004) focused on statistical inference as one of the aspects of statistical literacy, claiming that people should be able to understand and interpret the meaning of these reported findings or data-based claims. To do this, it is necessary to understand the logic of sampling, the need to infer from samples to populations, the notions of representativeness, and especially the notion of bias (Gal, 2004, p. 59).

Statistical inference is a process of finding characteristics of a population by extracting implied information from its samples, which are randomly drawn from the population (Lee, Lim, Seong, & So, 2000, p. 201). A main characteristic of statistical inference is to use probability for drawing more reliable conclusions (Kim et al., 2006, p. 195). Usually, statistical inference includes parameter estimation and hypothesis testing, both of which are taught in tertiary level. However, in recent studies in statistical education, researchers have extended the boundary of statistical inference with formal statistical inference and informal statistical inference (Watson, 2008; Zieffler, Garfield, delMas, & Reading, 2008). Formal statistical inference includes parameter estimation and hypothesis testing. Inferences about populations from

samples by using previous knowledge instead of following formal procedure is classified as informal statistical inference. In the case of informal statistical inference, students can recognize the variability of samples and relevance of samples and populations informally by making predictions about populations based on samples. Thus, they can subsequently approach formal statistical inference easily. Ben-Zvi and Gil (2010) described informal statistical inference as a bridge between descriptive statistics and formal statistical inference. In other words, informal statistical inference is significant in that it serves as a precedence stage of formal statistical inference.

The 2000 Principles and Standards for School Mathematics provided by National Council of Teachers of Mathematics (NCTM) and the Guidelines for Assessment and Instruction in Statistics (GAISE) Report provided by American Statistical Association (ASA) have presented the learning objective of informal statistical inference to primary and secondary students. NCTM (2000) states that all students from prekindergarten through grade 12 should be able to develop and evaluate inferences and predictions that are based on data. It delineates the specific learning goals that students should:

- Discuss events related to students' experiences as likely or unlikely.
- Propose and justify conclusions and predictions that are based on data and design studies to further investigate the conclusions or predictions.
- Use observations about differences between two or more samples to make conjectures about the populations from which the samples were taken.

- Use simulations to explore the variability of sample statistics from a known population and to construct sampling distributions.
- Understand how sample statistics reflect the values of population parameters and use sampling distributions as the basis for informal inference. (NCTM, 2000, pp. 400-401)

The GAISE Report, which set the guidelines for teaching statistics and assessment of K-12 students, recommends that students should learn about statistical inference by developing inferences based on data and inducing inference by using probability (ASA, 2005).

Even with the emphasis on the importance of informal statistical inference and the presentation of required learning objectives, the determination of content in school curricula and the method of teaching the objectives has been controversial in many countries (Watson, 2008). This is particularly true in South Korea, where, according to its mathematics curriculum, estimation should be taught shortly after learning the meaning of sample, population, and random sampling in high school (Ministry of Education and Science Technology, 2011). Woo (2000) stated that statistical inference had been only taught as a procedure and a tool in school mathematics, and more emphasis is actually placed on calculating probability in theory of probability distribution. The reason for the lack of emphasis on informal statistical inference can be attributed to the fact that statistics is classified as a sub-domain of mathematics in school, which focuses on the deductive over the inductive approach. In addition, it is

much more difficult to teach informal thinking than to teach concepts or formal procedure. However, unless students learn concepts such as sampling variability, which are the basis of formal statistical inference through informal statistical inference, students will face difficulties when encountering formal statistical inference. Therefore, it is necessary to think about how to teach informal statistical inference in school mathematics.

There are several research on informal statistical inference. In Korea, research is generally focused on analyzing the concepts, such as sample or probability distribution, examining students' understanding, and studying methods of teaching. Concepts related to informal statistical inference that have been studied include sample (Lee & Ji, 2005; Lee & Nam, 2005; Lee & Shin, 2011), sampling (Lee & Park, 2006; Ko & Lee, 2011), probability variable and probability distribution (Hwang & Yoon, 2011; Shin, 2012; Choi, Yun, & Hwang, 2014), and sampling distribution (Kim, 2006; Lee & Lee, 2010; Ko, 2012). Since each of these concepts has its own characteristics, researchers have tried to analyze each one didactically either by examining students' and teachers' understandings of the concept or by determining the most effective teaching methods. On the other hand, research on informal statistical inference outside Korea mostly targets primary and secondary students, focusing on defining and identifying components of informal statistical inference (Zieffler et al., 2008; Makar & Rubin, 2009; Makar, Bakker, & Ben-Zvi, 2011), examining students' understanding or development through teaching experiments, and determining methods of improvement (Ben-Zvi, 2006; Pfannkuch, 2007; Watson,

2008). Thus, many studies have emphasized the importance of informal statistical inference with a concentration on teaching students to understand the concepts.

With its importance in instructing statistical inference, assessment of instructed knowledge should also be discussed. As research on how to instruct informal statistical inference grows, understanding the process of measuring students' understanding is required. As previously stated, research on assessment for examining the relevance in understanding and thinking of concepts is further necessitated by the fact that informal statistical inference holds significant as a precedence stage of formal statistical inference.

Studies on assessment of statistics are ongoing include assessment of statistical concepts, statistical reasoning, and statistical thinking. These have been conducted by several research projects, such as Statistics Concepts Inventory (SCI; Allen, Stone, Reed-Rhoads, & Murphy, 2004), and Assessment Resource Tools for Improving Statistical Thinking (ARTIST; Garfield, delMas, & Chance, 2002). However, there is a need for more research on assessment of informal statistical inference. Recently, Park (2012) and Holcomb, Chance, Rossman, and Cobb (2010) studied the development of assessment tools for statistical inference. The target for both studies were tertiary level students, and the goal was to establish multiple-choice questions that would allow for the quantitative assessment of students' levels of understanding. However, the assessment suffers from limitations because it targets only tertiary level students and uses only multiple-choice questions, which is not verified yet as a proper way to assess informal statistical inference. Thus, it is necessary to find the most

fitting measures for assessing primary and secondary students' abilities on informal statistical inference.

Gal and Garfield (1997) underscored the need for new assessment methods that take into account both the aim of statistics education and the difference between statistics and mathematics. However, this distinction proved difficult due to the fact that curriculums typically classified statistics as a sub-category of mathematics in school and as such assessed similarly, i.e. the deductive, unique-answer. Although this deductive, unique-answer approach is the most convenient for assessing students, it is highly problematic considering the characteristics of statistics. First of all, statistics is a methodological discipline. Thus, the ability to apply its concepts is more important than demonstrating concepts' definitions. For example, in the case of the concept of an "average," it is important to be able to determine which average is appropriate among "mean," "median," or "mode" by considering characteristics of the data rather than knowing the definition of the term "average" or the procedure of its calculation (Watson, 2006). Secondly, context is crucial in statistics. If formal statistical inference is applied to determine the significance of an experiment, a key step lies in one's decision on where to set the significant level, 0.1 or 0.01, depending on the potential influence of the experiment to the people. If the result of an experiment is dangerous, one can perform the statistical inference with a smaller significance level. Lastly, inherent in statistics is the characteristic of uncertainty, for statistics addresses distribution, not ideal true value (Stigler, 1986). For these reasons, students' abilities to apply statistics cannot be appropriately measured using deductive, unique-answer

assessments.

Statistical inference uses induction, which is extended from a result of a special occasion or condition to guide to a principle of general situation based on characteristics of statistics and indeterminism epistemology (Kim, 2001, p. 173). To assess students' informal statistical inference, it is necessary to employ assessment methods that reflect the nature of informal statistical inference and present opportunities to explore it. Thus, the discussion to identify the nature of informal statistical inference how to assess it has become imperative. Consequently, assessment situations, environments, tasks and the role of the teacher in assessment should be examined.

The purpose of this study is to clarify the nature of informal statistical inference and to propose a fitting assessment method that reflects its nature. Because statistical inference depends on induction, which then results in conclusions about populations from samples, it can be described as a type of thinking process. To clarify the nature of informal statistical inference, it is necessary to examine the difference between the thinking process involved in statistical inference and other thinking processes. Thus, an epistemological analysis of general statistical inference is required to reveal its essence. It is then necessary to examine the didactical analysis of informal statistical inference. As previously mentioned, informal statistical inference is a precedence stage of formal statistical inference, such as estimation and hypothesis testing, and has the educational goal of supporting the understanding of formal statistical inference. Therefore, identification of the inherent characteristics of informal statistical

inference as verified in previous studies is needed, as is the examination of the treatment of essential concepts and thinking in statistical inference. It will also be required to reflect on the method for designing assessment. By integrating the results of the epistemological analysis of statistical inference and the didactical analysis of informal statistical inference, the nature of informal statistical inference and considerations that should be addressed in the assessment of informal statistical inference are induced. These results combined with the assessment research analysis will explicate the assessment model of informal statistical inference. The research questions of this study are as follows:

1. What is a nature of informal statistical inference?
2. What is an appropriate assessment method of informal statistical inference?

This dissertation consists of five chapters. Chapter 1 provides an overview of the research, drawing on the need for an assessment method to assess informal statistical inference. Chapter 2 describes the research methods used in this study. Chapter 3 reviews previous research of statistical inference and informal statistical inference and then deduces both the epistemological analysis of statistical inference and the didactical analysis of informal statistical inference. It answers research question 1 by presenting the nature of informal statistical inference. Chapter 4 answers research question 2 by outlining the appropriate assessment model for assessing informal statistical inference based on the nature of informal statistical inference and the

previous research analysis. Finally, chapter 5 provides a summary of the research findings and a discussion of the pedagogical implications of the results.

CHAPTER 2

METHODS

The purpose of this study is to clarify the nature of informal statistical inference and to suggest an assessment method, which reflects the nature of it. For this purpose, two research questions are presented. To answer both questions, a theoretical discussion based on previous research analysis is required. Thus, this study follows the document research method (Woo et al., 2006). Based on the document analysis of informal statistical inference, the nature of informal statistical inference is induced. By combining these findings and the assessment research analysis, the assessment method of informal statistical inference is drawn. Figure 1 illustrates the overall flow of the study.

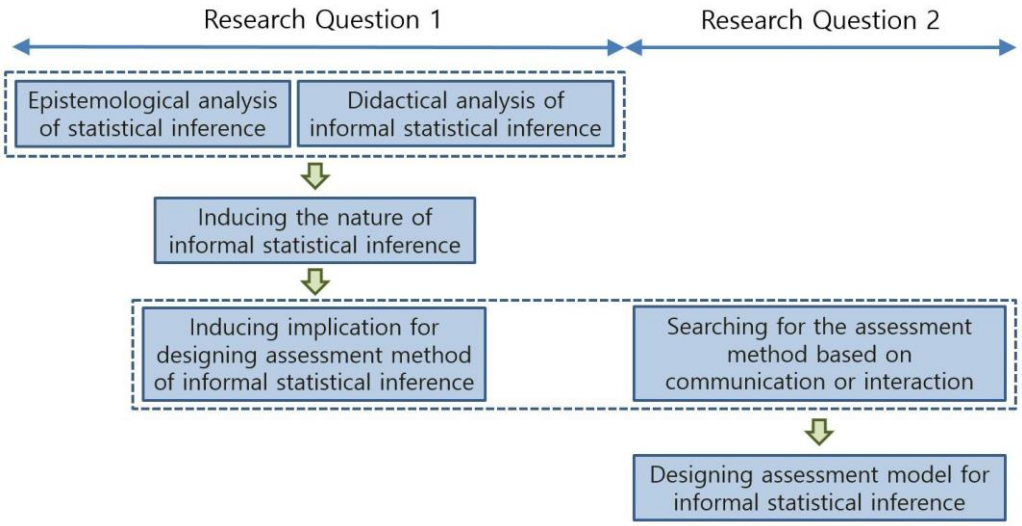


Figure 1 The overall flow of the study

To clarify the nature of informal statistical inference, didactical analysis is implemented. Didactical analysis is a type of methodology that analyzes the essence of a topic from multiple perspectives to identify useful pedagogical implications for organizing materials and lessons appropriate for teaching the topic of school mathematics (Woo et al., 2006, p. 31). Because this study aims to design an assessment method of informal statistical inference and because it is necessary to discuss informal statistical inference in terms of its multiple dimensions in order to induce the implications surrounding the most appropriate designing assessment method, didactical analysis proves fitting. First, previous research on statistical inference is analyzed epistemologically. Epistemological analysis is performed by inducing reasoning components included in statistical inference first by discriminating characteristics of descriptive statistics and inferential statistics and later by carefully analyzing each reasoning component's regard of characteristics of statistical inference. Next, previous research on instruction of informal statistical inference is analyzed. The examination of how the essential concepts and thinking in statistical inference are treated in informal statistical inference will reveal the inherent characteristics of informal statistical inference. By integrating the results of the epistemological analysis of statistical inference and the didactical analysis of informal statistical inference, implications that should be addressed in designing assessment method of informal statistical inference are indicated.

To design an assessment method of informal statistical inference, general educational research and mathematics education research related to assessment are

analyzed. Based on research that focuses on integrating instruction and assessment, the meaning of integrating instruction and assessment and the model for assessment in the situation of integrated instruction and assessment are examined. From the model, the focus is determining which components of the assessment model are important. By combining the results, the assessment model of informal statistical inference is generated.

CHAPTER 3

THE NATURE OF INFORMAL STATISTICAL INFERENCE

Statistical inference is an integral thinking process that induces conclusions about populations from samples. In statistical inference, a series of concepts, such as population and sample, parameter and statistics, sampling distribution, confidence interval, significant level, and null and alternative hypotheses are embedded (Batanero, 2000). The meaning of statistical inference has evolved over time, and there have been various philosophical issues on probability involved in statistical inference (Kalinowski, Fidler, & Cumming, 2008; Gigerenzer, Krauss, & Vitouch, 2004; Haller & Krauss, 2002). Some find statistical inference difficult and confusing due to the myriad aspects of issues one must consider.

To clarify the nature of informal statistical inference, it is necessary to examine the difference between the thinking process of statistical inference and other thinking processes because statistical inference is a thinking process. Thus section 3.1 presents the epistemological analysis of general statistical before the examination of informal statistical inference. Then, in section 3.2, the didactical analysis of informal statistical inference is presented to examine how informal statistical inference treats the essential concepts and thinking in statistical inference as it is necessary to reveal the characteristics inherent to informal statistical inference. Finally, in section 3.3, by integrating the results of the epistemological analysis of statistical inference and the didactical analysis of informal statistical inference, the nature of informal statistical

inference is summarized and implications that should be taken into account in the assessment of informal statistical inference are induced.

3.1. Epistemological Analysis of Statistical Inference

Indeterminism is a foundational epistemology of statistics. The key point of indeterminism lies in finding a way to deal with data while recognizing its uncertainty, variability, and chance. Statistical inference, with its basis of indeterminism, can be described as a thinking process to draw a conclusion regarding a population from a sample. For the epistemological analysis, it is necessary to draw essential thinking components involved in statistical inference. These components can be derived from the characteristics of inferential statistics that are distinct from descriptive statistics. Once these thinking components are determined, each can be thoughtfully analyzed in relation to the characteristics of statistical inference.

3.1.1. Thinking Components of Statistical Inference

Statistics is a methodological discipline which is appeared by an omnipresence of variability (Cobb & Moore, 1997). The two main categories of statistics, descriptive and inferential, are regarded as separate with each following its own methodology for managing variability of data (Rao, 1997, p. 63). Descriptive statistics relates to a summary of data by describing a given data. It includes calculating mean or standard

deviation for a given data and representing data visually using various graphs, such as bar graphs and histograms, to indicate the characteristics of data. In terms of variability, the variability in descriptive statistics involves the characteristics of the data itself, i.e., the distribution of data. On the other hand, inferential statistics aims to arrive at conclusions regarding a population based on information of a given sample. Variability in inferential statistics can be ascribed in two forms, the variability of the sample as drawn from the population and the sampling variability as occurred from a repetition of sampling. Rao (1997) named the methodology involved with the variability in descriptive data analysis and inferential data analysis as descriptive and inferential statistics respectively.

Unlike descriptive statistics, which employs a distribution to take into account variability for a given data set, inferential statistics addresses both the variability of a sample and sampling variability. As variability of sample is caused by the fact that a sample is drawn from a population, it is necessary to take this into account to derive a conclusion about the population from a given sample. Then, it is essential to have a process for examining the probability and the conclusion by asking, "What would happen if we used this method very many times?", in consideration of sampling variability by repetition of sampling from the same population (Cobb & Moore, 1997, p. 821). In other words, statistical inference, the fundamental of inferential statistics, can largely be divided into two components: the process of drawing a hypothesis regarding a population from a given sample with consideration of variability of the sample and the process of justifying a conclusion by calculating a probability

regarding a selected sample based on a sampling distribution from repetitive sampling. Among the two, the former relates to abduction while the latter relates to induction. Thus, in this study, each of the two main components of thinking used in statistical inference, abduction and induction, must be explained in more detail.

3.1.2. Statistical Inference as Induction

Statistical inference, a type of inductive inference, addresses the process of understanding an unknown population from a known sample. Inductive inference has been referenced as a method of scientific thinking (Gower, 1997; Salmon, 1967). Its main characteristic is that it contains a logical leap or uncertainty throughout its process of drawing a conclusion from premises. In this section, distinctive epistemological characteristics of an inductive inference in statistical inference will be discussed. In particular, the method for overcoming uncertainty over the course of the inference process will be highlighted. To start, it is important to understand the meaning of inductive inference and identify some attempts to solve the issue of uncertainty in statistical inference.

3.1.2.1. Meaning of inductive inference

Inductive inference is primarily known to have been introduced by Aristotle. Induction is a form of reasoning to understand a universal from a sense perception of particulars (Kwon, 1996). Induction is based on the fact that one can learn something

about the universe as a whole from a particular experience. The focus of induction is on the methodology or logic (Jeon, 2000). The methodology of induction, supported by Bacon (1561~1626), Whewell (1794~1866), Herschel (1792~1871), and Mill (1806~1873), concerns how the system derives a fitting conclusion from its given premises. On the other hand, the logic of induction, which is supported by Hume (1711~1776) and Carnap (1891~1970), concerns the validity of the inference that is drawn by a predetermined method.

For an inductive inference, one looks at whether the antecedent, $C_1, C_2, \cdots C_n$, is satisfied or not and then observes whether the consequent E occurred successively. From here, one finds a relativeness as '$C_1, C_2, \cdots C_n \rightarrow E$' to draw a universal rule (If C_i, then E) (Lee, 1988, p. 85).

- Observation: $C_1, C_2, \cdots C_n \rightarrow E$
- Universal rule: $L\ (C_i \rightarrow E)$

Inductive inference is inherently questionable because it draws broader, generalized conclusions from one or several relatively limited experiences. It is practically impossible to obtain a complete inductive inference, as it would require observing all cases for an infinite period of time and space (Lee, 1988).

Inductive inference can be viewed as an argument. An argument consists of more than two propositions where there is a relationship of "follow from" among different propositions. From an example of two propositions, p_1 and p_2, if p_2 is said to be

followed from p_1, the statement of p_2 is related to the statement of p_1. In such case, there are two ways for these propositions to relate to each other (Jeon, 2013):

- statement of p_2 is fully included in statement of p_1
- statement of p_2 is partially included in statement of p_1

The first case is called a deductive argument while the second is called an inductive argument. In other words, an inductive inference is a logical argument in which the content of the conclusion is partially included in the content of premises (Jeon, 2013, p. 54). Thus, it is possible for an inductive argument to contain statistical or probabilistic propositions in the premises or conclusion. According to Jeon (2013), propositions drawn from a given population and from its samples are applicable to this type of argument. It can be divided into three types: direct inference, drawing a conclusion for propositions of samples with premises from propositions of the population; reverse inference, drawing a conclusion for propositions of the population with premises from propositions of samples; and predictive inference, drawing a conclusion for propositions of samples with premises from propositions of different samples within the same population (Jeon, 2013, pp. 62-63).

Jeon (2013) has proposed three different meanings of induction (pp. 47-48). The first meaning incorporated the drawing of a conclusion beyond the premises based on empirical examples. For this, he used the term "inductive hypothesis construction." The second meaning of induction advocated the providing of appropriate evidence or

propositions to confirm a hypothesis or to provide an extent of confirmation that the evidence or propositions show. For this, he used the term "inductive hypothesis confirmation." The third meaning of induction involved the acceptance or rejection of a hypothesis by providing certain criteria to confirm formerly confirmed cases or any premises, which he named "inductive hypothesis acceptance/rejection." To explain these three meanings in terms of argument, inductive hypothesis construction provides IC (inductive conclusion) given IP (inductive premises) whereas inductive hypothesis confirmation and inductive hypothesis acceptance determine the extent to which a provided IP can be demonstrated as strong evidence for IC and whether IC can be considered as true or false in respect to IP.

The first key characteristic of an inductive inference is non-monotonicity (Jeon, 2013, p. 65). Consider a case where concerning the calculation of the probability of a coin landing on "tails" after being tossing a thirteenth time, with its outcome of the first 12 trials being 10 "heads" and two "tails." From deductive inference, which is based on an axiomatic system of probability, the probability of the thirteenth coin toss again will be a half, as previous attempts do not have any effect on the conclusion. On the other hand, in case of inductive inference, new premises will be added to any new events that could result in a change in conclusion. Consequently, the probability of the thirteenth coin toss will be less than a half as more "heads" have been drawn in the earlier attempts. This characteristic is called non-monotonicity. The second characteristic to consider is that an inductive inference uses all data. While for a deductive inference, it is permissible to choose a subset of premises to prove a

proposition, for an inductive inference, different subsets of data may lead to different and often contradictory conclusions (Rao, 1997, p. 56). Thus, all data should be used, and the editing or deleting of data should be performed by the inference process only if necessary. Finally, in inductive inference, the given data are the only elements that can be used. Any unverified assumptions or preconceived notions are forbidden to be used inputs (Rao, 1997).

3.1.2.2. Uncertainty in inductive inference

As discussed in the previous section, inductive inference is a process that aims to infer generalized rules for all the cases, including the ones yet to be experienced, from cases that have already been experienced. In addition, a conclusion drawn from an inductive argument is partially included by its premises in comparison to a deductive argument where a conclusion is fully included by its premises. By the fact that a generalized rule is derived from relatively limited experiences and that conclusions are only partially included by its premises, inductive inference is inherently uncertain. The necessary discussion on the uncertainty in inductive inference should be focused on the methods to overcome such uncertainty, methods that correspond to Hume's discussion of the justification of an inductive inference. Hence, one can find the unique characteristics of statistical inference as induction, which is a methodology of justification of an inductive inference addressing the uncertainty.

This study, observing the epistemological characteristics of statistical inference, aims to highlight the attempts to deal with the uncertainty of inductive inference in

statistical inference. The first attempt employs probability to quantify the uncertainty. Inferring a generalized statement with limited experiences that supports propositions is uncertain and prone to error. However, by using its probability, one can find the probabilistic truth of a generalized statement of an inductive inference (Lee, 1988). Similar to the existences of different philosophical perspectives in probability (Lee, 1996), there were various school of statistics based on which probability concept is being used to quantify the uncertainty of statistical inference. The second is to introduce *modus tollens* in statistical inference. *Modus tollens* is a form of logic in which one rejects the reverse of an original hypothesis by providing counterevidence of the reverse of the original hypothesis to overcome the limitation of generalization due to limited samples. In the following sections, two methods to surmount the uncertainty of statistical inference will closely be reviewed.

3.1.2.2.1. Quantification of uncertainty

Rao (1997) stated that the new paradigm, which is expressed in the following logical equation, was presented in the beginning of the 20th century. It was realized that although the knowledge created by any rule of generalizing from the particular is uncertain, it becomes certain knowledge, although of a different kind, once one can quantify its amount of uncertainty (Rao, 1997, p. 54).

Uncertain knowledge	+	Knowledge of the extent of uncertainty in it	=	Useable knowledge

Rao proposed that this new paradigm brought with it a new way of thinking based on the following assumptions:

- If we have to take a decision under uncertainty, mistakes cannot be avoided.
- If mistakes cannot be avoided, we better know how often we make mistakes (knowledge of the amount of uncertainty) by following a particular rule of decision-making (creation of new but uncertain knowledge).
- Such a knowledge could be put to use in finding a rule of decision making which does not betray us too often or which minimizes the frequency of wrong decisions, or which minimizes the loss due to wrong decisions. (Rao, 1997, pp. 54-55)

Like these, the quantification of uncertainty is the primary concept that has led to the codification of inductive inference (Rao, 1997). Inductive inference induces knowledge with uncertainty, but by quantifying the uncertainty, it induces probabilistic truth and useable knowledge. For example, consider forecasting tomorrow's weather. To do so requires the examination of the frequency of occasions on which it rained in the past on tomorrow's date when the atmospheric conditions on the previous day were as observed today (Rao, 1997). If one draws a conclusion that "there is 30% chance of rain tomorrow" from the information about frequency, then the number 30% is the amount of uncertainty about whether it will rain or not tomorrow, so the conclusion represents a statement of probabilistic truth. Eventually,

it becomes important to determine how to quantify the uncertainty of an inductive inference using a probability as one of the answers.

Various philosophical perspectives present themselves in regard to probability. Lee (1996) considered this probability from four different perspectives; classical, logical, subjective, and frequentist. Statistical inference can quantify uncertainty in various ways depending on the perspectives of probability. Through the application of a study conducted by Jeon (2013), the following section reviews Bayesianism, frequentism, and the likelihood approach in interpreting probability and quantifying uncertainty. Through this review, one can identify each perspective's attempts to solve dilemmas posed by the subjective nature.

In Bayesianism, a degree of support for a hypothesis can be derived from the posterior probability in Bayesian theorem (Jeon, 2013). Bayesian theorem is as follows:

$$P(h_r|e) = P(h_r) \times P(e|h_r) \,/\, P(e)$$

Posterior probability is computed with three different probabilities (Jeon, 2013) that are prior probability of hypothesis $P(h_r)$, likelihood of hypothesis $P(e|h_r)$, and probability of evidence $P(e)$. The prior probability of hypothesis, $P(h_r)$ is referring to the subjective probability of hypothesis that reflects a degree of a personal belief in the likelihood of the occurrence. In other words, the degree of inductive support is perceived as a degree of belief, which means the prior probability of hypothesis can

be freely defined as subjective. Bayesian theorem's second probability addresses likelihood, $P(e|h_r)$. Likelihood can be applied in the events of hypothesis h_r becoming true based on evidence e of samples drawn from a population of individuals. The probability becomes 1 in the case of a hypothesis being deterministic with hypothesis h_r logically implying the evidence e whereas in the case of a hypothesis being non-deterministic, the frequency of individuals that are satisfying the hypothesis becomes important as not all objects from the population supported by the hypothesis will satisfy it. Lastly, probability of evidence $P(e)$ can be obtained by expanding the hypothesis that is believed to be true, h_r, and the inverse, $\sim h_r$.

Frequentism was introduced as a critique of the subjectivity involved in the prior probability of hypothesis in Bayesianism. As inductive inference is inherently uncertain, it is easy for the reasoner to become biased. Frequentism is based on the belief that it is necessary to remain objective even in the case of uncertainty and this objectivity can be achieved through an emphasis on the frequency of data. With frequency of data, there can exist no possibility of inconsistency between reasoners, which precludes subjectivity. It is for this reason that they are called frequentists (Jeon, 2013).

Two leading methods used in frequentism to deal with relative frequency include Fisher's significance test and Neyman-Pearson's hypothesis testing. Accordingly to Fisher's significance test, frequentists first find the probability of relative frequency and define the criteria of probability to be considered statistically significant. It is possible to set these criteria at 0.05 (5%) or 0.01 (1%) as a commitment without any

theoretical evidence, and such probability is called significance level. However, such significant testing is not without its limitations, for it is possible to obtain the same outcomes from different experiments by null hypothesis based on which test statistics one uses (Jeon, 2013). Neyman-Pearson hypothesis testing takes these errors into account. Unlike the significance test in which the hypothesis is being validated by evidence from past occurrences on a predefined significance level, hypothesis testing works to minimize both the error of construing a wrong hypothesis from rejecting the correct hypothesis as well as the error of accepting a false hypothesis. Frequentists use p-value to describe the degree of inductive support for evidence that contradicts the problematic null hypothesis.

The likelihood approach was introduced to solve the limitations of frequentism. Subjectivity resides hidden in frequentism where significance level or probability of first order of error α is established in advance. In addition, a reasoner's subjectivity comes into play when calculating p-value where it can be changed with stopping criteria that can halt the trial that can create evidence such as observation and experiment. The key aspect of the likelihood approach is thorough objectivity regarding the experimental evidence (Jeon, 2013). Likelihood is determined by how objectively observed outcomes can make evidence either stronger or weaker in relation to a hypothesis. Likelihood involves the probability that a state's being will change, and it also can be perceived as conditional probability $P(e|h_r)$ in which evidence e is based on hypothesis h_r. Supporters of the likelihood approach implement an intuitive principle, a law and principle of likelihood that is based on

evaluating a hypothesis with likelihood. They are important in that the degree of support for a hypothesis based on evidence can be provided and compared to likelihood (Jeon, 2013). As the likelihood approach compares how more than two hypotheses are supported by each one's evidence, the degree of support changes, depending on the different hypotheses being compared. Therefore, it is important to understand the impact of the other hypothesis that will be compared with hypothesis of interest.

Bayesianism, frequentism, and the likelihood approach are perspectives that appeared to address uncertainty. However, they are different in each one's extent to which it accepts subjectivity, what approach each takes in regard to probability to solve the issue, and how each quantifies any uncertainty. It can be summarized that Bayesianism places emphasis on a prior probability that is subjective, frequentism aims to secure objectivity by recognizing probability as a limitation to relative frequency, and the likelihood approach works to compare the degree of support of hypotheses by observed evidence considered objective. With such different approaches to subjectivity, one cannot choose one over the others for all cases, but it remains essential to apply the appropriate approach given the particularities of a statistical inference situation. In today's world of statistics, the majority of statistical inference draws on the perspective of frequentism, meaning it maintains a strict objectivity based on relative frequency even with the possibility of the subjective setting of the significance level. However, at times, one may draw a more fitting inference by applying the Bayesianist approach. For instance, consider the following

case regarding HIV screening.

> In a population, the a priori probability of a person being infected by HIV $P(H_1)$ is 0.0001 while the probability of a test being positive for an infected person $P(D|H_1)$ and a non-infected person $P(D|H_2)$ is 0.999 and 0.0001 respectively. In this case, what is the probability that a person with a positive HIV test actually has the HIV virus? (Gigerenzer et al., 2004, p. 7)

It can be calculated using Baye's rule:

$$P(H_1|D) = \frac{P(H_1)P(D|H_1)}{P(H_1)P(D|H_1) + P(H_2)P(D|H_2)} = \frac{0.0001 \times 0.999}{0.0001 \times 0.999 + 0.999 \times 0.0001} = 0.5$$

Approaching the same problem using frequentism, the null hypothesis is defined as a person not infected by HIV and the probability of a person testing positive in the non-infected group is 0.0001, which is same as the level of significance. Therefore, it is considered highly unlikely thus rejecting the null hypothesis, which will result in accepting the alternative hypothesis that the person indeed has the HIV virus. This is somewhat different from the result of Bayesianism where the probability of that person being infected by HIV was 0.5. As demonstrated from this example, Bayesianism can sometimes be more appropriate in drawing a reliable conclusion by calculating the probability based on the priori probability for two hypotheses (Gigerenzer et al., 2004).

In summary, statistical inference as inductive inference introduces probability to quantify uncertainty occurring in the process of inference in attempt to give greater understanding. That is, the quantification of uncertainty is of key importance in the generalization of a statistical inference. Bayesianism, frequentism, and the likelihood approach each propose different approaches in regard to the quantification of uncertainty using probability. As essential information will differ depending on the context, it is crucial to select the proper approach for a given statistical situation rather than assuming and then accepting one as an absolute truth.

3.1.2.2.2. Introduction to *modus tollens*

The second attempt to overcome the uncertainty of inductive inference in statistical inference leads one to *modus tollens*. *Modus tollens* is a type of hypothetical syllogism that involves inferring a negation of an antecedent by denying the consequent (Lee, 1988). In other words, in "If a, then b," it draws a conclusion of "not b, thus not a." *Modus tollens* is a deductive logic fundamentally.

Fisher first introduced it in the history of statistics. Salsburg (2001) quoted Fisher to explain his interpretation of significance, true hypothesis, and p-values.

> For the logical fallacy of believing that a hypothesis has been proved to be true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistical than in other kinds of scientific reasoning ... It would, therefore, add greatly to the clarity with which the tests of significance are

> regarded if it were generally understood that tests of significance, when used accurately, are capable of rejecting or invalidating hypotheses, in so far as they are contradicted by the data: but that they are never capable of establishing them as certainly true ... (Salsburg, 2001, pp. 107-108)

The chi square goodness of fit test developed by Pearson, who was before Fisher, is valuable as a way of statistical inference, for it offers a trial for "proving" that data follow a particular distribution (Salsburg, 2001). Statistical inference involves the process of validating evidence to accept a hypothesis, but by the fact that evidence can be limited and imperfect, it does not provide absolute certainty but rather provides the state of a hypothesis being highly probable (Lee, 1988). Therefore, a hypothesis naturally follows the characteristic of a probabilistic hypothesis. Fisher tried to solve the problem of uncertainty in the process by *modus tollens*. Instead of validating numbers of evidence, he tried to draw a logical conclusion of the hypothesis being true by proposing a hypothesis rejected through falsification using *modus tollens*. Not only was Fisher among the first to apply the *modus tollens*, but he also he proceeded one step beyond Pearson's discovery in dealing with uncertainty in the process of inductive inference.

The introduction of *modus tollens* is related to a movement of induction that shifted from "induction by enumeration" to "induction by elimination." Induction studied by Keynes, Jeffereys, and Carmap can be characterized by enumeration, meaning the process of drawing a conclusion by spelling out evidence where its focus

is to find supportive evidence (Jeon, 2013). However, due to the fact that the statement of the conclusion is partially covered in the premises, induction by enumeration is exposed to the danger of the evidence and conclusion standing together even when the evidence may contain no logical implication of the conclusion. In contrast, induction by elimination not only considers the conclusion of interest but also considers other conclusions as alternate hypotheses and eliminates those (Jeon, 2013). As *modus tollens* is a reasoning process to determine a proposition that denies the reverse of the original hypothesis and reject it, it is widely used in induction by elimination. Bayesianism, frequentism, and the likelihood approach also provide perspectives that were developed using induction by elimination.

The usage of *modus tollens* in statistical inference is as follows: "If A implies B is very unlikely to happen. Then, if B happens, we deduce A is very unlikely to be true (Batanero, 2000, p. 86)." In statistical inference, evidence is collected from empirical observations and the data is drawn from samples rather than from an entire population. Therefore, it contains probabilistic propositions such as "highly unlikely" and does not necessarily derive a valid conclusion. Although an expectation can be more than what is drawn from a conclusion, as per Hume, it can be associated with the justification of an inductive inference and should be approached with more caution (Batanero, 2000). Nevertheless, introduction of *modus tollens* in statistical inference maintains significance in that it draws a probabilistic hypothesis and tries to overcome the limitation of induction by enumeration.

It is important to understand how *modus tollens* is used in hypotheses testing (Lee,

1988, pp. 89-90). Consider the following proposition, "With probabilistic hypothesis rule L_i, it is highly likely for E to happen observing the antecedent events of $C_1, \cdots C_n$" and the inferred proposition, "If probabilistic hypothesis rule L_i is true, the probability that event E will occur is p given the antecedent events of $C_1, \cdots C_n$." To test the hypothesis, whether rule L_i is true or not, one can apply *modus tollens*. Consider the following proposition for the *modus tollens*, "If probabilistic hypothesis rule L_i is true and the antecedent events of $C_1, \cdots C_n$ are given, then the probability that event E will occur is p." By providing the observation of the less likely occurrence of p, "the probability of event E will occur," one can prove that rule L_i is false. However, given the fact that the hypothesis that is being tested can only be falsified and cannot be proved according to Popper, the null hypothesis, which is the negation of original hypothesis, "Probabilistic hypothesis rule L_i is not true" must be introduced and the proposition changes to "If probabilistic hypothesis rule L_i is not true and the antecedent events of $C_1, \cdots C_n$ are given, then the probability that event E will occur is q." At this time, if the actual observed probability of event E occurring q' is not the same as q, one can infer that it is not likely that the conclusion of the null hypothesis, "probability that event E will occur is q" is true. Therefore, the "Probabilistic hypothesis rule L_i is not true" can also be inferred as not likely and thus, be rejected, which leads to the conclusion that the original hypothesis which is contradictory to the latter can accepted.

In statistical inference, the word "significant" is frequently used in relation to *modus tollens*. Salsburg (2001) charged that early in the development of statistical

inference idea, the word “significant” came to be used to indicate that a probability was low enough for rejection. That is, data became significant if they could be used to reject a proposed distribution (Salsburg, 2001). The word was used in its late 19th century English meaning, which is simply that a computation signified or showed something. As the English language entered the 20th century, the word “significant” began to take on other meanings until it developed its current meaning, implying that something very important. Statistical analysis continues to employ the word “significant” to indicate a very low probability computed under the hypothesis being tested. In that context, the word has an exact mathematical meaning. However, those who use statistical analysis often treat a “significant” test statistic as implying something much closer to the modern meaning of the word (Salsburg, 2001).

Thus, *modus tollens* is introduced to solve the problem of uncertainty by using justification in statistical inference and provided the background of the frequently use of the word “significant.” Statistical inference heavily relies on the *modus tollens* due to its close relation to rejecting hypotheses and the lack of the ability to justify a hypothesis in inductive inference even with much evidence. In summary, *modus tollens* maintains importance in statistical inference with its purpose in solving and justifying the occurrence of uncertainty.

3.1.3. Statistical Inference as Abduction

The process of generating a hypothesis of population drawn from data is included

in statistical inference. Among the three types of inference defined by Peirce, deduction, induction, and abduction, abduction is the one related to generating a hypothesis. While the discussion on inductive inference focused on how to deal with uncertainty in the process of inference, abduction is introduced to denote the importance of discerning the most likely explanation of a hypothesis.

3.1.3.1. Meaning of abduction

Inference, the thinking process of deriving a conclusion from one or more propositions, can be largely divided into two parts: deductive and inductive inference. Deductive inference involves the derivation of a specific conclusion or principles from general facts or principles whereas inductive inference involves the formulation of a more generalized conclusion from specifics. From here, Peirce introduced a new type of inference called abduction. Table 1 below offers an example to highlight the comparison between induction, deduction, and abduction.

Table 1 *Three kinds of inferences (Peirce, 1878)*

Deduction	Rule - All the beans from this bag are white. Case - These beans are from this bag. Result - These beans are white.
Induction	Case - These beans are from this bag. Result - These beans are white. Rule - All the beans from this bag are white.
Abduction	Rule - All the beans from this bag are white. Result - These beans are white. Case - These beans are from this bag.

From Table 1, each of the inferences is formed with three different propositions:

rule, case, and result. As shown in the Table 1, while deduction and induction move from a case to a result or rule, abduction derives a conclusion of a case. Similar to induction, abduction is considered as probable inference by the fact that a conclusion cannot be guaranteed by its premises being true. However, the difference between the two is that only the observed case and result are considered in induction whereas abduction is a process of inventing new theories with explanatory hypotheses (Kim & Lee, 2002). Abduction is "the inference to the best explanation" (Hacking, 2012, p. 408) and constitutes a process of inferring a hypothesis that best explains the given phenomenon.

Abduction, for the method of hypothesis, proposes a conjecture that explains a puzzling or interesting phenomenon (Hacking, 1990). However, it does not refer simply to a guess. Peirce included abduction as a type of inference because, according to him, one can rank the hypotheses that occur as capable of explaining a phenomenon one wants to understand in accordance with their plausibility (Gower, 1997). In other words, validity is determined by the human mind through the success of abduction.

Abduction usually consists of three stages (Hacking, 2012, p. 408):

1. Unusual phenomenon A is observed.
2. If B were true, A is no longer unusual.
3. Hence there is a reason to suspect that B is true.

In contrast to inductive inference where the hypothesis B becomes true based on

numerous occurrences of phenomenon A, the pattern that clearly explains the occurrence is more important than the facts themselves in abduction (Hacking, 2012, p. 408). With its key role in drawing a hypothesis that explains observed occurrences, abduction is widely used in solving criminals, determining medical diagnoses, and explaining human behaviors (Holland, Holyoak, Nisbett, & Thagard, 1989). Rao (1997) considered abduction similar to a case through which new theories are proposed without any data that are based purely on intuition or flash of imagination and are verified later by conducting experiments.

3.1.3.2. Generation of hypotheses from data

Ho (1994) explained the relation between exploratory data analysis (EDA) and abduction. As EDA acts as a producer of the model for confirmatory data analysis (CDA), abduction, the invention of an idea or hypothesis, can play a role in exploring the feasible paths of EDA. Hence, abduction supports EDA well. Ho (1994) showed that the objective of abduction is to develop a plausible hypothesis by finding a pattern from data.

Consider the following example of abduction from Peirce (1878).

> Suppose I enter a room and there find a number of bags, containing different kinds of beans. On the table there is a handful of white beans; and, after some searching, I find one of the bags that contains white beans only. I at once infer as a probability, or as a fair guess, that this handful was taken out of that bag. This

sort of inference is called making a hypothesis. It is the inference of a case from a rule and result. (Peirce, 1878, pp. 471-472)

In the example above, the hypothesis of the white beans on the table having originated from the bag that contains white beans becomes more plausible from the characteristic of data, i.e., the color of beans on the table is white. This example is similar to the process of validating the origin of samples on the table from many populations. Originally, under uncertainty, given data may have resulted from any one of possible hypotheses or causes (Rao, 1997). However, it is possible to correlate the data with a hypothesis based on its characteristics and patterns. Peirce (1878) described abduction as a process to reveal the cause of a phenomenon. The relationship between samples and population contains the characteristic of abduction where the hypothesis of population is being constructed from samples that are being recognized with the reason being their similarity to a particular population.

Although the previous view on EDA corresponds well with abduction, EDA has a limitation in that it does not allow for an interpretation beyond what is provided by the given data. The purpose of EDA is to explore the data, searching for interesting patterns, and conclusions from EDA are inferred based on what one sees in the data. The conclusions can apply only to the individuals and circumstances surrounding the data in hand (Ben-Zvi, Gil, & Apel, 2007). In other words, out of the three levels of graphical comprehension described by Curcio (1987), EDA cannot reached the third level of looking beyond the data, which means EDA cannot be used to make a

prediction or inference from the data. According to Peirce, induction infers the existence of phenomena that can be observed in cases that are similar while abduction rests on differing suppositions that are not related to observable phenomena and frequently are impossible observe directly (Peirce, 1878). Its significance in providing an explanatory hypothesis beyond the given data can also be found in statistical inference where the hypothesis regarding a population derives from samples.

Peirce clearly delineated the relationship between induction and abduction. As Hacking (1990) described it, "We frame hypotheses by abduction, and test them by induction" (p. 210). As previously mentioned, statistical inference begins with generating a hypothesis regarding a population from samples. As both the quantification of uncertainty and the use of *modus tollens* in statistical inference only make sense in terms of the existence of a hypothesis, statistical inference as abduction becomes essential. However, Peirce once illustrated that statistical inference can be better described with inductive inference by noting that "the induction is reasoning from a sample taken at random to the whole lot sampled" (Maker et al., 2011). When Peirce, a frequentist himself, explained induction with respect to probability and statistical inference as "a sample taken at random," "sampling," and "randomization," it can be assumed that he was focusing on the justification of statistical inference as inductive inference (Hacking, 1990). Ben-Zvi, Makar, and Bakker (2009) attempted to elaborate on abduction due to the fact that the term "statistical inference" typically refers to inductive inference.

In summary, induction as was reviewed in the previous section and abduction as

reviewed in this section each relate to different aspects of statistical inference, and as such, it is not correct to deem one superior over the other. Inductive inference relates to probability in regard to justifying a claim through multiple trials. Abduction can be associated with the determination of a hypothesis regarding a population from a sample before the occurrence of some phenomena. Both of these aspects are essential in the formation of the foundational components of the thinking process involved in statistical inference. Due to the fact that there is a difference between induction and abduction in regards to the use of probability (Hacking, 1990) and the close relationship between forming a hypothesis by abduction and validating it with induction, it is necessary to recognize them as two separate stages of inference.

3.2. Didactical Analysis of Informal Statistical Inference

Through the epistemological analysis of statistical inference as delineated in section 3.1, the foundational thinking components of statistical inference consist of induction and abduction. Statistical inference as induction can regulate the inherent characteristic of statistical inference in terms of its treatment of uncertainty, and it appears as the trial of quantifying uncertainty using probability and introducing *modus tollens.* Statistical inference as abduction is also significant because it draws conclusions about population from observing the characteristics and patterns of a sample. Through the defining of statistical inference from its thinking components, induction and abduction, essential concepts in statistical inference are revealed.

Statistical inference can be characterized both as the generation of a hypothesis about a population from several statistics of a sample (abduction) and the verification of the induced hypothesis by calculating the probability of a drawn sample using sampling distribution that is made from repeated sampling (induction). Thus, the related concepts include sample, descriptive statistics that correspond to several statistics, and sampling distribution that is the base of calculating probability. By analyzing the concept of a sample in detail, it becomes apparent that it consists of a relationship between a sample and a population, and the variability within the sample.

Because this study aims to suggest an assessment method based on the nature of informal statistical inference, it proves necessary to analyze previous research on informal statistical inference. Because informal statistical inference is a precedence stage of formal statistical inference and it has the educational aim of supporting the understanding of formal statistical inference, determination of the inherent characteristics of informal statistical inference as agreed upon in previous studies is imperative. From previous research, the meaning of informal statistical inference and how the essential concepts and thinking in statistical inference are treated in informal statistical inference will be examined. It will also prove necessary to consider the methods for designing assessments.

Section 3.2.1 provides the definition and components of informal statistical inference. Section 3.2.2 presents the essential concepts in statistical inference, that is, descriptive statistics, sample and population, a size of sample, and sampling distribution, and how they are treated in informal statistical inference. Section 3.2.3

reviews the treatment of the essential thinking in statistical inference, that is, abduction and induction. In particular, the examination of the treatment of abduction and induction in the regard of argumentation is examined because informal statistical inference closely relates to argumentation (Ben-Zvi, 2006; Paparistodemou & Meletiou-Mavrotheris, 2008).

3.2.1. Definition and Components of Informal Statistical Inference

Statistical education research has emphasized informal statistical inference in order to support the understanding of formal statistical inference, such as estimation and hypothesis testing, which are taught at the tertiary level (Watson, 2008; Ben-Zvi & Gil, 2010). Zieffler et al. (2008) claimed that because statistical inference integrates many important ideas in statistics, such as data representation, measures of center and variation, the normal distribution, and sampling, introducing informal statistical inference early could provide students with multiple opportunities to build the conceptual framework required to support statistical inference (p. 46). Reading (as cited in Zieffler et al., 2008, p. 46) have suggested that the presentation of informal statistical inference in schools engages students in the foundational ideas of statistical inference, such as generalizing to an appropriate population beyond a collected sample, basing inferences on evidence, choosing between competing models (i.e., hypothesis), expressing a degree of uncertainty in making an inference, and making connections between the results and the problem context. Because of these reasons, many studies

have been conducted to regulate the definition of informal statistical inference and to establish the components of it in order to support future teaching and learning implications (Ben-Zvi et al., 2007; Zieffler et al., 2008; Makar & Rubin, 2009; Rubin, Hammerman, & Konold, 2006).

Informal statistical inference is similar to EDA because both entail the observation of patterns of data in order to gain an understanding from data. However, EDA is classified as descriptive statistics and its conclusions are informal, inferred from what one can see in the data and thus apply only to the individuals and circumstances about which the data concern (Ben-Zvi et al., 2007). Cobb and Moore (1997) presented the distinctions between EDA and formal statistical inference (see Table 2).

Table 2 *Distinctions between EDA and formal statistical inference (Cobb & Moore, 1997, p. 808)*

Exploratory Data Analysis	Statistical Inference
Purpose is unrestricted exploration of the data, searching for interesting patterns.	Purpose is to answer specific questions, posted before the data were produced.
Conclusions apply only to the individuals and circumstances for which we have data in hand.	Conclusions apply to a larger group of individuals or a broader class of circumstances.
Conclusions are informal, based on what we see in the data.	Conclusions are formal, backed by a statement of our confidence in them.

Ben-Zvi et al. (2007) suggested that the positioning of informal statistical inference as the "bridge" between EDA and formal statistical inference (p. 2). From this point, they suggested a definition of informal statistical inference as follows:

> Informal statistical inference refers to the cognitive activities involved in informally drawing conclusions or making predictions about "some wider universe" from patterns, representations, statistical measures and statistical models of random samples, while attending to the strength and limitations of the sampling and the drawn inferences. (Ben-Zvi et al., 2007, p. 2)

Zieffler et al. (2008) defined informal statistical inference as the way in which students apply their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples (p. 44). In addition, Makar and Rubin (2009) defined it as a reasoned but informal process of creating or testing generalizations from data, that is, not necessarily through standard statistical procedures. Researchers presented various definitions of informal statistical inference, but the common point is that the essence of informal statistical inference is based on statistical knowledge about descriptive statistics and leads to conclusions about population from sample.

Based on this definition, researchers provided the components of informal statistical inference. Rubin et al. (2006) presented the related ideas of informal statistical inference as follows:

- Focusing on properties of aggregates such as signal and noise, and types of variability rather than properties of individual cases
- Considering the relation between sample size and accuracy of estimation of

population

- Controlling for bias when sampling from a population
- Distinguishing between claims that are always true and those that are often or sometimes true (Rubin et al., 2006, p. 2)

Zieffler et al. (2008) presented three components of informal statistical inference as follows:

- Making judgments, claims, or predictions about populations based on samples, but not using formal statistical procedures and methods (e.g., p-value, t tests)
- Drawing on, utilizing, and integrating prior knowledge (e.g., formal knowledge about foundational concepts, such as distribution or average; informal knowledge about inference such as recognition that a sample may be surprising given a particular claim; use of statistical language), to the extent that this knowledge is available
- Articulating evidence-based arguments for judgments, claims, or predictions about populations based on samples (Zieffler et al., 2008, p. 45)

Makar and Rubin (2009) determined three key principles to be essential to informal statistical inference. In the following the first of these principles is particular to the process of inference whereas the latter two are specific to statistics (Makar &

Rubin, 2009, p. 85):

- Generalization, including predictions, parameter estimates, and conclusions, that extend beyond describing the given data
- The use of data as evidence for those generalizations
- Employment of probabilistic language in describing the generalization, including informal reference to levels of certainty about the conclusions drawn (Makar & Rubin, 2009, p. 85)

While the above researchers presented the components of informal statistical inference through a theoretical lens, Pfannkuch (2006; 2007) provided the components through a practical lens in her examination of the elements foundational to teachers' and students' engagement in the informal statistical inference. Pfannkuch (2006) analyzed the teacher's communication when she compared two boxplot distributions by informal statistical inference. As a result, Pfannkuch extracted ten elements: hypothesis generation, summary, shift, signal, spread, sampling, explanatory, individual case, evaluative, and referent. The extraction from the teacher's informal statistical inference contributes to the research base by enhancing the understanding of the reasoning processes and raising issues about the links to formal inference, the nature of the informal statistical inference, and instructional practice (Pfannkuch, 2006, p. 43). In addition, based on the elements, Pfannkuch (2007) was able to construct a hierarchical model for students' abilities at informal

statistical inference and assessed students' responses to a task.

Meanwhile, Makar et al. (2011) argued that there are a number of interrelated key elements that are required to support students' informal statistical inference. By analyzing educational literature on informal statistical inference and philosophical literature about inference, researchers presented the key elements that support informal statistical inference: statistical knowledge, knowledge about the problem context, useful norms and habits, doubt and belief, and an inquiry-based environment that includes tasks, tools, and scaffolds. Detailed explanations of each element are described in Table 3.

A fundamental difference between Table 3 and components of informal statistical inference in other research lies in the fact that Table 3 includes social elements, such as collaboration norms. Other research placed emphasis on argumentation in respect to the drawing of generalizations based on data. However, Makar et al. (2011) established much more specific social elements, such as seeking peer consensus and clarification, privileging interesting outcomes, and basing conclusions on data.

Table 3 *Overview of elements supporting informal statistical inference (Makar et al., 2011, p. 160)*

Category	Cognitive and Social Elements	Examples
Statistical knowledge	Statistical concepts	Variability, distribution, inference, sampling
	Statistical ways of thinking	Aggregate thinking, thinking in terms of tendency
Contextual knowledge	Knowledge about problem context	Awareness of possible relationships between elements in the situation (e.g., relationship between gender and jumping distance)
Norms and habits	Collaboration norms	Seeking peer consensus and clarification
	Inquiry norms	Privileging interesting outcomes, basing conclusions on data
	Habits of mind	Seeking explanation; critical stance toward data-based claims, inquiry spirit, perseverance to achieve an inquiry goal, flexibility and creativity
	Habits of action	Organizing and plotting data and other actions common in exploratory data analysis
Inquiry drivers	Belief (hypothesis, expectation)	"Older students jump farther than younger students"
	Doubt (conflict)	Conclusion from a sample contradicts hypothesis
	Explanation	Coming up with an explanation for a small sample being non-representative of the population
Design elements	Task	Statistical investigation
	Computer tool	TinkerPlots, Excel
	Scaffolds	Teacher prompts, questions, peer collaboration

By analyzing previous research on informal statistical inference that focuses on its definition and components, several findings are induced. Firstly, the essence of informal statistical inference is based on statistical knowledge about descriptive

statistics and the ascertaining of a conclusion about a population from a sample. Secondly, components of informal statistical inference are divided into two components: concepts and thinking. The concepts components include concepts about descriptive statistics (such as expectation, variation, and distribution), sample and population, the size of sample, and the variability related to the size of sample. In addition, as recommended by NCTM (2000), a sampling distribution concept is required because students should use sampling distributions as the basis for informal statistical inference. These concepts are mostly similar to the essential concepts in statistical inference that were mentioned in the introduction of section 3.2; thus, one can argue that these concepts are emphasized in informal statistical inference as well. The thinking components include making generalizations based on the data, using data as evidence for generalizations, using probabilistic language, and including informal reference to levels of certainty. Also, according to Makar et al. (2011), the social aspect of argumentation is important in informal statistical inference.

3.2.2. Concepts Emphasized in Informal Statistical Inference

In the introduction of section 3.2., the essential concepts in statistical inference were ascertained through defining statistical inference from thinking components, induction, and abduction. These conceptions include descriptive statistics, such as expectation and variation, sample and population, a size of sample, and sampling distribution. In this section, information on the treatment of these concepts in informal

statistical inference and how students should understand each concept is presented.

3.2.2.1. Concepts of expectation and variation[1] in informal statistical inference

Descriptive statistical concepts such as expectation and variation are fundamental concepts in statistical inference. Sample mean is used when checking the significance of the difference of two groups. Hypothesis testing requires an understanding of the relationship between variability and probability distribution. Among the components of informal statistical inference, "properties of aggregates such as signal and noise, and types of variability" (Rubin et al., 2006), "formal knowledge about foundational concepts, such as distribution or average" (Zieffler et al., 2008), "summary," "signal," and "spread" (Pfannkuch, 2006) correspond to descriptive statistical concepts.

Watson, Callingham, and Kelly (2007) explored students' statistical understanding about expectation and variation. The value of this study is that it identified students' understanding about expectation and variation in the context of informal statistical inference and about the interaction between them. Researchers presented six contexts, including probability sampling, representation of temperature change, beginning inference, independent events, the relationship of sample and population, and description of variation. From the analysis of students' responses, they are classified into six levels as given in Table 4.

[1] The term "variability" is defined as the characteristic of the entity that is observable, and the term "variation" means the describing or measuring of that characteristic (Reading & Shaughnessy, 2004, p. 202). In school mathematics, variance or standard deviation is used primarily as a measure of variation, but in here variation will be described in the sense of presenting possible changes in response to the average, central tendency.

Table 4 *Levels of conceptual understanding of expectation and variation (Watson et al., 2007)*

	Level	Character
Level 1	Idiosyncratic	Little or no appreciation of either expectation or variation.
Level 2	Informal	Primitive or single aspects of expectation and/or variation and no interaction of the two.
Level 3	Inconsistent	Acknowledgement of expectation and variation, often with support, but few links between them.
Level 4	Consistent	Appreciation of both expectation and variation with the beginning of acknowledged interaction between them.
Level 5	Distributional	Established links between proportional expectation and variation in a single setting.
Level 6	Comparative Distributional	Established links between expectation and variation in comparative settings with proportional reasoning.

The levels were determined on the basis of the development of students' understanding of the interaction between expectation and variation. At level 1, students have little or no appreciation of either expectation or variation, and at level 2 students have attained primitive aspects of expectation and/or variation. Students at level 3 can recognize each concept but lack comprehension of the links between them. At level 4, students start to acknowledge the interactions between the two concepts. In the context of informal statistical inference, expectation is appreciated as "center" or "periodical tendency," and variation is considered in terms of "small change" or "random." At levels 5 and 6, students appreciate both expectation and variation in connection to "distribution." Expectation and variation are appreciated as "proportion" and "unexpected change," respectively. At level 5, students can explain the interaction in a single context while at level 6, they can do so across the various contexts.

Bakker, Kent, Derry, Noss, and Hoyles (2008) identified the importance of the

understanding of average (in relation to a target), variation (should be within certain limits), and distribution (roughly bell-shaped) from informal statistical inference regarding employees in a workplace. For this reason, they suggested that a comparison of data with a model, especially a comparison of data statistics with those of a hypothetical distribution, should be included as the key ingredients of statistical inference with generalizations, data as evidence, and the probabilistic language that Makar and Rubin (2007) presented.

Previous studies on informal statistical inference have emphasized expectation and variation as the predominant criteria of comparison in inference. It may be expected that expectation and variation can be easily applied to informal statistical inferences because the concepts were repeatedly emphasized in descriptive statistics. However, the results of studies do not confirm this conclusion. Therefore, it is necessary to emphasize that expectation and variation must be understood in relation to distribution and appreciated as the criteria of comparison for inference in the context of informal statistical inference.

3.2.2.2. Concepts of sample in informal statistical inference

A sample is a beginning and the most fundamental concept of statistical inference. Among the components of informal statistical inference, "making judgments, claims, or predictions about populations based on samples," "informal knowledge about inference such as recognition that a sample may be surprising given a particular claim" (Zieffler et al., 2008), "generalization that extend beyond describing the given data,"

and “the use of data as evidence for those generalizations” (Makar & Rubin, 2009) indicate the intrinsic importance of data, i.e., the sample. It is important to understand samples in relation to populations. Below, the key aspects of samples will be presented as they relate to the informal statistical inference.

Saldanha and Thompson (2002) distinguished two conceptions of sample as addictive and multiplicative from a teaching experiment about samples from a high school. In the addictive conception of sample, a sample is viewed as a subset of a population with the resemblance between sample and population not taken into consideration. In this conception, multiple samples are treated as multiple subsets of a population, and there is no information regarding an inference about the population. However, in the multiplicative conception, a sample is regarded a quasi-proportional version of a population. For the informal statistical inference, the multiplicative conception of a sample is important because the aim of statistical inference entails the possession of information from samples with quasi-proportionality when repeating several sampling processes.

Rubin, Bruce, and Tenney (1991) presented the results of an investigation of the concept of sample that is fundamental to understanding statistical inference. Researchers explained the concept of sample in terms of sample representativeness and sample variability. Sample representativeness constitutes the idea that a sample taken from a population holds characteristics similar to the population. For example, the proportion of objects is likely to be similar to the proportion found in the population. However, sample variability involves a contrasting idea. Samples taken

from a population are not always the same, thus don't apply in the same way as to the original population. For example, when it comes to a high school whose overall population of male and female students is evenly divided, it remains possible to have more boys than girls in a certain class. Sample representativeness and sample variability are contradictory according to a deterministic view. Rubin et al. (1991, p. 315) suggested that the key to understanding statistical inference is to balance the two ideas and to apply the meaning of "likely" in each as well.

In summary, a sample is important because it is the start and foundation of statistical inference. A proper informal statistical inference requires an understanding of quasi-proportionality between a sample and its population. As the expression "quasi-proportionality" indicates, it should be appreciated that samples can explain populations, not completely but partly. Furthermore, for an informal statistical inference, a balance is needed between sample representativeness and sample variability that are contradictory according to a deterministic view but complementary according to a non-deterministic view. The conception of sample should be considered when conclusions are based on samples, and justifications for those conclusions should be provided.

3.2.2.3. Concepts of size of sample in informal statistical inference

In the previous section, the importance of sample variability in conjunction with sample representativeness was demonstrated. While sample variability was explained in accordance with the relationship between sample and population, this section is

considers it in terms of its size. In statistical inference, the size of a sample highly relates to sampling variability. The type of hypothesis testing and the shape of a sampling distribution may differ according to its size. The importance of the size of a sample is presented in "considering the relation between size of sample and accuracy of estimation of population" and "controlling for bias when sampling from a population" (Rubin et al., 2006) among the components of informal statistical inference.

Rubin et al. (1991) claimed that the size of a sample is critical for establishing the relationship between sample representativeness and sample variability. Larger sample sizes depend greatly on sample representativeness whereas smaller ones depend more on sample variability. According to the law of large number, statistical probability approximates to mathematical probability as the size of a sample increases. The law seems to be intuitive. This means that the influence of size of sample is easily considered in statistical inference. Sedlmeier and Gigerenzer (1997) presented a group considering the size of the sample intuitively. When asked about the questions presented in an experiment of Piaget and Inhelder (as cited in Sedlmeier & Gigerenzer, 1997, p. 34) concerning what would happen if a ball was dropped in the simplest Galton board (device that has a funnel and is divided into two sections), students easily answered that the results would be closer to the uniform distribution as more balls were dropped. On the other hand, Pfannkuch (2008) suggested that it is easy to obtain information when presented with a sample larger size because the population distribution and sample distribution show results similarly. They assumed that the size

of the sample could serve as an axis when considering the sample variability.

The concept of the size of a sample mentioned so far does not directly relate to the formal statistical inference, in particular the sampling variability. Specifically, larger samples sizes provide more information about a population, which entail the drawing of inferences about populations on the basis of one sample. In general, however, a statistical inference is not based on only one sample because sampling distribution can only be obtained by repeated sampling and entails a probability calculation. Nevertheless, statistical inference about a population from one sample holds some importance in the drawing of an informal statistical inference. A representative instruction activity called “Growing samples” (Bakkar, 2004; Ben-Zvi, 2006, Ben-Zvi & Aridor, 2012) served to present the advantages and limitations of inferences about a population as the size of the sample increases. This activity comprehensively covers the concepts of data, distribution, variability, and sampling, and provides an opportunity to present a persuasive argument to the database or explanation to students concerning the uncertainty present in a statistical inference (Ben-Zvi & Aridor, 2012). Thus, key factor in regard to the size of a sample for an informal statistical inference is that as the size of a sample increases, the sample can provide both more information and more reliable information about a population.

3.2.2.4. Concepts of sampling distribution in informal statistical inference

Statistical inference about a population is not drawn from one sample but through comparison with a sample on the basis of sampling distribution, a distribution made

by infinite sampling, and afterward is used as an evidence of probability calculation. Thus, sampling distribution can be considered the most important concept for drawing a statistical inference. The previous section explored the influence of the size of a sample when an inference about a population is made by a given sample. To repeat, in formal statistical inference an inference about a population would never be determined from a single sample, a rule that informal statistical inference does not abide. The components of informal statistical inference with relation to the size of samples presented in the previous section are applicable in this context but are more appropriate for this section. Other than components from the studies about the teaching and learning of informal statistical inference, the "use sampling distribution as the basis for informal inference" (NCTM, 2000) also suggests the value of sampling distribution in informal statistical inference.

Sampling distribution is considered an integrated concept in several previous studies. Sampling distribution is a concept that requires integration with repeated random sampling, variability, and distribution (Saldanha & Thompson, 2002) and also can be thought of as an abstract concept to be integrated with sample, population, distribution, and variability (Chance, delMas, & Garfield, 2004). Lee and Lee (2010) classified statistical concepts as instrumental (distribution), validity (summary), and reliability (sample) and went on to denote sampling distribution as an integrated concept that should include aspects of those other types. Because of the integrated nature of sampling distribution, students frequently face difficulties in regard to it. It may be difficult for students who are accustomed to the inference with a sample to

detect the distribution of sample statistics. Furthermore, as can demonstrated from the nature of infinite sampling, sampling distribution is a theoretical distribution introduced by a deductive approach on the basis of probability theory (Lipson, 2002), which results in students often facing difficulties in even detecting the concept. Therefore, other approaches must be offered to resolve the difficulties that arise from the theoretical and probabilistic approaches and to introduce the necessity of sampling distribution. With a focus on these points, sampling distribution will be reviewed with regard to informal statistical inference.

Garfield and Ben-Zvi (2008) have suggested that make students recognize the necessity of sampling distribution by providing them with informal activities. Researchers noted activities both about sample variability and the idea of a random sample and suggested an activity to help reveal results can differ due to changes in the size of the sample (Garfield & Ben-Zvi, 2008, p. 249). Specifically, the activities address how samples can vary when sampling from the same population and how random sample can differ through repeated sampling. The activity that graphing the distribution of sample statistics for many samples from the same population allows students to see whether one sample is unlikely or not based on the comparison to the rest of the samples. This activity is also be considered as an informal precursor to the more formal notion of p-value in the future (Garfield & Ben-Zvi, 2008, p. 236). Garfield and Ben-Zvi suggested that the concept of sampling distribution can become evident to students once they can recognize its necessity, which can be achieved by presenting an open-ended problem, such as “How to measure the average height of a

high school's sophomore students" and asking students to consider how to produce a sample and how to choose the best method of extracting a sample that is similar to the population.

Pfannkuch (2005) proposed that students' informal statistical reasoning could be promoted by presenting them with an empirical sampling distribution. Pfannkuch understood this to be an alternative for students who experienced difficulty with the concept of sampling distribution as delineated in the theoretical and probabilistic approaches. The method she presented was called the resampling approach, which provides a theoretical foundation of statistical inference by considering what happens when sampling is repeated several times. Consider an example in which students are given a maximum temperature of the sample data for the two cities of Napier and Wellington and the difference between the medians of each sample was 2.2. From here a student describes the dilemmas as follows:

> Although I know that my particular samples for Napier and Wellington maximum temperatures have these particular medians and spreads, I know that if I repeat this study with the same size of sample, I will get different values. So, is this difference I see between Napier and Wellington maximum temperatures due to 'chance' (random or sampling variation) or is there a real difference? (Pfannkuch, 2005, p. 288)

In this example the resampling approach randomly reassigned all the maximum

temperature values to Napier and Wellington. Under the conditions that the value was determined arbitrarily, the difference in the median of the maximum temperature of the two cities can be said to appear by chance. Suppose a person constructed a sample data continuously for the maximum temperature of the two cities by using a computer. Then the person repeated it 1,000 times and each time calculated the difference between the median values. The resulting histogram (Shown in Figure 2) that represents the empirical distribution of the difference between the medians can be generated, and from it, one could calculate the p-value. From the histogram, students could notice the case that the difference of median greater than 2.2 is less than once in 1,000, which means the data has provided strong evidence for the difference in median temperatures in the two cities.

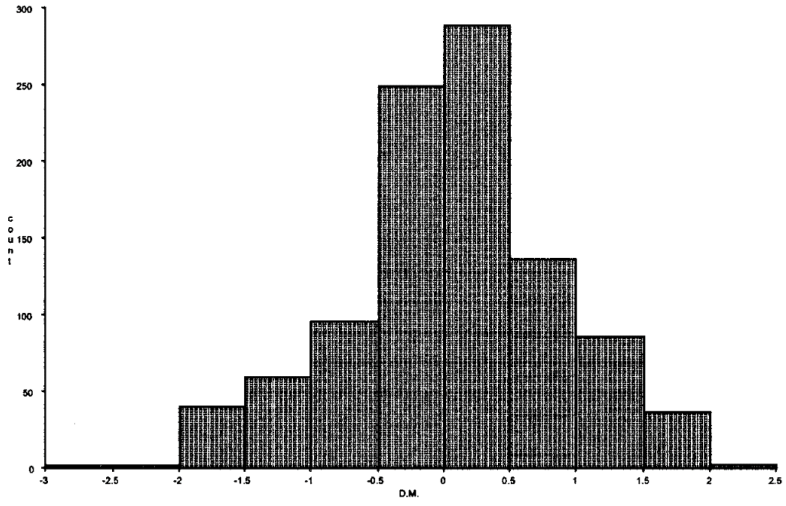


Figure 2 Empirical sampling distribution of the difference in medians for 1000 resamplings (Pfannkuch, 2005, p. 289)

Pfannkuch (2005) regarded the resampling approach as parallel with the concept of p-value based on the fact that the logic of p-value is based on the degree to which the outcome was surprising, which is more about assessing the strength of the evidence for the populations under consideration. The resampling approach that Pfannkuch presented is similar to the empirical approaches of sampling distribution suggested by Lipson (2002). Lipson proposed that key to understanding a statistical inference involved knowing that sampling distribution is a distribution of sample statistics based on the observation of several samples and then identifying the likelihood of the observed sample statistics as compared to the sampling distribution. Thus, students could easily access the sampling distribution by gaining a cognizance of the dynamic process of sampling.

The example Pfannkuch presented allows students the opportunity to witness the practical process of creating a sampling distribution. During the process, students could identify sampling distribution as a distribution of statistics. In fact, this study may not have only demonstrated the concept of sampling distribution but also the concept of null hypothesis, meaning that there was no difference of maximum temperature due to a random assignment of the data of maximum temperature to the two cities. Additionally it offered access to a hypothesis testing procedure in an informal manner. This study also informally offered students the opportunity to apply probability during the process of statistical inference and the combination of sampling distribution and statistical inference by allowing students to infer whether the difference of sample data occurred accidentally or not.

In Sedlmeier and Gigerenzer's (1997) research concerning the concept of size of sample, they found that people in the sampling distribution situation have difficulty even conceiving of the concept of the sample size. The following situation describes it.

> A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower. (Sedlmeier & Gigerenzer, 1997, p. 37)

Researchers argued that this problem is quite difficult because the inference is based on the sampling distribution, not the frequency distribution when comparing the results to those of Piaget and Inhelder's (as cited in Sedlmeier & Gigerenzer, 1997, p. 35) experiment. Therefore, understanding sampling distribution in relation to the size of a sample is essential. This also relates to the central limit theorem, which proposes that as the size of sample increases, the sampling distribution follows the normal distribution and the standard deviation at decreasing rates. Conversely, the hospitals where boys are born at a frequency greater than 60% are small hospitals because the deviation from the true proportion (50%) is more likely to be found in their smaller sizes of sample.

In summary, studies on sampling distribution in relation to informal statistical

inference revealed sample variability and promoted both the random sample and activities dealing with the changes that result by varying the sizes of sample. Through these students could become aware of the necessity of sampling distribution. In addition, the creation of the sampling distribution through a computer simulation offers easy access to the empirical sampling distribution concept that students often find difficult to understand because it is typically considered only through theoretical and probabilistic approaches. The creation of sampling distribution can lead to students' applying inference using probability in statistical inference. Students should be able to engage with the concept of the empirical sampling distribution and be cognizant of the necessity of sampling distribution. Students should also comprehend the relationship between the size of a sample and the sampling distribution.

3.2.3. Thinking Emphasized in Informal Statistical Inference

In the previous section, by epistemological analysis of the statistical inference, it has been determined that both abduction and induction are meaningful thinking components of statistical inference. Now, it is necessary to examine the treatment of these thinking components in regard to informal statistical inference and the type of ability required of students in applying each process.

Before addressing the thinking components, it is important to investigate the relationship between argumentation and informal statistical inference. Argument is understood as a string of claims each of which consists of a premise and a conclusion

with the purpose of an argument being the verification of a conclusion based on certain grounds (Kim, 2007, p. 26). Hacking (1980) considered an inference a concluding statement. For example, in arguments "S, so H," the inference is the conclusion H. Because statistical inference is composed of a process of drawing a conclusion about population based on the data regarding a sample by considering uncertainty, the transition from data to conclusion is important. From the view of Hacking (1980), in statistical inference, the whole argument process, encompassing the transition from data to conclusion, is greater importance than that of the inference as a conclusion.

In the case of informal statistical inference, there is no employment of formal statistical procedures and methods, such as p-value or t tests (Zieffler et al., 2008). Instead, it relies on data and context to draw a conclusion. Thus, the connection between data and a conclusion becomes quite important in informal statistical inference, which shows the close relationship between informal statistical inference and arguments. Informal statistical inference should be conducted based on argumentation. Ben-Zvi (2006) cited just such a similarity between informal statistical inference and argumentation in proposing that the derivation of logical conclusions from data should be accompanied by the need to provide persuasive arguments based on data analysis.

In summary, students should present informal statistical inference based on argumentation. For this reason, it is necessary to cultivate a classroom environment that encourages argumentation. More particularly, statistics classrooms should have norms articulating that informal statistical inferences should be carried out based on

argumentation. The norms of this type can be characterized as social norms in the classroom (Cobb, Stephan, McClain, & Gravemeijer, 2001).

In light of this background, the characteristics of informal statistical inference based on argumentation will be investigated in section 3.2.3.1, which induces the required situation for informal statistical inference. That situation will be the one that shows the thinking emphasized in informal statistical inference in section 3.2.3.2 and 3.2.3.3. Then the treatment of abduction and induction in regard to informal statistical inference based on argumentation will outlined. The conclusion will argue that in drawing an informal statistical inference, abduction is expressed as a construction of argumentation and induction as a verification of argumentation. Section 3.2.3.2 offers an exploration of the elements necessary for consideration when students construct an argumentation and section 3.2.3.3 explores possibilities for students to try when verifying their argumentation.

3.2.3.1. The characteristics of informal statistical inference based on argumentation

In statistics education, the importance of teaching and learning based on argumentation has been highlighted (Cobb & McClain, 2004; Garfield & Ben-Zvi, 2009). Cobb and McClain (2004) emphasized that teachers should achieve their instructional agendas by building on argumentation that students produce on the basis of data, and instruction should include communication that is based on the argumentation (p. 392). Garfield and Ben-Zvi (2009) argued that students should

participate in classroom discourse that is associated with the statistical reasoning with emphasis on the knowledge about data and context. Students should be required to participate in discussions by proposing an argumentation. As previously mentioned, informal statistical inference should be based on argumentation due to the importance of the transition from data to a conclusion. This necessitates a clear description of the situations for conducting an informal statistical inference based on argumentation. In this section, the characteristics of informal statistical inference based on argumentation will be investigated to induce the requirements for situations in which informal statistical inference occur.

The first characteristic of informal statistical inference based on argumentation is that it is carried out with natural language. In other words, students employ everyday language rather than formal terms or symbols when discussing informal statistical inference and that the language reflects students' statistical reasoning. For example, Bakker and Gravemeijer (2004) revealed that concerning the issue of distribution, students used informal words such as "hill" and "bump" to describe shape when they generate a hypothesis by comparing the two sets of data. Researchers commented that students use such meaningful but informal terms and argued that teachers should identify students' levels of understanding in their informal words. Teachers should try to learn about how students reason by listening and observing (p. 166).

The second characteristic is that argumentation is informal. Informal statistical inference is based on argumentation because of the importance of transition from data to a conclusion. However, the components of argumentation are not immediately

evident. According to Toulmin (1958), argumentation consists of the core components of data, warrant, and claim that through backing assures a warrant, but such a frame does not hold for informal statistical inference. Furthermore, it does not follow the symbols and logical operations used in formal argumentation. Informal statistical inference involves making an argumentation on the basis of a sample with prior knowledge such as knowledge about foundational statistical concepts and informal knowledge about inference as the guiding evidence (Zieffler et al., 2008, p. 45). It is also combined with natural language as mentioned above. Therefore, informality characterizes its nature.

The third characteristic is that it is based on the context. Context is a key component in the development of students' informal statistical inference (Ben-Zvi & Aridor, 2012). Context can greatly influence informal statistical inference. As data in statistics necessarily constitutes numbers embedded in particular contexts (Cobb & Moore, 1997), it is necessary to derive a plausible inference corresponding to the meaning of data. This context could provide language to support student discussion about statistical ideas as way of scaffolding students' development of informal statistical inference (Makar et al., 2011, p. 156) and aid them in finding meaning from observed patterns (Pfannkuch, 2011). Sometimes, a conflict between data and context could be resolved by a contextual-based explanation and was often substantiated by a statistical argument (Ben-Zvi & Gil, 2010).

A fourth characteristic is that argumentation occurs through interaction. Krummheuer (1995) described argumentation as a social phenomenon in which

individuals adjusted their intentions and interpretations by verbally presenting the rationale of their actions (p. 229). Makar et al. (2011) claimed that social factors have meaning in informal statistical inference. Argumentation as an informal statistical inference is not constructed by a person's comment, which suggests that students need to be explicit in explaining and justifying their insights before debating with peers (Makar et al., 2011). Through the interactive process, students can find more appropriate perspectives and evidence to support their conclusion. Students' collaborative work also assisted in this process especially in situations in which each inferred from a different perspective. This situation elicited fruitful discussions that contributed to the development of their reasoning about data and context (Ben-Zvi & Aridor, 2012). Teachers play a critical role as a member of interactions as well. They can establish the norms required in inquiry processes and scaffold students by prompting questions (Makar et al., 2011).

Because of its basis in natural language, informal argumentation, context, and interaction, informal statistical inference reveals itself more apparently in verbal than in written language. Written language has been recognized as a useful tool for learning mathematics. Writing requires students to use higher cognitive functions such as analyzing and synthesizing information (Albert, 2000). Albert (2000) considered writing a device for mediating cognitive development that enhanced self-assisted practice and promoted the product of self-regulatory processes. Albert is careful to note that writing relies on the formal meaning of words, and written language provides opportunities to use oral language out of a social context. These characteristics of

written language contradict the very nature informal statistical inference, which requires the use of natural language and is based on context. Thus, real informal statistical inference can only be achieved through verbal language. Given the characteristics of informal statistical inference based on the argumentation, requirements for informal statistical inference can be described as a "situation [in which] occurs communication based on argumentation using verbal language." It is through the situation that the thinking emphasized in informal statistical inference emerges.

By focusing on communication through argumentation in teaching and learning of informal statistical inference students are encouraged to use statistical concepts and to promote their development of the concepts. Bakker and Derry (2011) raised a question that which should be precede among statistical inference and concept learning. Researchers wanted to find an answer in inferentialism from the perspective of Brandom (as cited in Bakker & Derry, 2011, p. 6) that bases the essence of human knowing on inference. From the perspective of inferentialism, the meaning of a concept emerges through its use so that it is constructed through an activity or in social practice. According to inferentialism, students can participate in informal statistical inference even before they have fully formed the concept because it is during the process of using the concept that the meaning of the concept is formed and developed (Bakker & Derry, 2011). Therefore, the context of the informal statistical inference situation informs students' understanding and use of it.

3.2.3.2. Abduction and constructing argumentation

As previously argued, the meaning of abduction in statistical inference lies in the drawing of conclusions about populations based on characteristics and patterns of the sample. These aspects include "articulating evidence-based arguments for judgments, claims, or predictions about populations based on samples" (Zieffler et al., 2008), "generalization that extend beyond describing the given data," "the use of data as evidence for those generalizations," (Makar & Rubin, 2009) and "hypothesis generalization" (Pfannkuch, 2006). In other words, the process of drawing conclusions goes beyond the data as evidence, and from here argumentation is constructed.

Abelson (1995) suggested that the purpose of statistics is organizing a useful argument from quantitative evidence using rhetoric based on the principle. He argued that statistics, a field of dealing with uncertainty, should be used as a means of persuasion. To achieve this, one should address statistics as a narrative that includes rhetorical aspects. Abelson (1995) claimed, "The investigator must combine the skills of an honest lawyer, a good detective, and a good storyteller" (p. 16). In other words, the investigator should discover patterns from data while at the same time, through logical construction, formulate interesting claims. Therefore, he proposed that all the processes that lead to the development of explanation or interesting assertion from the data whether they describe the evidence as a systematic variation or chance variation are important. Based on Abelson's (1995) studies, Ben-Zvi (2006) proposed two critical dimensions in argumentation. One is an action or process about inference that

draws logical conclusions from the data, and the other is an action or process about rhetoric and narrative that provides a persuasive argumentation on the basis of the data analysis. He advocated that these two dimensions must also appear in informal statistical inferences. The processes are related to the constructing argumentation. From here, it becomes evident that a framework for students' construction of argumentation must be clearly articulated.

Paparistodemou and Meletiou-Mavrotheris (2008) identified how students' informal statistical inference most likely can emerge in the statistics class that employ argumentation from data. In this study, third grade students participated in an activity intended to draw various conclusions based on the data of a card containing 16 kinds of students' characteristics such as gender, grade level, hobbies, and so on. Then the participating students' resulting informal statistics inferences were categorized into three types: making argumentation based on data, generalizing though argumentation based on the data, and considering chance during generalizing though argumentation based on the data. The students progressed from focusing on data to focusing on populations by considering context, and from here they drew results beyond the data by gradually considering chance.

It is necessary to recognize the problem context as well as using data as evidence, which has already been outlined as a component of informal statistical inference, for constructing argumentation. Makar and Rubin (2007) stressed the importance not solely on drawing a conclusion from the data but also on extracting a conclusion to portray a situation that explains what the data means. Ben-Zvi and Gil (2010)

suggested that during the situation of informal statistical inference, a trend in the data or a conflict between the data and context can be resolved by a contextual-based explanation and further be substantiated by an argumentation. Ben-Zvi and Aridor (2012) also proposed that by considering data and context together, students could make connections with explanations about data, predictions, and informal statistical inferences.

In summary, argumentation needs to be the form of premise and conclusion to achieve conclusions on the basis of the data, and have persuasive characteristics. Therefore, to construct an argumentation as an informal statistical inference, students must utilize the data as evidence on the basis of its characteristics and patterns and draw conclusions that take into account the meaning of the data by considering the particular context.

3.2.3.3. Induction and verifying argumentation

In addressing the uncertainty that arises from the inference process, it was presented that induction could reveal the characteristics inherent to statistical inference. This claim was achieved through trials involving the quantification of uncertainty using probabilities and the introduction of *modus tollens*. It remains necessary to identify how informal statistical inference addresses these aspects. In informal statistical inference, this dilemma encompassed the matter of how to verify an incomplete connection between the data and the conclusion. Probability, more specifically, using probability models, is a critical component for resolving the

uncertainty in statistical inference. However, most students had not yet encountered probabilistic models when engaging with informal statistical inference, and even if they had learned the models, they had never engaged with any formal procedures for obtaining a p-value corresponding to a particular significance level. Pfannkuch (2005) claimed that the greatest difference between a formal and an informal statistical inference rests in the matter of using or not using probability, even at a basic level. Therefore, even students who were capable of drawing conclusions from a given data only on the basis of previous knowledge such as average, variation, and distribution, were not prepared to grapple with the issue of uncertainty.

Typically the first attempt to appear in such situation involved the application of probabilistic representation, which has been emphasized as a component of informal statistical inference. Rubin et al. (2006) claimed that it is necessary to apply a probabilistic representation when considering the possibilities through the component of "distinguishing between claims that are always true and those that are often or sometimes true." Makar and Rubin (2009) emphasized the important role of probabilistic representation through the component of "employment of probabilistic language in describing the generalization, including informal reference to levels of certainty about the conclusions drawn." Probabilistic representation is required for students who cannot use quantitative measures to address the uncertainty involved in drawing a statistical inference. Probabilistic language encompasses any language appropriate to the situation and level of students to express uncertainty in a speculated hypothesis, that a prediction is only an estimate, or that a conclusion does not apply

to all cases (Makar & Rubin, 2009, p. 87). For example, in using data to estimate the average height of an eight-year-old, students may suggest the typical height to be "around 130–138 cm" rather than reporting more precisely that the typical height "is 132 cm," and a six-year-old child may suggest that the most common way for children to travel to school "may be" by bus rather than stating, "is" by bus (Makar & Rubin, 2009). In fact, Ben-Zvi, Aridor, Makar, and Bakker (2012) explained that students when faced with uncertainty during drawing of an informal statistical inference could change from holding a perspective of extreme determinism or non-determinism in the beginning stages to employing probabilistic language to express sophisticated conclusions.

The second attempt to account for uncertainty is to present the need for more samples in addition to the given ones and to draw a conclusion from the result. From here the relevance of awareness about the importance of repeated sampling and random samples reveals itself. Statistical inference is achieved through a comparison of a data with a random sample on the basis of sampling distribution, a distribution of statistics that appears through infinitely repeated sampling. Saldanha and Thompson (2002) emphasized the multiplicative conception of sample, a stance that conceives a sample as a quasi-proportional, small-scale version of the population. This concept is important because it produces the image of repeated sampling and variability in this process. Rubin et al. (2006) suggested that variability must be recognized as a component of informal statistical inferences. One must be cognizant that samples may vary even when extracted from the same population. When drawing a conclusion from

a given sample in this way, it becomes necessary to elicit new evidence for a conclusion by considering uncertainty in a sample and sampling variability when extracting other samples.

The third method for addressing uncertainty with students centers on the establishment of norms for dealing with uncertainty and arriving at an informal statistical inference. Such norms should include "collaboration norms for the purpose of seeking peer consensus and clarification" (Maker et al., 2011) among the components of informal statistical inference. Bakker and Derry (2011) claimed that students should establish norms for dealing with the uncertainty that can be applied within specific practices, such as, "What is valid judgment or proper reasoning?" and "Is the concept applied correctly?" and from such norms, students can start the process of drawing an informal statistical inference. In a situation where communication takes place on the basis of argumentation, students can resolve uncertainty by looking for more appropriate perspectives and evidence to support a conclusion and by appropriately adjusting their claims as each new perspective, idea, or opinion is voiced. In Ben-Zvi and Sfard's (2007) study, two sixth grade students who solved the problem of inference concerning the determination of the mean of a population from the mean of random samples by predicting the mean of a population from two sample data and expressing the amount of certainty as a number from 1 to 10. The two students attempted to adjust their opinions using expressions such as "I am not 100% sure of this" or "I am certain of this … seven, approximately," until arriving at an agreed answer. Following such norms offers a solid method for dealing with uncertainty in

informal statistical inference and a reliable means to reach a consensus.

In summary, in drawing an informal statistical inference, students must acknowledge the incomplete connections present between data and a conclusion and verify them. For this purpose, students should apply probabilistic representations, draw a conclusion by recognizing the importance of repeated sampling, and attempt to validate the argumentation by establishing norms for dealing with uncertainty during their communications.

3.3. Discussion of the Nature of Informal Statistical Inference

3.3.1. Summary of the Nature of Informal Statistical Inference

Section 3.1 articulates that from its epistemological analysis, it is found that statistical inference consists of two thinking components, abduction and induction. Statistical inference as induction can regulate its inherent characteristic according to how it deals with uncertainty. It appears to be a trial of quantification of uncertainty using probability and introducing *modus tollens*. To quantify uncertainty, there were several approaches include Bayesianism, frequentism, and the likelihood approach with each offering a particular philosophical perspective in regard to the concept of probability. The approaches are distinguished by the extent of subjectivity that each accepts. It is important to choose an appropriate approach based on one's particular situation of statistical inference rather than to regard any single approach as the

absolute truth. *Modus tollens* was introduced to solve the problem that evidence for a certain hypothesis does not provide an absolute certainty but instead provides a probabilistic truth. *Modus tollens* provides a system of logic to show an original hypothesis is true by rejecting the inverse of that original hypothesis through the use of counterevidence to the inverse of the hypothesis. It is significant in that its introduction offers a resolution to the problem of verification in statistical inference. While statistical inference as induction focuses on verifying the conclusion by addressing uncertainty, statistical inference as abduction focuses on generating a plausible hypothesis about a population based on the characteristics and patterns of a sample. Both induction and abduction constitute the integral components of the thinking process in statistical inference. Thus, it is necessary to recognize them as two separate and crucial stages of statistical inference.

Through the didactical review of research on informal statistical inference in section 3.2, the meaning of informal statistical inference and how the essential concepts and thinking in statistical inference are treated in informal statistical inference were examined. The essential concepts include the expectation and variation concept, the sample and population concept, the size of a sample concept, and the sampling distribution concept. The treatment of these concepts in informal statistical inference was delineated, as well as the particular aspects of each concept students needed to understand. Also the section examined the essential thinking included in abduction and induction and the treatment of these thinking components in informal statistical inference. Informal statistical inference should be carried out based on

argumentation. It has the characteristics that it is an informal argumentation using natural language, is based on the context, and occurs within interactions. Thus, the situation lends itself most readily to achieving an informal statistical inference is the one in which communication occurs based on argumentation using verbal language. In this situation, teachers can observe how students use statistical concepts and how they construct and verify the argumentation when carrying out informal statistical inferences. Abduction is treated as constructing argumentation in informal statistical inference. Students should apply data as evidence based on the characteristics and patterns of the data and draw a conclusion considering the meaning of data by recognizing the context at hand. Induction is treated as verification of argumentation in informal statistical inference. Achieving verification of the necessarily imperfect connection between data and a conclusion has proved challenging to students. To solve this, students should verify the argumentation by using probabilistic language, drawing a conclusion using the recognition of the importance of repeated sampling, and establishing norms for dealing with uncertainty during communication.

In summary, the nature of informal statistical inference is based on argumentation and requires a situation in which communication occurs based on argumentation using verbal language. In such a situation, teachers can examine students' use of statistical concepts, construction of argumentation, and verification of argumentation. Abduction and induction that induced from the epistemological analysis of statistical inference are treated as constructing argumentation and verifying argumentation respectively in informal statistical inference. A statistical inference as abduction and

as construction of argumentation are quite similar because of their shared purpose: the drawing of a conclusion from data in consideration of a particular context. Conversely, a statistical inference as induction and as verification of argumentation in informal statistical inference hold vast differences in regard to each one's stance on the treatment of probability. Typically students cannot use the quantification approach because they have not adequately engaged with probability. Thus, it is necessary to address uncertainty in a manner fitting to a student's level.

3.3.2. Implication for Designing Assessment Model of Informal Statistical Inference

When people consider assessment in education, they typically envision the written assessment. Such an assumption holds true in the research of statistics education as well. Thus, when teachers plan an assessment, they consider how to make appropriate items than can reveal students' levels of understanding. However, no matter how seemingly strong the resulting assessment, accurate results will not be revealed if students' capabilities cannot be accessed through writing. The form of an assessment must align with the appropriate method for indicating students' understanding. Thus, the purpose of this study is designing an assessment method that aligns with the nature of informal statistical inference.

From the investigation of the nature of informal statistical inference, it is found that informal statistical inference should be conducted in situations that allow verbal

communication based on argumentation. In such a situation, teachers can examine students' use of statistical concepts, and construction and verification of argument. The purpose of assessing informal statistical inference eventually includes the examination of how students use statistical concepts, construct argumentation, and verify argumentation when they engage in informal statistical inference. Thus, to assess students' abilities accurately, assessment should be implemented when informal statistical inference occurs.

Students have shown difficulty in using language when they do informal statistical inference. Biehler (as cited in Pfannkuch, 2005, p. 269) claimed when comparing boxplots, students faced challenges in describing and interpreting the graphs verbally, for their language was inadequate. Ben-Zvi, Makar, Bakker, and Aridor (2011) characterized students' language describing uncertainty was not clear. In such a situation, it would not be appropriate to examine students' abilities at informal statistical inference through written language or traditionally styled test items because even in their verbal language, students were muddled. Instead, teachers should examine and draw on students' understanding from their verbal language.

Communication based on argumentation occurs in the process of teaching and learning. Thus, assessment should be implemented at the same time with teaching and learning. That is, teachers should examine students' informal statistical inference abilities actively during the process of communicating based on argumentation.

CHAPTER 4

DESIGNING AN ASSESSMENT MODEL OF INFORMAL STATISTICAL INFERENCE

In chapter 3, the implication for designing assessment methods of informal statistical inference was provided by articulating and deliberating about the nature of informal statistical inference. To assess students' informal statistical inference abilities, the situation when communication occurs based on argumentation using verbal language must be incorporated. In that situation, students' ways of using the concepts and constructing and verifying argumentation can be examined. Such a situation occurs in the teaching and learning process. Thus, assessment should occur in the same process, simultaneously with the teaching and learning. In that process, teachers can assess students' informal statistical inference abilities more actively. In this chapter, the meaning of integration of instruction and assessment will be provided. Then, an assessment model for the integration of instruction and assessment will be reviewed to propose an appropriate assessment model concerning informal statistical inference.

4.1. Integration of Instruction and Assessment

4.1.1. Changes of Assessment Perspectives According to Instructional Perspectives

In an educational assessment, assessment should solidly reflect the instructional aim (Shepard, 2000; Pellegrino, Baxter, & Glaser, 1999). As an instruction perspective evolves, the assessment perspective should evolve with it as well. In the following, changing views of knowledge, learning, and assessment from behaviorist, cognitive, sociocultural perspectives are summarized based on a research report from the National Research Council (NRC, 2001, pp. 61-64).

In the behaviorist perspective, knowledge is defined as the organized accumulation of stimulus-response associations that serve as the components of skills. One can learn by acquiring those associations and skills. Initially people acquire the simple components of a skill, and with more exposure and practice, they acquire more complicated skills that combine or differentiate the simpler ones. However, the behaviorist perspective does not address the underlying structures or representations of mental events and processes and the richness of thought and language. In assessment, knowledge is assessed in terms of its component information, skills, and procedures to be acquired. Assessment includes items that are considered significant knowledge in a particular domain. Performance on such items indicates the extent to master the domain.

According to the cognitive perspective, knowledge is more than the accumulation of factual information and routine procedures. It encompasses the ability to integrate information, skills, and procedures in ways that are useful for interpreting situations and solving problems. The cognitive perspective highlights how people develop structures of knowledge, including the concepts associated with a particular domain and the procedures for reasoning and solving problems. People can learn by actively constructing their understanding by trying to connect new information with their prior knowledge. Assessment purpose from this perspective is not only to determine what people know but also to assess how, when, and whether they use what they know. In order to assess cognitive structures and reasoning processes, more complex tasks are necessary which reveal information about thinking patterns, reasoning strategies, and growth in understanding over time.

In the sociocultural perspective, thought is not considered an individual response to task but a behavior at a different level of analysis, one oriented toward practical activity and context. Context indicates engagement in particular forms of practice and community. People learn to participate in the practices, goals, and habits of mind of a particular community. Assessment from this perspective measures the degree to which one can participate in a form of practice. Previous to the advent of the sociocultural perspective, test items did not present the actual contexts in which people use particular knowledge and instead presented test-takers with abstract contexts. However, in the sociocultural perspective, people's performance in the abstract context is not regarded as reflecting how well they would participate in organized,

cumulative activities. Thus, from the sociocultural perspective, assessment indicates observing and analyzing how people use knowledge, skills, and processes to participate in the real work of a community.

Today, both the individual development of knowledge emphasized by the cognitive approach and the social practices of learning emphasized by the sociocultural approach are considered important aspects of education (NRC, 2001, p. 64). Assessment practices from these perspectives do not focus on the individual skills and knowledge that characterize the earlier behaviorist perspectives. Instead, they encompass issues involving the organization and processing of knowledge, including participatory practices that support the knowing, understanding, and embedding of knowledge in social contexts (NRC, 2001, p. 65). Knowledge is often embedded in particular social and cultural contexts, including the context of the classroom, and encompasses understandings about the meaning of specific practices such as asking and answering questions. Thus, assessments should examine how well students engage in communicative practices that are appropriate to a particular domain of knowledge and skills, what they understand about those practices, and how well they use the tools appropriate to that domain (NRC, 2001, p. 92).

As NRC (2001) indicated, assessment from both the cognitive and sociocultural perspectives is emphasized in today's world of education, and there are studies that compare assessments from before and after the implementation of these perspectives. Shepard (2000) compared measurement theory from the behaviorist perspective and assessment theory from the cognitive and constructivist perspectives. Torrance and

Pryor (2001) compared assessments from the behaviorist and social constructivist perspectives as well. Also, Van den Heuvel-Panhuizen and Becker (2003) compared the psychometric model of assessment and the didactic model of assessment.

Shepard (2000) claimed that objective testing was preferred from the behaviorist perspective while classroom assessment was advocated according to the cognitive and constructivist perspectives. While the objective test more conveniently facilitates the scientific measurement of ability and achievement, the classroom assessment provides tasks to elicit higher order thinking and addresses learning processes as well as learning outcomes. Classroom assessment is an ongoing process that is integrated with instruction. In addition, it is used formatively in support of student learning. Brookhart (2003) presented three foundational principles of the classroom assessment: 1) in classroom assessment, there is a psychosocial context; 2) classroom assessment and instruction are integrated; and 3) classroom assessment is primarily formative (Brookhart, 2003, pp. 6-8). Integration incorporates the assessment into the environment itself to become a part of the fabric of students' learning. Brookhart stated that the process of classroom assessment itself is a part of learning, and with it both teachers and students become members of the assessment practice. Students often view classroom assessment as a demonstration of what they were "supposed to learn," and they recognize the language of the assessment as the language of instruction.

Torrance and Pryor (2001) named the behaviorist perspective of assessment "convergent assessment" and assessment from the social constructivist perspective

"divergent assessment." In convergent assessment, they ascribed the prevalent aim as the determination of whether or not the learner knows, understands, or can do a predetermined task. However, in divergent assessment, the prevalent aim is the discovery of what the learner knows, understands, and can do. They viewed convergent assessment as a repeated summative assessment or continuous assessment while divergent assessment they considered more close to contemporary theories of learning and as formative assessment (Torrance & Pryor, 2001, p. 617). A detailed description of each assessment is provided in Table 5.

Van den Heuvel-Panhuizen and Becker (2003) compared the psychometric model of assessment and the didactic model of assessment. From the psychometric approach, tests are developed under classical test theory or modern test theory design. Thus, the psychometric approach has non-ambiguous characteristics and assumes that every problem has one correct answer and the answer can always be identified without question (Van den Heuvel-Panhuizen & Becker, 2003, p. 702). However, this approach does not properly connect instructional aim and assessment. Van den Heuvel-Panhuizen and Becker claimed that subject matter experts are necessary than statisticians for the classroom assessment. As an alternative, researchers proposed the didactical model of assessment. They emphasized that assessment should play an integral role in teaching and learning, and that instruction and assessment should be epistemologically consistent. Assessment should support and align with the teaching and learning process, as well as emphasize teachers' assessment abilities.

Table 5 *Convergent and divergent assessment (Torrance & Pryor, 2001, p. 617)*

	Convergent Assessment	Divergent Assessment
Aims	To discover if the learner knows, understands or can do a predetermined thing	To discover what the learner knows, understands or can do
Practical Implications	- Precise planning and an intention to stick to it - Tick lists and can-do statements - An analysis of the interaction of the learner and the curriculum from the point of view of the curriculum - Closed or pseudo-open questioning and tasks - A focus on contrasting errors with correct responses - Judgmental or quantitative evaluation - Involvement of the student as recipient of assessments	- Flexible planning or complex planning which incorporates alternatives - Open forms of recording - An analysis of the interaction of the learner and the curriculum from the point of view both of the learner and of the curriculum - Open questioning and tasks - A focus on miscues - aspects of learner's work which yield insights into their current understanding, and on prompting metacognition - Descriptive rather than purely judgmental evaluation - Involvement of the student as initiator of assessments as well as recipient
Theoretical Implications	- A behaviorist view of learning - An intention to teach or assess the next predetermined thing in a linear progression - A view of assessment as accomplished by the teacher	- A social constructivist view of learning - An intention to teach in the zone of proximal development - A view of assessment as accomplished jointly by the teacher and the student

Even though views on assessment requirements have changed along with those of instructional perspectives, the changes in real classrooms have not kept pace. Appropriate curriculum and scientific measurements required from the behaviorist perspective should be moved to proper curriculum and classroom assessments from

the cognitive and constructivist perspectives (Shepard, 2000). However, instruction drawn from the emergent paradigm and testing drawn from the traditional paradigm continue to coexist. Shepard (2000) illustrated this phenomenon as Figure 3.

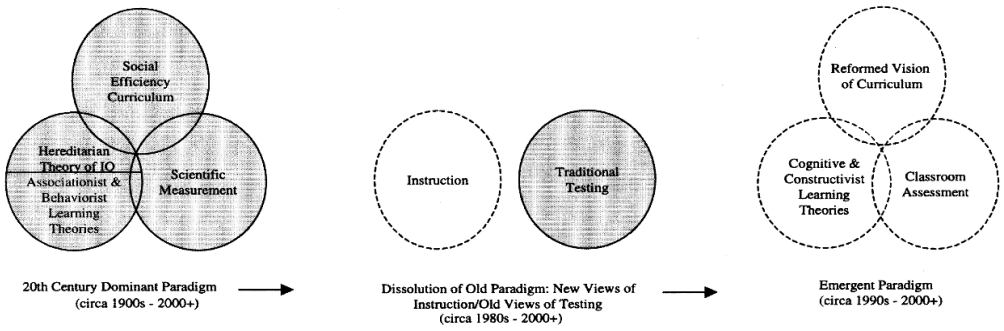


Figure 3 Historical overview explain the current incompatibility between instruction and testing (Shepard, 2000, p. 5)

In summary, as the instructional perspective has changed, required assessment has changed as well; for example, from objective testing to the classroom assessment and from convergent assessment to divergent assessment. However, required assessments have not been implemented in the classroom; thus, new views of instruction paired traditional views of testing coexist in a single classroom (Shepard, 2000). In the case of informal statistical inference, it occurs in a teaching and learning situation in which communication based on argumentation appears. However, it suffers the similar problem as Shepard (2000) mentioned in that statistics assessments continue to follow traditional ways of testing. Students should use related statistical concepts in the process of informal statistical inference, and they need to participate in the discourse with established norms when they constructing and verifying

argumentations. Thus, informal statistical inference requires assessments designed from the cognitive and sociocultural perspectives. For example, NRC (2001) provided the following case concerning a mathematics assessment from the sociocultural perspective.

> For example, to assess performance in mathematics, one might look at how productively students find and use information resources; how clearly they formulate and support arguments and hypotheses; how well they initiate, explain, and discuss in a group; and whether they apply their conceptual knowledge and skills according to the standards of the discipline. (NRC, 2001, p. 64)

To assess students' abilities at informal statistical inference from the cognitive and sociocultural perspectives, teachers can observe how students use related statistical concepts in the process of informal statistical inference; how they construct argumentations based on data and context; and how they verify their argumentation in peer groups by using probabilistic language, recognizing the importance of repeated sampling, and establishing norms for dealing with uncertainty.

4.1.2. Meaning of Integration of Instruction and Assessment

Assessment can be distinguished as "assessment of learning" and "assessment for learning," according to its purpose (Wiliam, 2007; Bennett, 2011). The purpose of

assessment of learning is to certify students' level of competence, and the purpose of assessment for learning is to support students' learning. Section 4.1.1 acknowledged that as the instructional perspective changed, the emphasized assessment would become the classroom assessment or divergent assessment. Because these assessments focus on learners and learners' improvement in learning, they align closely with the assessment for learning. Klenowski (2009) presented the definition of assessment for learning in the International Conference on Assessment for Learning recently.

> Assessment for learning is part of everyday practice by students, teachers, and peers that seeks, reflects upon, and responds to information from dialogue, demonstration, and observation in ways that enhance ongoing learning. (Klenowski, 2009, p. 264)

Swaffield (2011) claimed that assessment for learning mainly concerns with the present and near future and with the learners and teacher directly involved with the present activity. Also, assessment for learning is in itself a learning process. Harlen (2006) claimed that assessment for learning helped to check the next learning stage more than assess what has been accomplished. Assessment for learning takes place in the classroom lesson at the same time, and judgments of achievement are based on student references. As a result, both students and teachers receive feedback. Therefore, interaction between the teacher and students is critical.

By reviewing the changes in assessment perspectives according to changes in instructional perspectives in the former section and the characteristics of assessment for learning in the present section, it becomes apparent that both advocate the integration of instruction and assessment. Because informal statistical inference should occur in the integration of an instruction and assessment situation as well, it is necessary to clarify what an integration of instruction and assessment means.

Firstly, an integration of instruction and assessment involves a consensus of perspective between instructional theory and assessment (Shepard, 2000; Pellegrino et al., 1999). Shepard (2000) highlighted the incompatibility between new views of instruction and traditional views of testing, and thus the resulting inconsistency in such a practice. Reflecting the emphasis of instruction in an assessment, that is, to align an instructional aim with the assessment's aim, composes the basic condition for integrating instruction and assessment. Secondly, as have many researchers articulated about assessment for learning, the integration of instruction and assessment means that assessment occurs in the present or near future, is in itself a learning process, and takes place simultaneously in the classroom lesson (Swaffield, 2011; Harlen, 2006). In this case, the participants of the assessment include both the students and their teacher all of whom directly relate to the assessment activity.

Integration of instruction and assessment is appropriate for assessing from the cognitive and sociocultural perspectives. In the cognitive perspective, it is important to assess students' cognitive processes, such as how students structuralize and organize their knowledge (Webb, 1992; Black & Wiliam, 2009). From this

perspective, process ability is more emphasized than individual knowledge. Students need to show what they understand when they engage in problem solving, reasoning, and communicating. If a student's process ability is assessed separately from the instruction, there might be a possibility that another ability is actually being assessed rather than the intended one (Watson, 2000). Thus, it is necessary to accumulate information on what and how much students understand by integrating the instruction and assessment situations, and through this integration, student improvement can be monitored.

In the sociocultural perspective, students are required to engage in particular forms of practice and community. Thus, an assessment from this perspective would measure the degree to which the student can participate in a particular form of practice (NRC, 2001). It is necessary to examine how well students engage in communicative practices appropriate to a domain of knowledge and skill, what they understand about those practices, and how well they use the tools appropriate to that domain (NRC, 2001, p. 92). Since the sociocultural perspective focuses on participation and collective activity, instruction and assessment cannot be separated. Uchiyama (2004) claimed that because learning takes place in the social setting prior to being internalized by the individual, learning is mediated by language and discourse, and because teaching assists performance by scaffolding from the sociocultural perspective, it is necessary that assessment be conducted during instruction, thus calling for a situation of integrated of instruction and assessment.

In order to integrate instruction and assessment, one must examine the critical

elements of simultaneity and interaction between a teacher and students. Black and Wiliam (2009) claimed that formative assessment is concerned with the creation of and capitalization upon "moments of contingency" in instruction for the purpose of the regulation of learning processes (p. 10). It must be noted that these moments of contingency can be synchronous. Examples of synchronous moments include the teacher's real time adjustments during one-on-one teaching and in whole class discussion (Black & Wiliam, 2009, p.10). In addition, Black and Wiliam (2006) emphasized the interactive aspects between a teacher and students because interactive feedback is a critical feature in determining the quality of a learning activity. In a classroom discussion, the feedback will stand in relation to the needs of the subject-classroom as a whole and may work as an immediate intervention in the flow of classroom discussion (Black & Wiliam, 2009, p. 11). Black and Wiliam (2009) provided the concept of a formative interaction, which is an interaction between an external stimulus and feedback, and an internal production by the individual learner. This kind of interaction influences cognition. Three aspects considered in this situation are the external, the internal, and their interactions. Black and Wiliam (2009) illustrated three aspects as Figure 4.

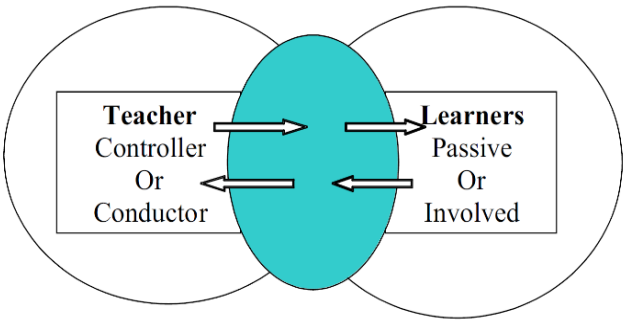


Figure 4 Three interacting domains of pedagogy (Black & Wiliam, 2009, p. 11)

When a teacher addresses the learner with a task, the learner responds to this, and the teacher then composes a further intervention, in the light of that response (Black & Wiliam, 2009). This basic structure can be described as initiation-response-evaluation or I-R-E. Black and Wiliam claimed that the model is meant to apply beyond one-on-one tutoring. The shaded area in Figure 4 stands for the classroom where many learners are involved through hearing the exchange, perhaps by joining in, so there would be many arrows in all directions in this area (Black & Wiliam, 2009, p. 11).

In the interaction between a teacher and students, their role is regulated as following. Black and Wiliam (2009) stated that teachers should build a model of how students learn to provide feedback to students. They said that the validity of such models constitutes a necessary condition for the effectiveness of the feedback (p. 13). Black & Wiliam claimed that teachers not only support students' learning but also support students to be better learners. That is, teachers emphasize learning how to learn (LHTL) and learning to learn (LTL) to support students to be better learners in

regard to competency improvement and the self-regulation learning while they give information related to learning contents for improvement in learning. Swaffield (2011) described students' roles as active participants in that they are participants in a learning process and accept feedback actively as well.

Wiliam (2007) drew the important instructional processes for assessment for learning based on Ramaprasad's definition of feedback. Ramaprasad (1983) defined feedback as the information about the gap between the actual level and the reference level of a system's parameter that is used to alter the gap in some way (p. 4). Three key instructional processes are as follows:

- Establishing where the learners are in their learning,
- Establishing where they are going,
- Establishing what needs to be done to get them there. (Wiliam, 2007, p. 1064)

Based on these processes, Wiliam presented the framework by assigning roles to the teacher, the learner, and the peers. Then he suggested five essential strategies to implement assessment effectively in the framework (Wiliam, 2007; Black & Wiliam, 2009; Wiliam & Thompson, 2008) that are illustrated in Table 6.

Table 6 *Aspects of assessment for learning (Wiliam, 2007, p. 1064)*

Where the learner is going		Where the learner is right now	How to get there
Teacher	Clarifying learning intentions and sharing and criteria for success	Engineering effective classroom discussions and tasks that elicit evidence of learning	Providing feedback that moves learners forward
Peer	Understanding and sharing learning intentions and criteria for success	Activating students as instructional resources for one another	
Learner	Understanding learning intentions and criteria for success	Activating students as the owners of their own learning	

In the framework, the role of the teacher can be summarized as one that clarifies and shares learning intentions and criteria for success, engineers effective classroom discussions, questions, and learning tasks that elicit evidence of learning, and provides feedback that moves learners forward (Wiliam, 2007, p. 1064).

In summary, for the successful integration of instruction and assessment, consensus should be reached regarding each aspect's perspectives and aims for each aspect need to have consistency. In this case, assessment is in itself a learning process with both a teacher and students acting as participants in the assessment. Integration of instruction and assessment is a necessary and sufficient condition for assessing from the cognitive and sociocultural perspectives. For integrating instruction and assessment, assessment should occur in whole class discussions simultaneously with the discussion, all the while interactive feedback should be emphasized. It is an interaction between a teacher and a whole class of students, so the teacher can give immediate interventions in the flow of classroom discussion and by this the interaction

can influence students' cognition. The process can be described as initiation-response-evaluation. A teacher should implement an assessment to make students improve from their actual levels to reference levels. Several roles of teacher were described. In the case of informal statistical inference, where communication based on argumentation is required, assessment from the cognitive and sociocultural perspectives is necessary. Thus, it is important to integrate instruction and assessment in practicing assessment. The roles of both a teacher and students in the interaction during the class discussion provided in this chapter should be noted.

4.1.3. Models of Integration of Instruction and Assessment

4.1.3.1. General assessment model

Assessment has several components or stages, including establishing the aims of assessment, designing a task, presenting a task to students, and interpreting their responses. NRC (2001) provided the assessment triangle model to aid in the implementation of an effective assessment. NRC (2001) claimed that an assessment is a tool designed to observe students' behavior and produce data that can be used to draw reasonable inferences about what students know (p. 42). They named the process of collecting evidence to support the types of inferences one wants to draw reasoning from evidence. By subdividing the process of reasoning from the evidence into its essential elements, they composed the assessment triangle model. The three elements in the model are cognition, observation, and interpretation as illustrated in Figure 5.

The model has been applied in various areas according to various purposes, such as analyzing current assessments and designing future assessments. In the following, the detailed aspects of the assessment triangle model will be examined, based on the research of NRC (2001, pp. 44-51).

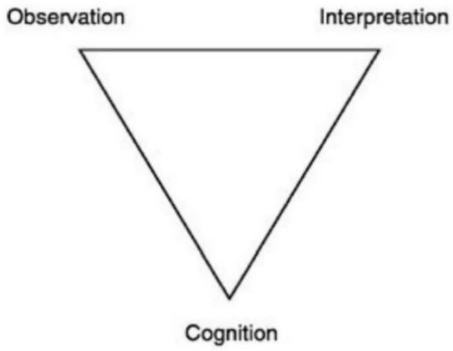


Figure 5 The assessment triangle model by NRC (2001, p. 44)

The three elements of the assessment model include: a model of student cognition and learning in the domain, a set of beliefs about the kinds of observations that will provide evidence of students' competencies, and an interpretation process for making sense of the evidence (NRC, 2001, p. 44). Researchers stated that an assessment could not be designed and implemented without taking into some degree of consideration each of the three elements. Each element is connected to and dependent on the other two, which is the reason for the triangle shape with each element resting in a corner of the triangle. NRC (2001) claimed that for an assessment to be effective, the three elements must act in synchrony.

The first element, cognition refers to a theory or set of beliefs about how students

represent knowledge and develop competence in a subject domain (NRC, 2001, p. 44). To assess students' competencies within a certain domain, a theory of learning in the domain is required to identify the set of knowledge and skills that are important to measure. Assessment would be effective through the teacher's clear and explicit engagement with a conceptualized cognitive model of learning. The model can include several levels and types of knowledge representation such as social and contextual components because it can be based on several assessment perspectives. In addition, the model concerns tendencies in behavior, conceptions of phenomena, available strategies, or levels of development, all of which could be expressed in terms of numbers, categories, or some mix. If the cognitive model is described in detail, teachers can diagnose particular difficulties that students have. Any model of learning underlying an assessment necessarily will be a simplification of what is going on in the mind of the students and in the social situation within which the assessment takes place. The point of basing an assessment on a cognitive model is to focus the assessment on those competencies that are most important to measure in regard to the desired inferences about student learning.

The second element, observation represents a description or set of specifications for assessment tasks that will elicit illuminating responses from students (NRC, 2001, p. 48). The tasks should be designed carefully to provide evidence that is linked to the cognitive model of learning. In addition, they should support the kinds of inferences and decisions that will be based on the assessment results. For example, in case of a one-to-one situation, the observation framework would describe what the learner says

and does, does not say and do, or says or does with specific kinds of support or scaffolding (NRC, 2001, p. 48). The tasks for observation should be developed with the purpose of the assessment.

The third element, interpretation includes all the methods and tools used to reason from observation. Every assessment is based on certain assumptions and models for interpreting the evidence collected from observations (NRC, 2001, p. 48). Interpretation represents how the observations derived from assessment tasks become evidence about the knowledge and skills being assessed. In the context of the classroom assessment, the interpretation is often made less formally by the teacher and is typically based on an intuitive or qualitative model rather than a formal statistical one (NRC, 2001, p. 49).

NRC (2001) stated that each of the three elements of the assessment triangle must connect to each of the other two elements in a meaningful way to lead to an effective assessment and thus to sound inferences (p. 51). For example, a cognitive theory of how people develop competence in a particular domain provides clues about the types of situations that will elicit evidence about that competence and about the types of interpretation methods that are appropriate for transforming the data about student performance into assessment results. In the case of connections between observation and interpretation, knowing the possibilities and limitations of various interpretation helps in designing tasks for observations that is effective and efficient. The interpretation model expresses how the observations from a given task constitute evidence about the performance being assessed as it bears on the targeted knowledge

(NRC, 2001, p. 51).

The assessment triangle model provided by NRC (2001) can serve as a basis for designing a particular assessment model or a framework to analyze the assessment or assessment tasks. Because the model consists of three comprehensive components as a necessary condition for designing an effective assessment, there are studies that have constructed more particular assessment models based on the assessment triangle model. The Berkeley Evaluation and Assessment Research (BEAR) assessment system presented four principles for the valid assessment: a developmental perspective; a match between instruction and assessment; the generating of high-quality evidence; and management by instructors to allow appropriate feedback, feed-forward, and follow-up (Wilson & Carstensen, 2007). When researchers linked these principles with the assessment triangle model, they found that cognition corresponds to the developmental perspective, observation to the match between instruction and assessment, and interpretation to the high-quality evidence and management by teachers. Lyon (2011) used the assessment triangle model as a framework to analyze a teacher's classroom assessment. He applied the model to individual and group problem solving to analyze the relation between the assessment beliefs, practices, and reflection of the teacher. Ekmekci (2013) constructed a conceptual framework by applying the assessment triangle model to the design of PISA assessment items. Based on the assumption that each component of the model is interconnected, the researcher analyzed the items to see the relation between the PISA assessment framework and the resulting interpretation.

4.1.3.2. Assessment model based on interaction

The integration of instruction and assessment necessitates that the aims for instruction and assessment should be consistent, that assessment is in itself a learning process, and that it occurs in the whole classroom lesson situation simultaneously. Participants of assessment include both a teacher and students, and the interaction between them is significantvital. By integrating instruction and assessment, students' cognitive processing abilities can be measured from the cognitive perspective (Black & Wiliam, 2009). Also, how well students engage in communicative practices, what they understand about those practices, and how well they use the tools appropriate to that domain can be examined from the sociocultural perspective (NRC, 2001). In the situation of the integration of instruction and assessment, teachers examine students' cognitive abilities and the degree of participation in a form of practice and give them appropriate feedback through interaction between the teacher and students during the process of instruction, thus implying that the results of the assessment are immediately connected to instruction.

In the situation of the integration of instruction and assessment, teachers take a greater role because they need to provide feedback immediately during the interaction. Tasks become very important. In Van den Heuvel-Panhuizen and Becker's (2003) didactic model for assessment, they proposed the conditions for an assessment as follows: it should involve tasks having multiple solutions; it should involve dependent problems such that students can use the solution of the first part of the problem to solve the second part; and it should require the application of strategy. Researchers

emphasized that the didactic model should support the instruction process, and teachers should implement assessments that required teachers' assessment ability.

There are important elements for implementing a valid assessment in the situation of integrating instruction and assessment. In this section, assessment models based on interaction will be reviewed to investigate particular elements that were articulated in the previous studies that focus on the integration of instruction and assessment. In particular the analysis will focus on how teachers understand students' responses and provide feedback during interactions. Because the integration of instruction and assessment was found to hold a similar meaning as formative assessment, the formative assessment model also will be investigated.

Torrance and Pryor (2001) developed a formative assessment model for practice by integrating the descriptive and analytic frameworks for classroom assessment, which was developed in their previous research project, entitled Teacher Assessment at Key Stage (TASK) with the results of teachers' practices that generated descriptions and interpretations about their assessment based on the framework. The model situated in an intersubjective, social process and achieved through the interaction between teachers and students. Through its placement of clear criteria at the core of classroom practice and through the interaction of questioning, observation, and feedback, the model was constructed holistically and dynamically. The model is illustrated in Figure 6.

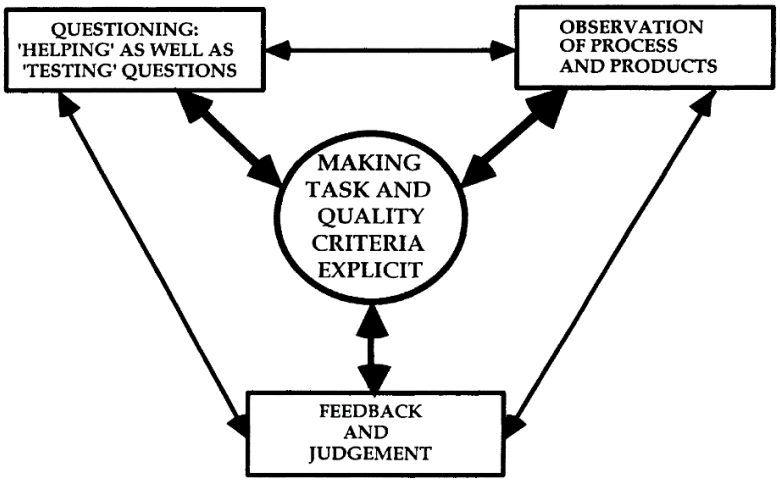


Figure 6 Assessment model by Torrance & Pryor (2001, p. 623)

In the model, the first element involves the explicit articulation of criteria. These criteria reside in two components: task criteria and quality criteria. Task criteria relate to both the particular task at hand and the overall "social rules" of the classroom, which refers to the teacher's expectations of behavior and the implicit rule-following (Torrance & Pryor, 2001, p. 623). Making task criteria explicit is important in establishing the context for a classroom assessment. Attention to quality criteria is also important. Researchers suggested that the expression of quality criteria could be accomplished through the teacher's interactions with individual students through questioning. The second element in the model is questioning. There are two kinds of questions: "helping" questions and "testing" questions. Different forms of questioning need to be used in various ways. Sometimes questioning can be seen as problematic because students can regard it as an indication of a wrong answer. Thus the cultivation of a situation in which children regard questions as opportunities for reflecting and

growing their thinking is essential. The third element, observation is the fundamental way in which teachers can obtain information about what children know, understand, and can do (Torrance & Pryor, 2001, p. 624). Observation of students' changes after questioning is considered even more important than observation of their initial stage, i.e., before the interaction. The last element, feedback, is intended to identify clearly to the extent the task has been completed and the extent and ways its quality could be improved (Torrance & Pryor, 2001, p. 628). For example, if students have a conversation about their accomplishments, quality would be enhanced.

Cowie and Bell (1999) presented a model of formative assessment in the science classroom. They regarded formative assessment as the process used by teachers and students to recognize and respond to student learning to enhance that learning while the learning is happening (Cowie & Bell, 1999, p. 101). Researchers found that teachers used two kinds of formative assessment: planned and interactive formative assessment. Planned formative assessment involved the teacher eliciting assessment information through planned, specific assessment activities and from there, interpreting and taking action on the information (Cowie & Bell, 1999, p. 114). This type of assessment tended to be used with the whole class. Interactive formative assessment involved the teacher noticing students' thinking and processing in the context of the learning activities, recognizing, and then responding (Cowie & Bell, 1999, p. 114). This type of assessment tended to be used with individual students or small groups. The comparison between these two kinds of formative assessment is described in Table 7.

Table 7 *Planned and interactive formative assessment (Cowie & Bell, 1999, p. 114)*

Planned formative assessment	Interactive formative assessment
The parts of the process are eliciting, interpreting, and acting	The parts of the process are noticing, recognizing, and responding
Tended to be carried out with all the students in the class	Tended to be carried out with some individual students or small groups
Could occur over an extended time frame	Happened over a short time frame
Purposes are mainly science-referenced	Purposes are science-, student-, and care-referenced
Responsive to 'getting through the curriculum'	Responsive to student learning
What is assessed is mainly science	What is assessed is science, personal, and social learning
The assessment information obtained is product and process	The assessment information obtained is product and process but ephemeral
Interpretations are norm-, science-, and student-referenced	Recognizing are science-, norm-, and student- referenced
Actions are science-, student-, and care-referenced	Responses are science-, student-, and care-referenced
Relied on teachers' professional knowledge	Relied on teachers' professional knowledge

The main distinction between the two kinds of formative assessment is the degree and type of planning carried out by the teachers (Cowie & Bell, 1999, p. 114). Since the purpose of planned formative assessment is to obtain information from the whole class about progress in the learning of specific disciplines, it is planned in advance how to elicit students' understandings and skills. On the other hand, since the purpose of interactive formative assessment is to mediate in the learning of individual students with respect to a wider range of learning outcomes, such as the learning of a discipline, social learning, and personal learning, it requires more immediate actions to practice. Teachers can attain information from students' verbal and non-verbal language. Verbal language includes students' comments and questions, and non-verbal language

includes how they performed in practical activities, how they interacted with others, the tone of discussions, and their body language (Cowie & Bell, 1999, p. 108). Both of assessments should be contextualized, responsive, on-going, and carried out during instruction to improve learning.

Cowie and Bell (1999) claimed that the two kinds of formative assessment could be connected through the purpose of formative assessment. Figure 7 below illustrates this connection. Some teachers moved from planned formative assessment to interactive formative assessment and back. In most cases, this kind of movement occurred when a teacher noticed something, such as students' alternative conceptions or misconceptions. The focus of this movement shifted from the whole class to an individual.

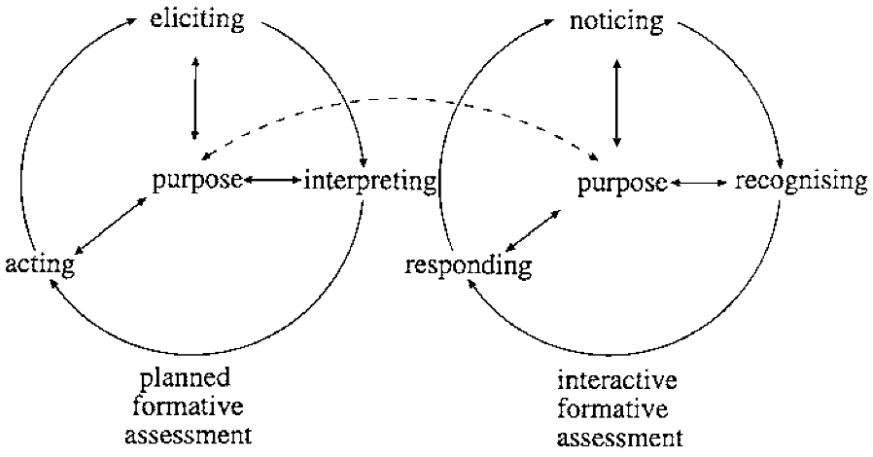


Figure 7 Assessment model by Cowie & Bell (1999, p. 113)

Herman (2013) proposed a formative assessment model based on previous studies

on assessment. Starting from the assessment triangle model provided by NRC (2001), the researcher added elements of assessment tools and processes that can support learning. The model, illustrated in Figure 8, aimed to provide a starting point for considering how to observe student learning relative to the intended goal and the design of tasks that simultaneously engage students in and can be used to elicit evidence of their learning relative to the goals. The task can be designed to promote students' learning based on the cognitive theory. The interpretation framework should be designed and applied to student responses or interactions to identify where students are relative to the goal and to provide diagnostic information for moving forward. These elements, the goal, the task, and the interpretation framework, parallel the three elements in the assessment triangle model. However, to apply the model as a formative assessment, the interpretation should be transformed by teachers to give students informative feedback, which students could subsequently act upon to help them reach the intended goal. The entire assessment process is embedded in a classroom culture and structures that support learning goals, collaboration, and accountability for learning (Herman, 2013, p. 16).

Herman (2013) advocated that in the context of formative processes, feedback occurs interactively, involving teachers, students, and peers. And it is connected to the next instruction. For example, in responding to a student's misconception during a classroom discussion, a teacher can ask other students to propose alternative answers and engage them in instructional interactions to coach them through cycles of scaffolding-response-feedback to enable students to reach a higher level of

understanding (Herman, 2013, pp. 16-17). The model illustrated in Figure 8 is a simplified version that shows just a single assessment. In the classroom, it would be applied in a system of assessment.

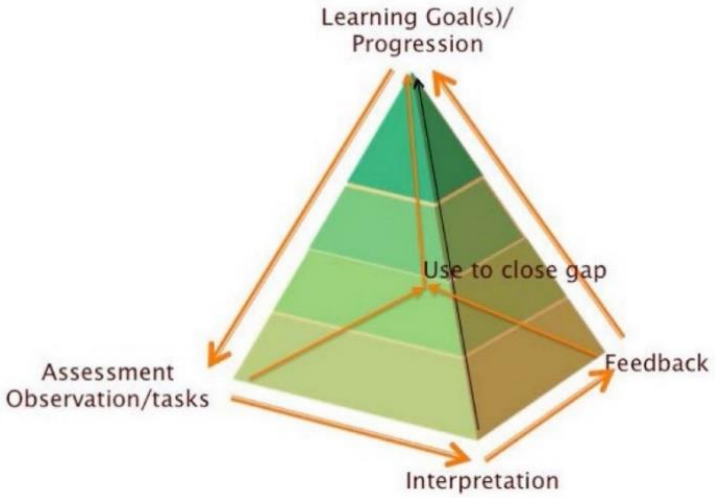


Figure 8 Assessment model by Herman (2013, p. 15)

In section 4.1.3., assessment models that include the general assessment triangle model (NRC, 2001) and several assessment models based on the interaction from the integration of instruction and assessment were investigated. By combining the results of that investigation, an implication for designing an assessment model for informal statistical inference can be derived as follows. First, from the assessment triangle model, the three components, cognition model for assessment, observation to induce students' understanding, and interpretation of students' responses emerge as very important. These components are also found in the assessment models based on interaction. In addition, feedback should be considered as a component of assessment

as well a part of the interaction. This component is named in Torrance and Pryor (2001) and Herman (2013) “feedback” and in Cowie and Bell (1999) “acting” or “responding.” In the assessment model of Torrance and Pryor (2001), four components including feedback are formed interactively and dynamically because their assessment model is based on the interaction between teachers and students. Herman (2013) also considered feedback an important component and included it in the assessment triangle model given by NRC (2001). Therefore, in assessments based on interactions, feedback should be regarded both as one of the components and also as one of the stages in the assessment procedure.

Cowie and Bell (1999) presented two kinds of formative assessment, planned formative assessment and interactive formative assessment. Planned formative assessment, based on assessing students’ understanding of a subject, tends to be implemented on the whole class. In the case of interactive formative assessment, individual student’s personal and social understanding is addressed; thus, it tends to be implemented on individual students. In both assessments, the teacher plays a leading role, for it is the teacher who is responsible for assessing students’ cognitive processes and their participation of activity either by planned formative assessment or by interactive formative assessment. Namely, teachers must assess students’ cognitive aspects and social aspects back and forth.

In the assessment based on interaction, there is a basic cycle of assessment that continuously repeats. Heritage (2010) provided a six-stage feedback cycle, which includes eliciting evidence of learning, interpreting the evidence, identifying gaps,

engaging in inquiry or feedback, learning or acting, and bridging new learning. There exists a basic unit of assessment like six-stage. Black and Wiliam (2009) presented the process of interaction between teacher and learner as: “teacher addresses to the learner a task, the learner responds to the task, and teacher compose a further intervention in the light of the response.” They described this basic structure as initiation-response-evaluation. Herman (2013) also presented a basic structure of assessment, even though assessment is implemented as a continuous system. In the assessment based on interaction, it is necessary to focus on that basic unit of assessment.

4.2. Assessment Model for Informal Statistical Inference

4.2.1. Design of Assessment Model

Based on the implications for designing an assessment model of informal statistical inference in chapter 3, it is proposed that the assessment of informal statistical inference should be conducted through communication based on argumentation. And communication based on argumentation would be emerged in the teaching and learning process. In this situation, students’ use of statistical concepts and construction and verification of argumentation can be assessed actively. From the analysis of the general assessment triangle model and several assessment models based on the interaction that results from the integration of instruction and assessment,

it is induced that the essential components in the assessment models include the cognition model, observation, interpretation, and feedback. In addition, when instruction and assessment are integrated, assessment can be carried out based on the interaction between a teacher and students. Thus, a teacher should assess both the cognitive aspect and social aspect back and forth at the same time. Finally, assessment models can be constructed based on the basic assessment structure defined as initiation-response-evaluation.

Considering these implications, the assessment of informal statistical inference should be implemented through communicating based on argumentation and at the same time as the process of instruction. The components of assessment include the cognition model, observation, interpretation, and feedback. Since cognition corresponds to the ability to be assessed, it can be regarded as assessment elements. In addition, since observation corresponds to the task of examining students' understanding, it can be regarded as assessment tasks. Based on components and considering the procedure of assessment, the assessment model of informal statistical inference can be illustrated in a diagram as in Figure 9.

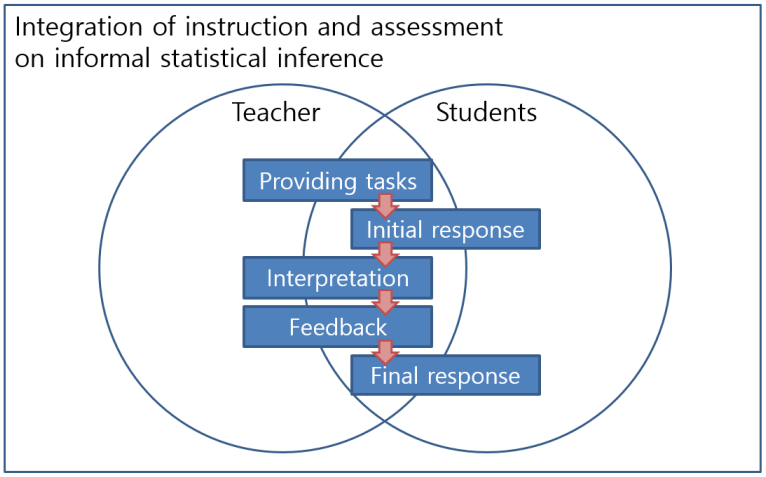


Figure 9 Assessment model of informal statistical inference

First, to present that the assessment of informal statistical inference is carried out based on the interaction between a teacher and students when instruction and assessment are integrated, the integration of instruction and assessment is represented in the diagram as a universal set with the teacher and students represented as two intersecting sets. Since assessment is conducted at this intersection, the procedure of assessment is represented in the intersection. Assessment begins when a teacher provides assessment tasks, and then students present initial responses. A teacher will interpret students' initial responses based on the assessment element. Even though both the assessment element and interpretation are components of assessment, it is solely interpretation that is the behavior of teacher and is included as a procedure because both of them occur at the same time in the assessment. A teacher provides feedback in light of his or her interpretation of the assessment elements, and students in turn will present their final response, which can be used to illustrate their

improvement. Figure 9 can be considered a basic assessment structure, and in a real assessment situation, the interactive process described on the diagram would be repeated continuously.

4.2.2. Characteristics of Components in Assessment Model

The assessment situation, which was presented as the universal set in Figure 9, will be explained in detail below. Among the five stages in the procedure, providing tasks, interpretation, and feedback have been selected because they related to the teacher's behavior. In section about providing the task, the characteristics of proper assessment tasks for informal statistical inference and some examples will be described. In the section about interpretation, the assessment elements that can be applied in assessing students' statistical concepts and argumentation in the process of informal statistical inference will be presented. In the section about feedback, the proper method for providing feedback in interaction of informal statistical inference will be discussed.

4.2.2.1. The assessment situation

The characteristics of informal statistical inference include that it is an informal argumentation using natural language, that it is based on context, and that it occurs within an interaction. Therefore, the realization of informal statistical inference demands a situation of teaching and learning processes in which communication

occurs based on argumentation using verbal language. To examine students' informal statistical inference abilities, assessments should be carried out at the same time as teaching and learning processes. As described in the meaning of the integration of instruction and assessment, such integration produces a good situation for the implementation of the assessment from both cognitive and sociocultural perspectives. In the classroom, an environment that encourages the use of argumentation is required. Thus, the norms necessary for conducting informal statistical inference should be based on argumentation and should be established as social norms too in the classroom. In this environment, effective communication and interaction between a teacher and students should occur.

4.2.2.2. Providing tasks

A task is one component in the assessment model and is necessary to convey students' understanding. To assess students' abilities to draw authentic informal statistical inferences, tasks reflecting the nature of informal statistical inference should be designed. The concepts and thinking that are essential in informal statistical inference have been investigated in the didactical analysis of informal statistical inference. Concepts that include expectation and variation, sample and population, a size of sample, and sampling distribution, and construction and verification of argumentation should be included in the tasks. Above all, as mentioned in the assessment situation, the task that facilitates student communication based on argumentation is necessary.

To induce proper tasks for assessing informal statistical inference, studies that have been conducted about the characteristics of tasks for statistics and informal statistical inference should be reviewed. Burrill (2007) cited openness as an important characteristic for formative assessment tasks in statistics. Open-ended tasks that allow discussion enable the teacher to provide feedback and students to present various responses. Also, open-ended tasks have the potential to reveal much information about students' conceptions and misconceptions. In addition, Burrill suggested that tasks needed to be approached by students at different levels of understanding. They should be used to develop students' conception, visualization, various solving strategies, critical thinking, connection with other conceptions, and gradual formation.

Reading (as cited in Zieffler et al., 2008, p. 46) claimed that tasks for informal statistical inference should be used to examine how students integrate components of informal statistical inference and to grasp their ideas concerning statistical inference. In addition, tasks should be designed to draw various argumentations. Based on Reading's suggestion, Zieffler et al. (2008) presented the types of tasks that can be used to demonstrate informal statistical inference as depicted in Table 8. Zieffler et al. connected the components of informal statistical inference, such as making judgments or predictions, using or integrating prior knowledge, and articulating evidence-based arguments, and the types of task to formulate nine forms of tasks.

Table 8 *Types of tasks in informal statistical inference (Zieffler et al., 2008, p. 52)*

Type of Task	Informal statistical inference component		
	Make judgments or predictions	Use or integrate prior knowledge	Articulate evidence-based arguments
Estimate and draw a population graph	Predict characteristics of a population (shape, center, spread) that are represented in a student-constructed graph	Bring in intuitive or previously learned knowledge and language to predict the characteristics of a population	Requires an explanation of how the characteristics of the population graph were chosen
Compare two samples of data	Judge whether there is a difference between two populations; based on similarities or differences in samples of data	Bring in intuitive or previously learned knowledge and language to compare two samples of data	Requires an explanation of why students determined whether or not there is a difference in the two populations
Judge between two competing models	Judge whether sample data provide more support for one model than another	Bring in intuitive or previously learned knowledge and language to judge between two competing models	Requires an explanation of why students chose one model over the competing model

Three of Zieffler et al.'s (2008) types of tasks presented in Table 8 should be examined. The first task type is to estimate and draw a population graph. It begins with investigating the sample, which is related to "growing sample tasks." "Growing sample tasks" has been greatly emphasized in many studies (Bakker & Gravemeijer, 2004; Ben-Zvi, 2006; Ben-Zvi & Aridor, 2012). Bakker and Gravemeijer (2004) stated that tasks related to growing a sample encourage students to reason about stable characteristics when sample grows. Ben-Zvi and Aridor (2012) claimed that the task of growing a sample allows students to connect the worlds of data and context, draw

a conclusion from the data, explain their informal inference, and support their conviction. Cobb and McClain (2004) claimed that according to Bakker and Gravemeijer's (2004) findings, repeating the process of data generation or growing sample activities are necessary to improve students' reasoning abilities concerning statistical inferences. Through the activities, students can ensure predicted characteristics of data sets stabilize as the size of a sample increases.

The second type of task involves the comparison of two samples of data. Comparing data sets activity encourages students to focus on statistical concepts and components (Watson & Moritz, 1999; Watson, 2008). Especially, when the comparing data sets are samples, it is directly related to make statistically significant decisions or infer about data and situations (Ciancetta, 2007). Also, since it provides the opportunity to draw a conclusion by statistical inference, the data set comparison activity provides a rich context that allows students to experience foundational concepts of inferential statistics (Makar & Confrey, 2002). When comparing data sets, tasks typically involve comparing two graphs such as bar graphs or boxplots. A boxplot is a graph that shows five statistics including the minimum value, maximum value, the first quartile, median, and the third quartile. Instead of representing data value specifically, a boxplot organizes data by dividing it in sets of the same number of value. It encourages the understanding of distribution by students (Bakker & Gravemeijer, 2004). In comparison contexts, boxplots can serve multiple functions, such as comparing the same statistics or different statistics. Therefore, it is applied frequently in studies.

The last type of task entails making a judgment between two competing models. In the process of drawing a formal statistical inference, when a certain sample is obtained, the calculation of p-value of the sample results and the decision about rejecting a hypothesis or not is based on the comparison of the sample result with a sampling distribution that formulated by infinite sampling. Students' abilities with informal statistical inferencing typically do not extend to calculating statistics rigidly. However, students can still communicate about whether the sample result is unusual or not or how much they can trust the result. Since these processes support the initial understanding of p-value (Garfield & Ben-Zvi, 2008), they are helpful for instructing formal statistical inferences later. Notably, these types of tasks are meaningful because they allow students to consider sampling distribution or variability including chance variability, thus enabling students to present argumentation and explanation (Zieffler et al., 2008).

By considering the types of tasks given by Zieffler et al. (2008) and the conditions for tasks suggested by Burrill (2007) and Reading (as cited in Zieffler et al., 2008, p. 46), tasks for facilitating informal statistical inference can be induced. Tasks may reflect the types of tasks given by Zieffler et al. as contexts of each problem and can be designed specifically by including concepts and thinking from the didactical analysis of informal statistical inference in each of their problem situations. Because the tasks will be applied in a classroom in which students communicate using argumentation, tasks should be open-ended to help students at differing levels present argumentation with various ideas and approach conceptions from multiple directions.

Table 9 shows the overview of the task that can be used in approximately six lessons of instruction. The task consists of three contexts: “to infer about population from one sample,” “to infer about population from two samples,” and “to infer based on the probability model.” In the case that students have learned the probability model, the complete contexts of the task can be used. If not, the first two contexts of task can be used. Tasks are designed based on the slight modification of tasks from previous studies (Zieffler et al., 2008; Shaughnessy, Ciancetta, Best, & Canada, 2004; Pfannkuch, 2007; Chance et al., 2004; Rossman, 2008) to ensure they are fitting for assessing informal statistical inference. The complete set of tasks is presented in the Appendix. Concepts that should be assessed from the tasks are relatively clear. In the context of “to infer about population from one sample” and “to infer about population from two samples,” concepts of expectation and variation, sample and population, and variability of the sample constitute the focus. In the context of “to infer based on the probability model,” the concepts of sampling distribution and sampling variability according to the size of a sample are the key points. Also, in the first two contexts, the task that advocates the notion of sampling distribution, i.e., sampling repeatedly to complement the uncertainty of induction, can be included though it does not address sampling distribution directly.

Table 9 *Overview of the 6-lesson tasks for assessment of informal statistical inference*

Lesson	Context and main concepts	Source
1	"Infer about population from one sample 1" Main concepts: descriptive statistics, sample and population, the size of sample	Zieffler et al. (2008)
2	"Infer about population from one sample 2" Main concepts: descriptive statistics, sample and population, the size of sample, the idea of sampling distribution	Saldanha & Thompson (2002)
3	"Infer about population from two samples 1 - comparison of bar graphs" Main concepts: descriptive statistics, sample and population, the size of sample and the variability of sample	Shaughnessy et al. (2004); Ciancetta (2007)
4	"Infer about population from two samples 2 - comparison of boxplots" Main concepts: descriptive statistics, sample and population, the size of sample and the variability of sample	Pfannkuch (2007); Zieffler et al. (2008)
5	"Infer based on the probability model 1" Main concepts: sampling distribution, descriptive statistics, the size of sample and sampling variability	CAOS, Chance et al. (2004)
6	"Infer based on the probability model 2" Main concepts: sampling distribution, descriptive statistics, the size of sample and sampling variability	Rossman (2008)

Examples of tasks for assessing students' argumentation will be presented below. Included is a task that requires the construction of argumentation from the data and a task that requires verification of argumentation by focusing on uncertainty. Figure 10 shows the task of the former type. From the comparison of two boxplots of sample, students should construct an argumentation. The task asks participants to infer a conclusion based on the sample. By considering the properties and patterns of samples and the context, students should be able to deduce a probable conclusion and the ground.

S company and K company, which are cell phone companies, provide text message services. From S company, a user can send 500 text messages for 10,000 (won) per month, and from K company, a user pays for each text message at a cost of 20 (won) each. Jinyoung wants to compare the number of text messages that users from each company have sent. She surveyed 100 people from each company and represented the results using boxplots. What can she infer from this data? Explain the reasons, giving as many as possible.

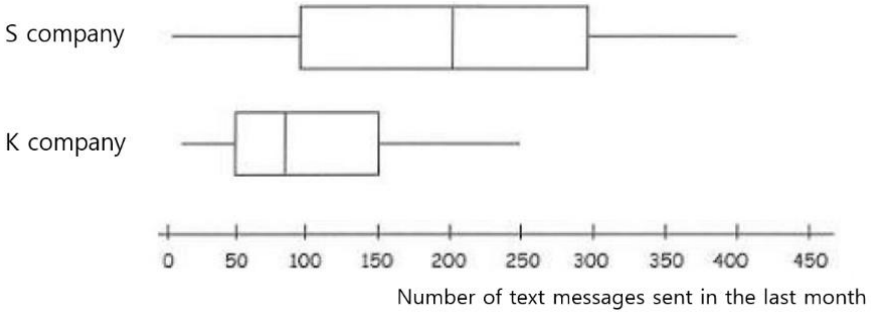


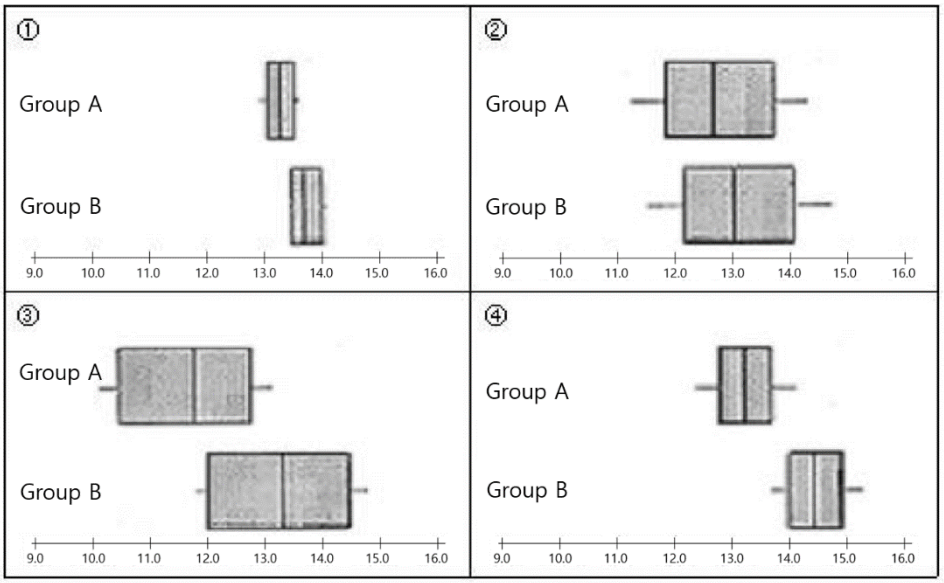
Figure 10 A task for constructing argumentation (Pfannkuch, 2007, p. 157)

Figure 11 depicts a task of the latter type, one requiring a verification of argumentation by focusing on uncertainty. The task randomly classifies athletes into two groups and provides group A with an additional program. The graphs show the difference between the two groups. Students should rank the differences according to how convincing each is in making an argument that the additional program is effective, which means that the difference between two graphs are significant. Each pair of graphs is different in view of mean as well as spread. Therefore, students should be able to compare them in regard to the distribution of each graph. To explain the reasoning behind a ranking is related to justification since this situation includes uncertainty, which also suggests the possibility of various answers for this task. Thus, students' argumentation should include an expression of probability language. Also, agreement procedures between students' answers are required, and students should

establish norms regarding the validation of argumentations.

Suppose that there is a special summer camp for track athletes. There is one group of 100 athletes that run a particular race, and they are all pretty similar in height, weight, and strength. They are randomly assigned to either group A or group B. Group A gets an additional weight-training program. Group B gets a regular training program without weights. All the students from both groups run the race before and after the camp, and their times are recorded, so that the data could be used to compare the effectiveness of the two training programs.

- Presented below are some possible graphs that show boxplots for different scenarios where the running times are compared for the students in the two groups. Rank the four pairs of graphs on how convincing they are in making an argument that the weight-training program was more effective in decreasing athletes' times (from the least convincing to the most convincing evidence). Explain your reasoning.



- For the pair of graphs that provide the most convincing evidence, would you be willing to generalize the effects of the training programs to all similar athletes on track teams based on these samples? Why or why not?

Figure 11 A task for verifying argumentation (Zieffler et al., 2008, p. 49)

4.2.2.3. Interpretation

Interpretation is the process of understanding evidence from given tasks. From students' responses regarding the assessment tasks, teachers should interpret both the cognitive aspect, that is the conceptual understanding and argumentation, and the sociocultural aspect, that is interaction based on argumentation. Interpretation can be conducted based on assessment elements. As presented in chapter 3, assessment elements can be induced from the essential concepts include descriptive statistics as expectation and variation, sample, the size of a sample, sampling distribution and from the essential thinking include the construction of argumentation and the verification of argumentation. Assessment elements relate to how students understand each concept and how they treat argumentation and overcome uncertainty.

Assessing students' conceptual understanding entails examining how students use each concept in the context of informal statistical inference rather than whether they know the meaning of concepts or not explicitly. When teachers examine students' use of each concept, they should make sure to take note of how students recognize the concept according to a framework for understanding. Researchers usually present frameworks of understanding levels in research. Based on the didactical analysis of informal statistical inference, the aspects of each concept that should be assessed are described below.

Firstly, descriptive statistics concepts such as expectation and variation can be interpreted based on the characteristics given Table 4 (p. 49) adapted from research (Watson et al., 2007). Under level 3, students have an incomplete understanding, that

they cannot understand the meaning of each concept or the interaction between them. In the context of informal statistical inference, students need to understand expectation and variation as a "center" or a "periodical tendency" and as "small change" or "random," respectively, and interaction between them. Such an understanding first appears in level 4. In levels 5 and 6, students understand expectation and variation as "proportion" and "unexpected change," respectively. This understanding is important because expectation and variation are related to distribution. The difference between levels 5 and 6 involves whether students can understand the interaction in a single context or across the various contexts. Also, an essential element in understanding is that these concepts should be considered criterions of comparison in inference and considered in the distribution to show the difference of concepts (Bakker et al., 2008).

Secondly, for a sample and population, students should understand the quasi-proportionality between them, which refers to the balance between sample representativeness and sample variability (Saldanha & Thompson, 2002; Rubin et al., 1991). To summarize the findings of previous studies, the understanding level of sample and population can be described as follows: consider a sample and population identically; consider sample representativeness only; consider sample variability only; understand the relation between sample and population according to proportional reasoning; balance between sample representativeness and sample variability (Watson et al., 2007; Ben-Zvi & Aridor, 2012; Saldanha & Thompson, 2002).

Thirdly, for the size of sample, it is necessary to understand that as the size

increases, the sample tends to explain the greatest part of the population, which is sample representativeness. For example, in the "growing samples" activity, students need to recognize what the stable characteristics are when the size increases (Rubin et al., 1991; Bakker, 2004; Ben-Zvi, 2006). The size of sample should be understood in two ways according to the context. For one sample as a part of a population, a large size of sample tends to depend on the sample's representativeness and a small size of sample tends to depend on the sample variability. For samples derived from repeated sampling from a population, if the size of sample is increased, then the sampling variability of sampling distribution tends to move closes to the normal distribution that has a smaller standard variation (Chance et al., 2004; Saldanha & Thompson, 2002). The former understanding can be related to the sample concepts described above, and the latter understanding can be related to the sampling distribution described below. When a teacher interprets students' responses, they must first consider the context.

Fourthly, for sampling distribution, it is important for students to know its meaning and the necessity of it at the primitive level in informal statistical inference. Students should be able to predict the sampling variability according to the size of sample and apply it in the inference (Garfield & Ben-Zvi, 2008; Pfannkuch, 2005; Sedlmeier & Gigerenzer, 1997). Related to the informal statistical inference, the following components for sampling distribution should be understood (Lipson, 2002; Saldanha & Thompson, 2002; Garfield & Ben-Zvi, 2008).

- Distinguish the distribution of a sample itself and the distribution of the statistics of sample. Sampling distribution is a distribution of the statistics of sample.
- Understand that a sampling distribution is a product of infinite sampling. Even though in the empirical approach, it is a distribution from the result of a repeated sampling trial.
- Sampling distribution is the basis of locating a given sample through a statistical inference. That is, the fact that a sampling distribution is itself a distribution and the necessity of it is important.

In the case of the thinking involved in informal statistical inference, interpretation should be based on the argumentation that students present. It can be interpreted in the two aspects: the construction of argumentation and the verification of argumentation.

When students induce a hypothesis that represents data well, they have to take into consideration the characteristics and patterns of data and the context to draw a conclusion. They then need to construct an argumentation based on the grounds between the data and conclusion. Therefore, to interpret the construction of an argumentation, teachers can examine data, conclusions, and grounds, which is inference rules, in argumentation. From here, they can interpret logical connections have been made between each of them. Teachers should focus most notably on the hypothesis and its level of probability in consideration of the data and context.

While an interpretation of construction of argumentation entails an examination of the logical connection between each of data, conclusion, and grounds which are presented explicitly, an interpretation of verification of argumentation more closely relates to an assessment of the understanding of uncertainty inherent in statistical inference. Teachers should interpret it based on how students approach the uncertainty. Students can use probabilistic representations, draw a conclusion by recognizing the importance of repeated sampling, or attempt to validate the argumentation by establishing norms for dealing with uncertainty during communication. For probabilistic representations, teachers should discern quantitative language and language that contains probable words from students' responses. Also, an expression of the degree of belief regarding their conclusion is also included in the probabilistic representation. For the recognition of importance of repeated sampling, students may present doubt about the uncertainty. They can show the difficulty in drawing conclusions from one sample and show recognition of the necessity to search for alternative methods. For the establishment of the norms, while students may criticize argumentations that contain probability and refute them, they may also make an agreement about the verification of a claim.

Interpreting students' construction and verification of argumentation is a challenging task because it must be assessed immediately in the interaction between a teacher and students. A teacher can start from the examination of data, conclusion, and grounds and then question other students about their peer's argumentation to expose differences and incite validity. To encourage students to discuss which data or

conclusion is more valid is also a way of interpreting responses.

4.2.2.4. Feedback

In statistical reasoning, it is important to consider which aspects students focus upon in a given data set, which would serve as the basis of designing instruction or determining the level of students' reasoning (Bakker, 2004; Reading & Reid, 2006; Pfannkuch, 2006). When Bakker (2004) designed the hypothetical learning trajectory of distribution, he decided to shift the focus from the center and spread to the overall shape of distribution. Reading and Reid (2006) constructed a framework for the level of reasoning by elaborately considering the treatment of important components in the reasoning of distribution. In relation with informal statistical inference, Pfannkuch (2006) extracted ten elements of reasoning that teachers paid attention to when they compared boxplots using informal inferential reasoning.

To implement an assessment from the view of emphasizing particular components on which students should direct their focus in the deducing of an informal statistical inference, it is necessary for the teacher to guide students' focus toward the appropriate target through the provision of feedback or questions. To do this, "attention," which is the concept given by Watson (2007), can be considered. Watson (2007) emphasized "attention" as a way of facilitating students' mathematical development through teachers' actions. Orientations or directions of attention are set up by the teacher performing certain acts, expressing certain ideas, initiating discussion about certain aspects, or asking learners to undertake certain acts. By

actions, she meant the ways in which individuals direct changes in objects, whether these are abstract ideas, visible things, or symbolic constructions (p. 118). Watson expected the shift of orientation towards concepts, methods, properties, and relationships to occur as a result of directing learners' attention. If the concept of attention is applied in the feedback for students' informal statistical inference, it is necessary to see what aspects students focus on before and after.

Because the assessment is based on the interaction between a teacher and students in this study, how the teacher connects students' responses is important. Cobb and McClain (2004) and Garfield and Ben-Zvi (2009) characterized the role of the teacher in developing statistical discourse in the following ways.

- Know the various approaches that students infer about data
- Predict the important statistical concept that can be a topic of discussion
- Use open-ended questions to facilitate students' thinking
- Require the verification of an inference. Ask other students whether they agree or not with reasons
- Establish classroom norms for representing students' perspectives comfortably

Ben-Zvi and Aridor (2012) emphasized the intervention of a teacher and the situation of social interaction between students as a place of improvement in informal statistical inference. Social interactions between students can contribute to the

development of the inference by presenting each student's different perspective. The role of teacher in such an intervention is as follows:

- "Predicting questions" can provide students an opportunity to imagine the properties of a population based on data and context.
- The situation that has a conflict between data and context is a catalyst to improve informal statistical inference. Teachers can provide evidence based on data and abduction explication continuously to improve students' predictions.

The "attention" concept of Watson (2007) combined with the role of teachers suggested from Cobb and McClain (2004), Garfield and Ben-Zvi (2009), and Ben-Zvi and Aridor (2012) offers a strategy that can be applied to teachers' feedback when it integrates teaching and learning in regard to informal statistical inference. Using these strategies of feedback, teachers can indicate the more advanced levels of object, concept, and uncertainty, levels greater than the students' present level by using understanding level framework that used in interpretation. Also, by considering data, conclusions, and grounds in argumentation, teachers can identify any parts of a student response that lacks validity.

CHAPTER 5

SUMMARY AND CONCLUSION

Statistics has its unique characteristics and styles of reasoning, when compared to mathematics. It is a methodological discipline basis on an indeterministic epistemology based in the variability of data. However, in the teaching of statistics in schools, it has been considered as a sub-domain of mathematics with students' learning expectations concentrated on the procedure of calculating means or confidence intervals through the application of algorithms and mechanisms. This contradicts the aim of statistics education and also is not appropriate in terms of assessing students' abilities in statistics. There has been a need of assessment methods that reflect and align with the characteristics of statistics, but substantive discussion about such assessments are still lacking. Thus, this study identified the nature of informal statistical inference and proposed appropriate assessment methods for measuring students' abilities regarding informal statistical inference.

Through an epistemological analysis of statistical inference, it is found that a statistical inference consists of two thinking components, abduction and induction. Statistical inference as induction can regulate its inherent characteristic according to how it deals with the uncertainty. To tackle this issue, statistical inference uses the quantification of uncertainty using probability. In addition, to validate the conclusion drawn from relatively limited experiences, falsification through *modus tollens* is employed in statistical inference. While the discussion on statistical inference as

induction is focused on how to address uncertainty in the process of inference to verify the conclusion, statistical inference as abduction is introduced to denote the importance of generating the simplest and most likely explanation of a hypothesis based on the characteristics and patterns of the sample and the context. Both induction and abduction serve as a component of thinking in regard to statistical inference and need to be recognized as different stages.

By reviewing research on informal statistical inference didactically, the meaning of informal statistical inference and the treatment of essential concepts and thinking were examined. The essential concepts include the expectation and variation concept, the sample and population concept, the variability of sample concept, and the sampling distribution concept. The treatment of each concept in informal statistical inference were examined, and the particular aspects necessary for students to understand each concept were decided. The essential thinking includes abduction and induction, and how these thinking components are treated in informal statistical inference was examined. Informal statistical inference should be carried out based on argumentation due to its characteristics, such as an informal argumentation using natural language, its basis in context, and its tendency to occur within interaction. Thus, the situation that best suits informal statistical inference is one that necessitates communication based on argumentation using verbal language. In informal statistical inference, abduction and induction are carried out as construction of argumentation and verification of argumentation, respectively. Statistical inference as abduction and construction of argumentation are quite similar because the purpose of both of them

is drawing a conclusion from data considering the characteristics and patterns of the data and the context. On the other hand, statistical inference as induction and verification of argumentation in informal statistical inference greatly differ in terms of their use or lack of use of probability. Students cannot use the quantification approach if they did not learn probability. Thus, it is required to address the uncertainty that is appropriate for students' level. Consequently, it was found that students could use probabilistic representations, draw a conclusion by recognizing the importance of repeated sampling, and attempt to validate the argumentation by establishing norms for dealing with uncertainty during the communication.

Due to the nature of the informal statistical inference, situations where communication based on argumentation must happen in the teaching and learning process. Therefore, assessments must occur in parallel with the teaching and learning process. For this reason, the meaning of integration of instruction and assessment was examined and several assessment models, such as the general assessment triangle model by NRC (2001) and assessment models based on interaction, were analyzed. As a result, an assessment model for informal statistical inference was developed. The assessment model includes the integration of instruction and assessment as a universal set and interaction between a teacher and students as two intersecting sets. The model includes teacher's providing tasks, students' initial responses, teacher's interpretation based on an assessment element, teacher's feedback, and students' final responses, according to the procedure of the assessment. Assessment tasks have to be open-ended and approached by multiple aspects from the different levels of students. Tasks have

three contexts: “to infer about population from one sample,” “to infer about population from two samples,” and “to infer based on the probability model.” Assessment elements are presented as conceptual understanding frameworks and as assessment criteria for argumentation based on the concepts and thinking induced from the nature of informal statistical inference. Finally, feedback strategies were induced based both on the concept of attention provided by Watson (2007) and on the role of teachers in statistical discourse presented in previous studies.

The significance of this study lies in its exploration of the possibility of assessing statistics in the aspect of finding the appropriate assessment method for informal statistical inference rather than employing the existing method of writing-based assessments. Most of the studies in assessment concentrate on the simple execution, focusing on how to construct writing or multiple-choice questions to reflect well the assessment elements whereas this study is focusing on how to assess based on the nature of informal statistical inference. This supports Gal and Garfield’s proposal (1997) about the need of developing an assessment tool that applies the characteristics of statistics as well as Shepard’s assertion (2000) that instruction and assessment should coincide in their goals. As informal statistical inference is emerged and promoted within the communication through argumentation, assessment cannot take place outside of the scope of informal statistical inference. Especially from the sociocultural point of view, the focus of assessment is in looking at students’ participation during the teaching-learning process. It is distinct from the previous way of separating teaching and learning independently, and its significance rests in

informing the nature of assessment in an integrated situation.

As assessment of informal statistical inference occurs during the process of communication through argumentation, the verbal language of students needs to be observed quite closely. There has been much discussion around the importance of verbal language in the area of statistics. Nevertheless, most concentrate on the challenges posed by language. For example, in the studies of lexical ambiguity that is caused by the disruption between statistical concepts and natural language, there is a movement to restrict those confusing terms (Kaplan, Fisher, & Rogness, 2009; 2010). On the contrary, in this study, the discussion about informal statistical inference equating with informal argumentation based on the natural language is intensified with its characteristics of argumentation. In other words, many important observations can be derived from the verbal language; thus, any lexical ambiguity should be accepted as is. The verbal language of students is bound to be incomplete and unclean in the process of combining context and natural language. Nevertheless, there are some meaningful elements in it, so it is important to know how to extract them. Therefore, it is required to actively look for an argumentation reflecting context, a use of probabilistic expression, and a norm of justification in the meaning of students' words.

The integration of instruction and assessment is a new trend that can sometimes result in difficulty for educators who need to conduct the assessment (Webb, 1992). Watson (2000) examined the complexity that instructors face when evaluating their students' performance informally in classrooms. The complexity arose because

teachers not only had to understand the content thoroughly, but also had to implement new assessments at the level of students' knowledge. When integration of instruction and assessment occurs from the interaction between teachers and students, expertise in assessment is a prerequisite to all teachers. In the case of assessment in the cognitive perspective, teachers are overlooked on what to ask, what to look for in performance, how to estimate student's knowledge, and how to adjust to the level of knowledge in the case of lack of understanding in the subject (Morgan & Watson, 2002). In the case of this study's main subject, informal statistical inference, the integration of instruction and assessment can be perceived as even more difficult to educators who are familiar with teaching mathematics. Therefore, understanding the concepts and thinking process of informal statistical inference is essential and assessment must be based on this understanding.

Meanwhile, assessing from the sociocultural point of view, the appropriate role of the educator is also important. In informal statistical inference, students' seeking peer consensus and clarification, privileging interesting outcomes, and basing conclusions on data are underscored as social components. Hence, it became necessary for the teacher to form a method of socio-mathematical or socio-statistical norms to define students' reasoning (Makar et al., 2011). In the sociocultural perspective, the teacher can provide the most influence by improving students' understanding of concepts and argumentation during the course of leading a discourse through the communication between both parties in a classroom (Sfard, 2008). With this interaction, the educator can guide students to an awareness of informal statistical

inference by forming definitions and leading the discussion. Therefore, teachers are expected to be familiar with statistical discourses and provide statistical argumentation beyond the thinking processes used in mathematic classes.

This study provides a theoretical foundation by discussing an assessment method in statistics education based on the nature of informal statistical inference. In future research, the assessment model should be applied in the situation of integration of instruction and assessment and revised consistently by analyzing the assessment process in the classroom. It becomes particularly necessary to elaborate on both assessment elements and feedback strategies based on the students' responses from the result of teaching experiment analysis.

REFERENCES

Abelson, R. P. (1995). *Argumentation as Principled Argument*. Hillsdale, NJ: Lawrence Erlbaum.

Albert, L. (2000). Outside-in, inside-out: seventh-grade students' mathematical thought processes. *Educational Studies in Mathematics, 41*, 109-141.

Allen, K., Stone, A., Reed-Rhoads, T., & Murphy, T. J. (2004). The statistics concept inventory: Developing a valid and reliable instrument. In *Proceedings of the American Society for Engineering Education Annual Conference and Exposition*. Salt Lake City.

American Statistical Association. (2005). *Guidelines for Assessment and Instruction in Statistical Education: College Report*. Alexandria, VA: Author.

Bakker, A. (2004). *Design Research in Statistics Education: On Symbolizing and Computer Tools*. Utrecht, The Netherlands: CD Beta Press.

Bakker, A., & Derry, J. (2011). Lessons from inferentialism for statistics education. *Mathematical Thinking and Learning, 13*, 5-26.

Bakker, A., & Gravemeijer, K. (2004). Learning to reason about distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 147-168). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bakker, A., Kent, P., Derry, J., Noss, R., & Hoyles, C. (2008). Statistical inference at work: statistical process control as an example. *Statistics Education Research*

Journal, 7(2), 130-145.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning, 2*(1&2), 75-97.

Bennett, R. (2011). Formative assessment: a critical review. *Assessment in Education: Principles, Policy, & Practice, 18*(1), 5-25.

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman and B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics (ICOTS-7)*. Salvador, Bahia, Brazil.

Ben-Zvi, D., & Aridor, K. (2012). Children's wonder how to wander between data and context. In *Proceedings of the 12th International Congress on Mathematics Education (ICME-12)*. Seoul, Korea.

Ben-Zvi, D., Aridor, K., Maker, K., & Bakker, A. (2012). Students' emergent articulations of uncertainty while making informal statistical inferences. *ZDM the International Journal on Mathematics Education, 44*(7), 913-925.

Ben-Zvi, D., & Gil, E. (2010). The role of context in the development of students' informal inferential reasoning. In C. Reading (Ed.), *Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS-8).* Ljubljana, Slovenia.

Ben-Zvi, D., Gil, E., & Apel, N. (2007). What is hidden beyond the data? Helping young students to reason and argue about some wider universe. In D. Pratt & J. Ainley (Eds.), *Proceedings of the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5)*. University of

Warwick, UK.

Ben-Zvi, D., Makar, K., & Bakker, A. (2009). Towards a framework for understanding students' informal statistical inference and argumentation. In K. Makar (Ed.), *Proceedings of the Sixth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-6).* University of Queensland, Brisbane, Australia.

Ben-Zvi, D., Makar, K., Bakker, A., & Aridor, K. (2011). Children's emergent inferential reasoning about samples in an inquiry-based environment. In M. Pytlak, T. Rowland, & E. Swoboda (Eds.), *Proceedings of the Seventh Congress of the European Society for Research in Mathematics Education.* University of Rzeszow, Poland.

Ben-Zvi, D., & Sfard, A. (2007). Ariadne's thread, daedalus' wings, and the learner's autonomy. *Education & Didactique, 1*(3), 123-142.

Black, P., & Wiliam, D. (2006). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and Learning* (pp. 81-100). London: Sage Publications.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability, 21*, 5-31.

Brookhart, S. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice, 22*(4), 5-12.

Burrill, G. (2007). The role of formative assessment in teaching and learning statistics.

In *Proceedings of the IASE/ISI Satellite Conference on Assessing Student Learning in Statistics.* Guimaraes, Portugal.

Chance, B., delMas, R., & Garfield, J. (2004). Reasoning about sampling distribution. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 295-323). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Choi, J. -S., Yun, Y. -S., & Hwang, H. -J. (2014). A study on pre-service teachers' understanding of random variable. *Journal of Korea Society of Educational Studies in Mathematics: School Mathematics*, *16*(1), 19-37. (in Korean)

Ciancetta, M. A. (2007). *Statistics Students Reasoning when Comparing Distributions of Data* (Unpublished doctoral dissertation). Portland State University, Portland.

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

Cobb, P., & McClain, K. (2004). Principles of instructional design for supporting the development of students' statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 375-395). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Cobb, P., Stephan, M., McClain, K., & Gravemeijer, K. (2001). Participating in classroom mathematical practices. *The Journal of the Learning Sciences, 10*(1&2), 113-163.

Cowie, B., & Bell, B. (1999). A model of formative assessment in science education. *Assessment in Education: Principles, Policy & Practice, 6*(1), 101-116.

Curcio, F. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education, 18*(5), 382-393.

Ekmekci, A. (2013). *Mathematical Literacy Assessment Design: A Dimensionality Analysis of Programme for International Student Assessment (PISA) Mathematics Framework* (Unpublished doctoral dissertation). The University of Texas at Austin, Texas.

Gal, I. (2004). Statistical literacy: Meanings, components, responsibilities. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 47-78). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Gal, I., & Garfield, J. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J. Garfield (Eds.), *The Assessment Challenge in Statistics Education* (pp. 1-13). Amsterdam, The Netherlands: IOS Press.

Garfield, J., & Ben-Zvi, D. (2008). *Developing Students' Statistical Reasoning: Connecting Research and Teaching Practice*. Dordrecht, The Netherlands: Springer.

Garfield, J., & Ben-Zvi, D. (2009). Helping students develop statistical reasoning: implementing a statistical reasoning learning environment. *Teaching Statistics, 31*(3), 72-77.

Garfield, J., delMas, R., & Chance, B. (2002). *Assessment Resource Tools for*

Improving Statistical Thinking. Retrieved from https://apps3.cehd.umn.edu/artist/index.html

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual what you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences* (pp. 391-408). Thousand Oaks, CA: Sage Publications.

Gower, B. (1997). *Scientific Method: A Historical and Philosophical Introduction*. London: Routledge.

Hacking, I. (1980). The theory of probable inference: Neyman, Peirce, and Braithwaite. In D. H. Mellor (Ed.), *Science, Belief, and Behaviour: Essays in honour of R. B. Braithwaite* (pp. 141-160). Cambridge: Cambridge University Press.

Hacking, I. (1990). *The Taming of Chance*. Cambridge: Cambridge University Press.

Hacking, I. (2012). *The Taming of Chance*. (H. Geong, Trans.). Seoul: Badabooks Publications. (Original work Published 1990) (in Korean)

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research Online, 7*(1), 1-20.

Harlen, W. (2006). On the relationship between assessment for formative and summative purposes. In J. Gardner (Ed.), *Assessment and Learning* (pp. 103-117). London: Sage Publications.

Heritage, M. (2010). Formative assessment and next-generation assessment systems:

are we losing an opportunity? *Paper prepared for the Council of Chief State School Officers*. Washington, D. C.

Herman, J. (2013). Formative assessment for next generation science standards: a proposed model. *Paper presented at Invitational Research Symposium on Science Assessment*. Washington, D. C.

Ho, Y. C. (1994). Abduction? Deduction? Induction? Is there a logic of exploratory data analysis? *Paper presented at the Annual Meeting of American Educational Research Association*. New Orleans, Louisiana.

Holcomb, J., Chance, B., Rossman, A., & Cobb, G. (2010). Assessing student learning about statistical inference. In C. Reading (Ed.), *Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS-8)*. Ljubljana, Slovenia.

Holland, J., Holyoak, K., Nisbett, R., & Thagard, P. (1989). *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: MIT Press.

Hwang, S. -G., & Yoon, J. -H. (2011). A study on teaching continuous probability distribution in terms of mathematical connection. *Journal of Korea Society of Educational Studies in Mathematics: School Mathematics*, *13*(3), 423-446. (in Korean)

Jeon, Y. -S. (2000). Inductive Logic of Carnap. In C. Lee (Ed.), *Inductive Logic and Philosophy of Science* (pp. 79-105). Seoul: Chulhak goa Hyunsilsa. (in Korean)

Jeon, Y. -S. (2013). *When Can We Make an Inductive Leap?* Seoul: Acanet. (in

Korean)

Kalinowski, P., Fidler, F., & Cumming, G. (2008). Overcoming the inverse probability fallacy: A comparison of two teaching interventions. *Experimental Psychology, 4*(4), 152-158.

Kaplan, J., Fisher, D., & Rogness, N. (2009). Lexical ambiguity in statistics: what do students know about the words association, average, confidence, random and spread? *Journal of Statistics Education, 17*(3). 1-19. Available online: http://www.amstat.org/publications/jse/v17n3/kaplan.html

Kaplan, J., Fisher, D., & Rogness, N. (2010). Lexical ambiguity in statistics: how students use and define the words: association, average, confidence, random and spread. *Journal of Statistics Education, 18*(2). 1-22. Available online: http://www.amstat.org/publications/jse/v18n2/kaplan.pdf

Kim, H. -K. (2001). *Statistical Inference*. Seoul: Acanet. (in Korean)

Kim, K. -S. (2007). *Logic and Critical Thinking*. Seoul: Chulhak goa Hyunsilsa. (in Korean)

Kim, S. -H., & Lee, C. -H. (2002). Abduction as a mathematical reasoning. *Journal of Educational Research in Mathematics, 12*(2), 275-290. (in Korean)

Kim, W. -C., Kim, J. -J., Park, B. -U., Park, S. -H., Song, M. -S., Lee, S. -Y., ... Cho, S. S. (2006). *General Statistics (1st Ed.)*. Seoul: Youngchi Moonhwa Sa. (in Korean)

Kim, Y. -H. (2006). A study of using the terminology of sampling error and sampling distribution. *Journal of the Korean School Mathematics Society, 9*(3), 309-

316. (in Korean)

Klenowski, V. (2009). Editorial: Assessment for learning revisited: An Asia-Pacific perspective. *Assessment in Education: Principles, Policy, & Practice, 16*(3), 263-268.

Ko, E. -S. (2012). A comparison of mathematically talented students and non-talented students' level of statistical thinking: statistical modeling and sampling distribution understanding. *Journal of Gifted/Talented Education, 22*(3), 503-525. (in Korean)

Ko, E. -S., & Lee, K. -H. (2011). Pre-service teachers' understanding of statistical sampling. *Journal of Educational Research in Mathematics, 21*(1), 17-32. (in Korean)

Krummheuer, G. (1995). The ethnography of argumentation. In P. Cobb & H. Bauersfeld (Eds.), *The Emergence of Mathematical Meaning: Interaction in Classroom Cultures* (pp. 229-269). Hillsdale, NJ: Lawrence Erlbaum.

Kwon, C. -E., (1996). Inductive inference of Aristotle. In H. Yeo (Ed.), *Logic and Truth* (pp. 211-224). Seoul: Chulhak goa Hyunsilsa. (in Korean)

Lee, J. -M. (1988). Logic of experiment: scientific explanation and inference. In Korean Psychological Association (Ed.), *Introduction to Methodologies of Experimental Psychology: Hypothesis generating, Experimental Designs, and Analysis* (pp. 73-116). Seoul: Seung Won Sa. (in Korean)

Lee, K. -H. (1996). *A Study on the Didactic Transposition of the Concept of Probability* (Unpublished doctoral dissertation). Seoul National University,

Seoul. (in Korean)

Lee, K. -H., & Ji, E. -J. (2005). Pedagogical significance and students' informal knowledge of sample and sampling. *Journal of Educational Research in Mathematics, 15*(2), 177-196. (in Korean)

Lee, M. -S., & Park, Y. -H. (2006). A case study on understanding of the concept of sampling and data analysis by elementary 6th graders. *Journal of Korea Society of Educational Studies in Mathematics: School Mathematics, 8*(4), 441-463. (in Korean)

Lee, O. -S., Lim, Y. -B., Seong, N. -G., & So, B. -S. (2000). *Introduction to Statistics (2nd Ed.)*. Seoul: Kyung Moon Sa. (in Korean)

Lee, Y. -H., & Lee, E. -H. (2010). The design and implementation to teach sampling distributions with the statistical inferences. *Journal of Korea Society of Educational Studies in Mathematics: School Mathematics, 12*(3), 273-299. (in Korean)

Lee, Y. -H., & Nam, J. -H. (2005). An epistemological inquiry on the development of statistical concepts. *Journal of Korea Society of Mathematics Education Series A: The Mathematics Education, 44*(3), 457-475. (in Korean)

Lee, Y. -H., & Shin, S. -Y. (2011). Features of sample concepts in the probability and statistics chapters of Korean mathematics textbooks of grades 1-12. *Journal of Educational Research in Mathematics, 21*(4), 327-344. (in Korean)

Lipson, K. (2002). The role of computer based technology in developing understanding of the concept of sampling distribution. In *Proceedings of the*

Sixth International Conference on Teaching Statistics (ICOTS-6). Cape Town, South Africa.

Lyon, E. (2011). Beliefs, practices, and reflection: exploring a science teacher's classroom assessment through the assessment triangle model. *Journal of Science Teacher Education, 22*, 417-435.

Makar, K., Bakker, A., & Ben-Zvi, D. (2011). The reasoning behind informal statistical inference. *Mathematical Thinking and Learning, 13*(1-2), 152-173.

Makar, K., & Confrey, J. (2002). Comparing two distributions: investigating secondary teachers' statistical thinking. In *Proceedings of the Sixth International Conference on Teaching Statistics (ICOTS-6)*. Cape Town, South Africa.

Makar, K., & Rubin, A. (2007). Beyond the bar graph: Teaching informal inference in primary school. In D. Pratt & J. Ainley (Eds.), *Proceedings of the Fifth International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5)*. University of Warwick, UK.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82-105.

Ministry of Education and Science Technology. (2011). *Mathematics Curriculum*. Seoul: Author. (in Korean)

Morgan, C., & Watson, A. (2002). The interpretative nature of teachers' assessment of students' mathematics: issues for equity. *Journal for Research in Mathematics Education, 33*(2), 78-110.

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Research Council. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, D. C.: National Academy Press.

Paparistodemou, E., & Meletiou-Movrotheris, M. (2008). Developing young students' informal inference skills in data analysis. *Statistics Education Research Journal, 7*(2), 83-106.

Park, J. (2012). *Developing and Validating an Instrument to Measure College Students' Inferential Reasoning in Statistics: An Argument-Based Approach to Validation* (Unpublished doctoral dissertation). University of Minnesota, Minnesota.

Peirce, C. S. (1878). Deduction, induction, and hypothesis. *Popular Science Monthly, 13*, 470-482.

Pellegrino, J., Baxter, G., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education, 24*, 307-353.

Pfannkuch, M. (2005). Probability and statistical inference: how can teachers enable learners to make the connection? In G. Jones (Ed.), *Exploring Probability in School: Challenges for Teaching and Learning* (pp. 267-294). New York, NY: Springer.

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal, 5*(2), 27-45.

Pfannkuch, M. (2007). Year 11 students' informal inferential reasoning: a case study about the interpretation of box plots. *International Electronic Journal of Mathematics Education, 2*(3), 149-167.

Pfannkuch, M. (2008). Building sampling concepts for statistical inference: A case study. In *Proceedings of the 11th International Congress on Mathematics Education (ICME-11)*. Monterrey, Mexico.

Pfannkuch, M. (2011). The role of context in developing informal statistical inferential reasoning: A classroom study. *Mathematical Thinking and learning, 13*, 27-46.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*(1), 4-13.

Rao, C. R. (1997). *Statistics and Truth: Putting Chance to Word (2nd Ed.)*. Singapore: World Scientific.

Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: From a variation perspective. *Statistics Education Research Journal, 5*(2), 46-68.

Reading, C., & Shaughnessy, M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning, and Thinking* (pp. 201-226). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Rossman, A. (2008). Reasoning about informal statistical inference: One statistician's view. *Statistics Education Research Journal, 7*(2), 5-19.

Rubin, A., Bruce, B., & Tenney, Y. (1991). Learning about sampling: Trouble at the core of statistics. In D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics. Vol. 1. School and General Issues* (pp. 314-319). Voorburg, The Netherlands: International Statistical Institute.

Rubin, A., Hammerman, J., & Konold, C. (2006). Exploring informal inference with interactive visualization software. In *Proceedings of the Seventh International Conference on Teaching Statistics (ICOTS-7)*. Salvador, Bahia, Brazil.

Saldanha, L., & Thompson, P. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51,* 257-270.

Salmon, W. C. (1967). *The Foundations of Scientific Inference*. Pittsburgh, PA: University of Pittsburgh Press.

Salsburg, D. (2001). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. New York, NY: W. H. Freeman.

Sedlmeier, P., & Gigerenzer, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making, 10*(1), 33-51.

Sfard, A., (2008). *Thinking as Communicating: Human Development, the Growth of Discourse, and Mathematizing.* Cambridge: Cambridge University Press.

Shaughnessy, J. M., Ciancetta, M., Best, K., & Canada, D. (2004). Students' attention to variability when comparing distribution. *Paper presented at the 82nd*

Annual Meeting of the National Council of Teachers of Mathematics. Philadelphia, PA.

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Shin, B. -M. (2012). A study on a didactic transposition method and students' understanding for the normal distribution. *Journal of Educational Research in Mathematics, 22*(2), 117-136. (in Korean)

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Swaffield, S. (2011). Getting to the heart of authentic assessment for learning. *Assessment in Education: Principles, Policy, & Practice, 18*(4), 433-449.

Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: using action research to explore and modify theory. *British Educational Research Journal, 27*(5), 615-631.

Toulmin, S. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.

Uchiyama, M. (2004). *Teachers' Use of Formative Assessment in Middle School Reform-Based Mathematics Classrooms* (Unpublished doctoral dissertation). University of Colorado, Colorado.

Van den Heuvel-Panhuizen, M., & Becker, J. (2003). Towards a didactic model for assessment design in mathematics education. In A. J. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick & F. K. S. Leung (Eds.), *Second International Handbook of Mathematics Education* (pp. 689-716). Dordrecht, The

Netherlands: Kluwer Academic Publishers.

Watson, A. (2000). Mathematics teachers acting as informal assessors: practices, problems and recommendations. *Educational Studies in Mathematics, 41*, 69-91.

Watson, A. (2007). The nature of participation afforded by tasks, questions and prompts in mathematics classrooms. *Research in Mathematics Education, 9*(1), 111-126.

Watson, J. (2006). *Statistical Literacy at School: growth and goals*. Mahwah, NJ: Lawrence Erlbaum Associates.

Watson, J. (2008). Exploring beginning inference with novice grade 7 students. *Statistics Education Research Journal, 7*(2), 59-82.

Watson, J., Callingham, R., & Kelly, B. (2007). Students' appreciation of expectation and variation as a foundation for statistical understanding. *Mathematical Thinking and Learning, 9*(2), 83-130.

Watson, J., & Moritz, J. (1999). The beginning of statistical inference: comparing two data sets. *Educational Studies in Mathematics, 37*(2), 145-168.

Webb, N. (1992). Assessment of students' knowledge of mathematics: steps toward a theory. In D. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 334-368). New York, NY: Macmillan.

Wiliam, D. (2007). Keeping learning on track: Classroom assessment and the regulation of learning. In F. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 1053-1098). Charlotte, NC:

Information Age Publishing.

Wiliam, D., & Thompson, M. (2008). Integrating assessment with learning: what will it take to make it work? In C. A. Dwyer (Ed.), *The Future of Assessment: Shaping Teaching and Learning* (pp. 53-82). Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M., & Carstensen, C. (2007). Assessment to improve learning in mathematics: the BEAR assessment system. In A. Schoenfeld (Ed.), *Assessing Mathematical Proficiency* (pp. 311-332). New York, NY: Cambridge University Press.

Woo, J. -H. (2000). An exploration of the reform direction of teaching statistics. *Journal of the Korea Society of Educational Studies in Mathematics: School Mathematics, 2*(1), 1-27. (in Korean)

Woo, J. -H., Chong, Y. -O., Park, K. -M., Lee, K. -H., Kim, N. -H., Na, G. -S., & Yim, J. -H. (2006). *Research Methods in Mathematics Education*. Seoul: Kyung Moon Sa. (in Korean)

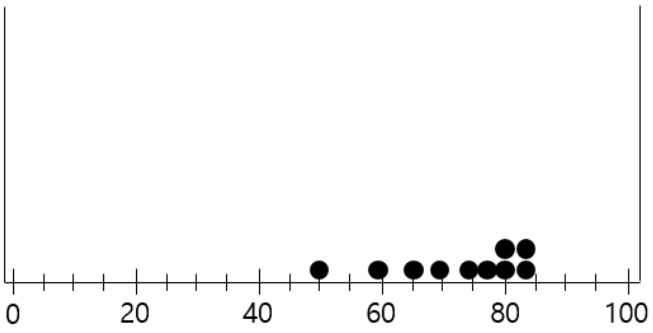
Zieffler, A., Garfield, J., delMas, R., & Reading, C. (2008). A framework to support research on informal inferential reasoning. *Statistics Education Research Journal, 7*(2), 40-58.

APPENDIX

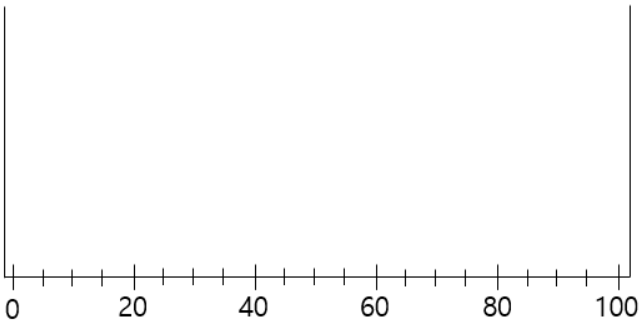
Assessment Tasks for Informal Statistical Inference

"To infer about population from one sample"

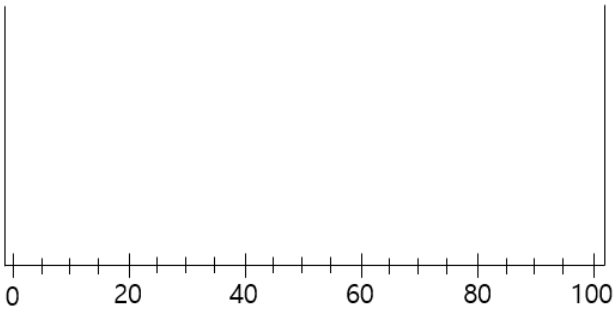
1. 1,000 students who learn mathematics from the same teacher took a final test. The test scores for a random sample of 10 students are shown in the dot plot below.



   1) Consider a random sample of 25 students. What does a graph look like? Sketch a dot plot of 25 scores and explain your reasoning as much as possible.



2) Consider the whole 1,000 students. What does a graph look like? Draw an outline of the distribution and explain your reasoning as much as possible.



3) To infer the distribution of the whole 1,000 students' scores more accurately, what information is necessary?

2. Suji has a pocket filled with 1,000 candies. The colors of the candies are red, yellow, and blue. Suji wants to know the ratio of each color of candies.
   1) What way can Suji apply to investigate the ratio of each color of candies?
   2) Suji decides to take a random sample of 100 candies to investigate the ratio of each color of candies. What are the advantages and disadvantages of this way?
   3) Suji comes up with two ways to take a random sample of 100 candies. Which do you think is more proper? Explain your reasoning.
      - Take a random sample of 100 candies at once and investigate the ratio of each color.
      - Take a random sample of 10 candies at once, put them back, and repeat ten times. Investigate the ratio of each color every time.

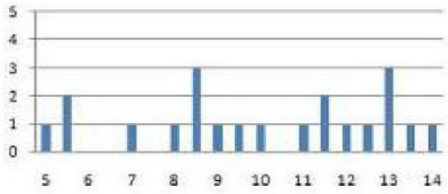
"To infer about population from two samples"

1. Recently, movie theaters show commercials and previews before a movie begins. The wait-time for a movie is the difference between its advertised start time and the actual start time. 21 students decided to investigate the wait-times at two popular movie theaters. Each student attended two movies, a different movie in each theater, and recorded the wait-times in minutes below.

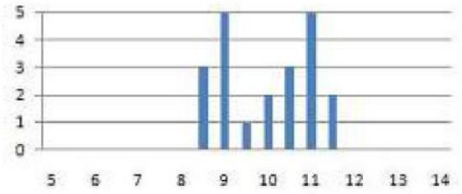
Movie Theater A :						
5.0	12.0	13.0	5.5	9.5	13.0	5.5
11.5	8.0	8.5	14.0	13.0	8.5	7.0
8.5	12.5	13.5	11.5	9.0	10.0	11.0
Average = 10 minutes						

Movie Theater B :						
11.5	11.0	9.0	10.5	8.5	11.0	9.0
10.5	9.5	8.5	10.0	11.5	10.5	8.5
9.0	11.0	11.0	9.5	10.0	9.0	11.0
Average = 10 minutes						

Joonsu drew the following two graphs using the above data.



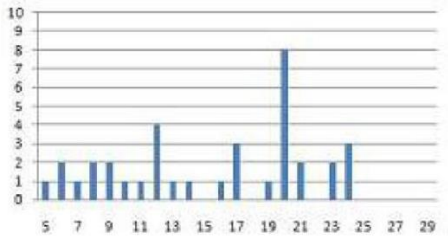
Movie Theater A



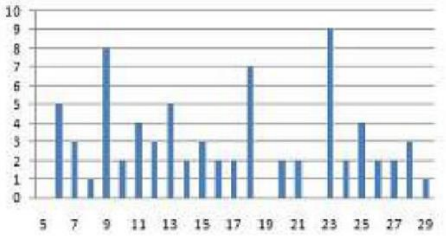
Movie Theater B

1) Joonsu concluded that there is no difference in wait-times for movies in both theaters since the averages are the same. Do you agree or disagree? Explain your reasoning.
2) Which of these theater would you choose to see a movie? Explain your reasoning.

2. There are two hospitals in the area of Jia's school. The school board had to make a decision about which one of the two hospitals to call when emergencies arise at their school. The response-time is the difference between the time of the call and the actual arrival time. Students obtained the most recent 36 response-times for hospital A and the most recent 74 response-times for hospital B. Jia made the following two graphs from the data.



Hospital A

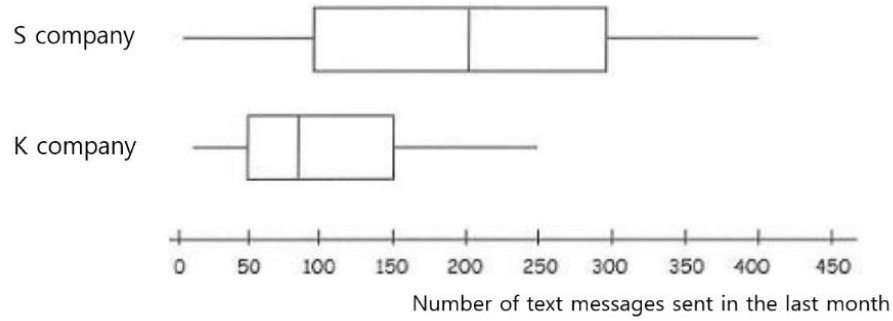


Hospital B

1) Compare the response-times of the two hospitals and present the conclusions of comparison.
2) Which of these hospitals would you choose to call when emergencies arise? Explain your reasoning.

3. S company and K company, which are cell phone companies, provide text message services. From S company, a user can send 500 text messages for 10,000 (won) per month, and from K company, a user pays for each text message at a cost of 20 (won) each. Jinyoung wants to compare the number of text messages that users from each company have sent. She surveyed 100 people from each company and represented the results using boxplots. What

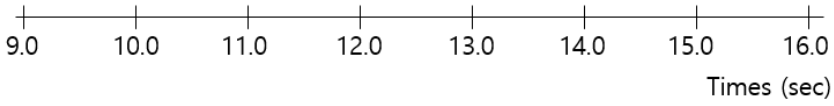
can she infer from this data? Explain the reasons, giving as many as possible.



4. Suppose that there is a special summer camp for track athletes. There is one group of 100 athletes that run a particular race, and they are all pretty similar in height, weight, and strength. They are randomly assigned to either group A or group B. Group A gets an additional weight-training program. Group B gets a regular training program without weights. All the students from both groups run the race before and after the camp, and their times are recorded, so that the data could be used to compare the effectiveness of the two training programs.
   1) Predict the median of the running time records of each group before the camp.
   2) If the additional weight-training program is not effective, what would you expect in the running time records of both groups? Present it using boxplots and explain your reasoning.

Group A

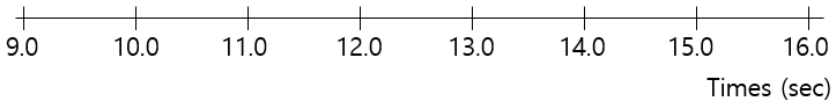
Group B



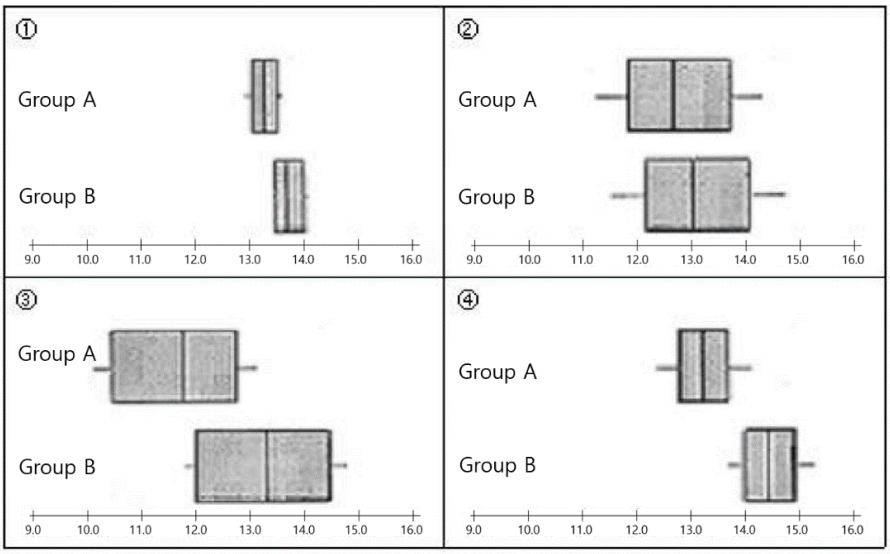
3) If the additional weight-training program is effective, what would you expect in the running time records of both groups? Present it using boxplots and explain your reasoning.

Group A

Group B



4) Presented below are some possible graphs that show boxplots for different scenarios where the running times are compared for the students in the two groups. Rank the four pairs of graphs on how convincing they are in making an argument that the weight-training program was more effective in decreasing athletes' times (from the least convincing to the most convincing evidence). Explain your reasoning.



5) For the pair of graphs that provide the most convincing evidence, would you be willing to generalize the effects of the training programs to all similar athletes on track teams based on these samples? Why or why not?

"To infer based on the probability model"

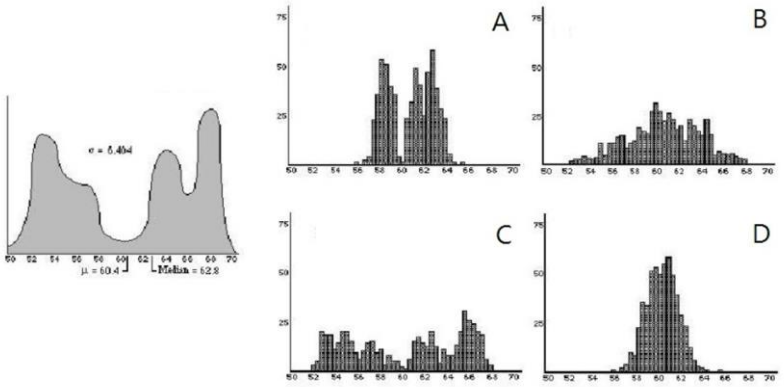
1. The result of an entrance examination shows that the distribution of mathematics test scores is a normal distribution with an average score of 48 and a standard deviation of 20. Would you expect to find a difference between the probability that a single student scores above 68 and the probability that in a random sample of 25 people, the sample mean score is above 68? Explain your reasoning.

2. In a certain country, males must take a physical examination at a local public health center when they turn 18. In addition to other information, the height of each male is measured. The national average height for 18-year-old males is 1 m 75 cm. Every day for one year, about 5 men visited a small local public health center, and about 50 men visited a large local public health center. At the end of each day, a clerk at each health center computed and recorded the average height of the men who had visited it that day. Which health center would you predict regarding the number of days on which the average height for the day was more than 1 m 80 cm? Explain your reasoning.

3. Scores on a particular college entrance exam are not normally distributed. The distribution of test scores is greatly skewed toward lower values with a mean of 20 and a standard deviation of 3.5. A research team plans to take a simple

random sample of 50 students from different high schools across the United States. What would be the shape of the sampling distribution of average test scores for a sample of size 50?

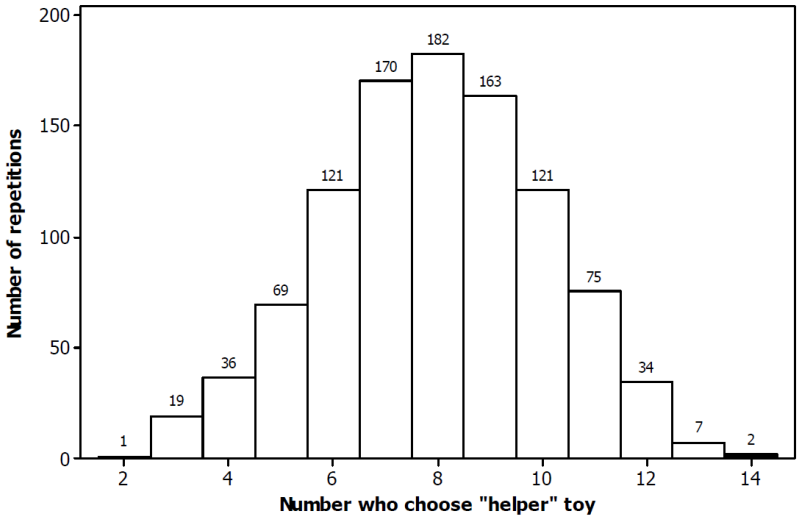
4. The graph below on the left represents a population distribution, and the four graphs below on the right represent possible empirical sampling distributions for 500 samples. Which distribution of sample means best represents an empirical distribution for 500 samples of size 4 and also for size 25? Justify your choice of graphs and explain your reasoning.



5. A research institute conducts an experiment with infants. The purpose of the experiment is to investigate whether infants take into account an individual's actions towards others in evaluating that individual as appealing or aversive. 10-month-old infants were shown a "climber" character (a piece of wood with "googly" eyes glued onto it) that could not make it up a hill in two tries. Then they were shown two scenarios for the climber's next try: one where the

climber was pushed to the top of the hill by another character (“helper”) and one where the climber was pushed back down the hill by another character (“hinderer”). Each infant was alternately shown these two scenarios several times. Then the child was presented with both pieces of wood (the helper and the hinderer) and asked to pick one to play with.

1) In a sample of 16 infants, 14 of them chose the helper toy. Based on this result, can you conclude that infants consider an individual’s actions towards others in evaluating that individual as appealing?
2) By flipping a fair coin 16 times and counting the number of heads and repeating this process 1,000 times, we obtained the results presented in the below figure. Based on this figure and the fact that “in a sample of 16 infants, 14 of them chose the helper toy,” can you conclude that infants consider an individual’s actions toward others in evaluating that individual as appealing?



국문초록

비형식적 통계적 추리의 평가

최근 통계교육 연구에서는 대학 수준의 통계적 추리를 돕기 위한 선행 단계로서 비형식적 통계적 추리의 교수-학습을 강조하고 있다. 이에 따라 비형식적 통계적 추리의 의미를 밝히고, 학생들의 비형식적 통계적 추리를 향상시키기 위한 방법을 탐색하는 것에 주목해왔다. 교수-학습에 대한 연구와 더불어, 학생들의 비형식적 통계적 추리를 어떻게 평가할 수 있을 지가 핵심적인 문제이다. 그 동안 통계교육 연구에서는 수학과 구별되는 통계학의 본질적인 특성을 반영한 평가 방안이 필요하다는 주장이 제기되었음에도 불구하고 아직까지 그에 대한 실질적인 논의는 부족하였다. 본 연구는 비형식적 통계적 추리의 본질을 밝히고, 이를 반영한 비형식적 통계적 추리의 평가 방안을 탐색하는 것에 목표를 둔다.

통계적 추리의 인식론적 분석을 통해 통계적 추리는 크게 귀납추리와 가추라는 사고 요소를 가지고 있음을 확인하였다. 귀납추리로서의 통계적 추리는 불확실성을 어떠한 방식으로 다루는지에 따라 통계적 추리의 고유한 특성을 규정하며, 이는 확률을 사용한 불확실성의 계량화와 부정논법의 도입이라는 시도로 나타난다. 가추로서의 통계적 추리는 표본의 특성 및 패턴을 바탕으로 맥락을 고려하여 이를 잘 설명할 수 있는 모집단에 대한 가설을 생성하는 측면에서 중요하다. 귀납추리와 가추는 각각 통계적 추리의 부분적인 사고 요소로서 의미를 가지며, 서로 다른 단계로 파악하는 것이 필요함을 확인하였다.

비형식적 통계적 추리에 대한 교수-학습 관련 선행연구를 분석하여, 통계적 추리에서 필수적인 개념 및 사고 요소들이 비형식적 통계적 추리에서 어떻게 다루어지며 학생들이 각 요소에 대해 무엇을 이해해야 하는지를 중심으로 확인하였다. 개념 요소로는 대푯값 및 변이를 포함한 기술통계 개념, 표본과

모집단, 표본의 크기, 표본분포에 대해, 사고 요소로는 가추와 귀납추리에 대해 분석하였다. 가추와 귀납추리는 비형식적 통계적 추리에서 각각 논증의 구성, 논증의 타당화로 다루어지며, 특히 학생들은 논증의 타당화를 위해 확률적 표현을 사용하고, 반복 표집의 중요성을 인식하며, 불확실성을 다루기 위한 규범을 정립해야 함을 확인하였다. 비형식적 통계적 추리는 자연언어를 바탕으로 하는 비형식적 논증이며, 맥락을 기반으로 하고 교실에서의 상호작용을 바탕으로 일어난다는 특징을 가진다. 이에 따라 비형식적 통계적 추리를 위해 필요한 상황은 논증을 바탕으로 의사소통이 일어나는 상황임을 확인하였다.

비형식적 통계적 추리의 본질을 확인함으로써 이를 평가하기 위해 교수-학습과 평가의 통합이 필요함을 확인하였다. 교수-학습과 평가의 통합의 의미를 밝히고, 선행연구에서 제시된 평가모델들을 분석한 결과를 바탕으로 비형식적 통계적 추리의 평가 모델을 고안하였다. 평가 모델은 비형식적 통계적 추리에 대한 평가가 교수-학습과의 통합 상황에서 교사와 학생의 상호작용을 바탕으로 이루어져야 한다는 점을 반영한다. 또한 평가 모델은 평가 절차를 도식화하며, 그 절차는 교사의 과제 제시, 학생의 초기 반응, 평가 요소를 바탕으로 한 교사의 해석, 교사의 피드백, 학생의 최종 반응으로 이루어진다. 비형식적 통계적 추리의 평가를 위한 과제의 특성, 평가 요소, 피드백의 특성을 도출하였다. 마지막으로 본 연구의 교수학적 시사점을 논의하고, 본 연구에서 고안한 평가 모델을 바탕으로 하는 향후 연구 방향을 제안하였다.

주요어: 비형식적 통계적 추리, 평가 모델, 귀납추리, 가추, 논증, 교수-학습과 평가의 통합

학 번: 2010-30404