PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 82, No. 4, 2024

QUANTITATIVE TECHNIQUES WITH SMALL SAMPLE SIZES: AN EDUCATIONAL SUMMER CAMP EXAMPLE

Trina Johnson Kilty, Kevin T. Kilty
University of Wyoming, USA
E-mail: trina.j.kilty@gmail.com, kkilty1@uwyo.edu

Andrea C Burrows Borowczak, Mike Borowczak
University of Central Florida, USA
E-mail: Andrea.Borowczak@ucf.edu, mike.borowczak@ucf.edu

Abstract

A computer science camp for pre-collegiate students was operated during the summers of 2022 and 2023. The effect the camp had on attitudes was quantitatively assessed using a survey instrument. However, enrollment at the summer camp was small, which meant the well-known Pearson's Chi-Squared to measure the significance of results was not applied. Thus, a quantitative analysis method using a multinomial probability distribution as a model of a multilevel Likert scale survey was used. Exact calculations of a multinomial probability model with likelihood ratio were performed to quantitatively analyze the results of questionnaires administered to participants in two cohort groups (combined N=17). Probabilities per Likert categories were determined from the data itself using Bayes theorem with a Dirichlet prior. Each cohort functioned as part of a homogenous sample, thus allowing cohorts to be pooled. Post-test revealed significant changes in participants' attitudes after camp completion. Using this technique has implications for studies with small sample sizes. Using exact calculation of the multinomial probability model with the use of likelihood ratio as a statistical test of evidence has advantages: a) it is an exact value that can be used on any size sample, although it offers a quantitative analysis option for small sample size studies; b) depends only on what was observed during a study; c) does not require advanced calculation; d) modern spreadsheet and statistical package programs can calculate the analysis; and e) likelihood ratio employed in Bayes theorem can update prior beliefs according to evidence. Utilizing small sample size quantitative analysis can strengthen insights into data trends and showcase the importance of this quantitative technique.

Keywords: *multinomial probability model, likelihood ratio, small sample study, survey research, quantitative analysis, summer camp*

Introduction

Researchers engaged in educational experiments or in other settings, such as innovative classrooms or informal learning camps, would often like to assess the effectiveness of their efforts. One assessment strategy is to use a paired pre-/post-activity survey. However, a common issue encountered is that the camp or classroom involves a small sample of participants. The common method of measuring significant differences from pre- to post-activity is through Pearson's Chi-Square. However, this test of significant difference requires a sample size often larger than a camp or classroom provides. Thus, some other method is needed.

The problems that small sample studies present are not unusual in educational research involving educational activities (Boddy, 2016; Delice, 2010; Fugard & Potts, 2015). In self-contained classrooms, the average class size in U.S. public middle schools and high schools is slightly under seventeen (NCES, n.d.). Thus, if a typical class were to be used as a cohort in a

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 82, No. 4, 2024

study it would be a small-numbers cohort. While the number of studies documenting a positive influence of class size on student performance is small, and any conclusions drawn from them are tentative, there is likely to be continued pressure for reduced class size in the future by reason of this being among common-sense ideas (Chingos & Whitehorst, 2011). Similarly, more intensive learning activities like camps are size constrained for many reasons, including but not limited to staff-to-students ratio requirements. They may involve small-numbers cohorts (Talafian et al., 2019). Thus, an examination of statistical tools for analyzing small-numbers cohorts may be useful.

Research Problem

Measuring the effectiveness of summer camps with school-aged children is often done through survey research of camp participants answering prompts by way of a Likert Scale, the results of which are analyzed quantitatively (Baek & Touati, 2020; Cheng et al., 2021; Chou, 2020; Fiorella et al., 2019). A Chi-Squared analysis approach is generally considered appropriate if done with a sample size (*N*) greater than 25. However, many summer camps have fewer than 25 participants, or researchers would like to compare groups containing fewer than 25 individuals (Talafian et al., 2019). Thus, a problem exists of how to analyze groups of fewer than 25 quantitatively. The research team in this study engaged with two years of summer camps in 2022 and 2023. Predicting that not enough participants would enroll in camp to justify using Pearson's Chi-Squared, novel methods were considered early.

Research Focus

In early 2022, a cryptocurrency exchange and bank based in the United States approached the university where this study took place through its not-for-profit Foundation to offer a gift in support of a variety of educational and research initiatives. A portion of the funds were earmarked to support engaging camp-like educational activities pertaining to a broader set of topics and technology related to cryptocurrency for pre-collegiate children ages 10-14, henceforth referred to as campers. Based on the initial suggestion to expose campers to the central technology that enabled cryptocurrency, the camp focused on distributed trust, foundations of encoding and transforming information, and ultimately blockchain structures.

The camps were designed to use an unplugged approach of developing computational thinking implicitly rather than programming skills. Both years of camp were facilitated by graduate students in computer science and assisted by undergraduates majoring in computer science. The unplugged activities consisted of several games and puzzles which indirectly taught concepts of consensus, trust, chains, and blocks. Unplugged activities appeared to have captured the attention of the campers because there was little attrition throughout the week-long course. However, the course was under-subscribed with only about one-half of the planned enrollment of 20. From those ten campers, nine were able to complete both pre and post questionnaires. The camp was funded again during the summer of 2023, and eight campers participated in the 2023 cohort.

Because summer camp is an informal learning activity, assessments and other content knowledge tests were avoided. Moreover, at the time of camp, this study took place in a state that had just adopted standards for middle and high school computer science for the 22-23 school year, so benchmarks to measure content knowledge learning gains were not readily available (Northrup & Burrows Borowczak, 2023). Thus, it was decided to measure attitude (affect) towards computer science and to measure a 21st Century skill of teamwork, composed of trust and consensus building of campers in small groups.

https://doi.org/10.33225/pec/24.82.507 ISSN 1822-7864 (Print) ISSN 2538-7111 (Online)

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 82, No. 4, 2024

Research Aim and Research Questions

The purpose of this study was to examine a quantitative analysis approach of participants' responses to questionnaire prompts in a group of fewer than 25 participants. To fulfill this purpose, the following research question was pursued: how might a multinomial probability model be used to quantitatively analyze a small group (*N*), e.g., participants in a summer computer science camp?

Research Methodology

General Background

For the past several years, people have noted a need for informal education about science, technology, engineering, and mathematics (commonly called STEM), especially computer science and computer engineering (NRC, 2015). Many educators have offered summer camps for school-aged children as one way to meet this need (Cui & Ng, 2021). Informal education by way of summer camps geared towards school-aged children has been used for more than a decade to raise awareness of computer science, engineering, and other STEM disciplines (Decker & McGill, 2019).

Summer camps often have the purpose to raise awareness in children about programming through robotics (Amo et al., 2021; Anwar et al., 2019; Darmawansah et al., 2023) or about computational thinking through subtler means such as unplugged activities (Chen et al., 2023). In addition to discipline-specific knowledge and skills, broader 21st Century Skills such as communication and collaboration have been integrated into STEM camp, as well as foundational skills common to multiple disciplines, such as computational thinking (Kim et al., 2019; Wiebe et al., 2020; Wing, 2006). Camp objectives may articulate computational thinking skills explicitly by way of programming activities, e.g., robotics (Yilmaz Ince & Koc, 2021). There is also an implicit approach, that of unplugged activities (Delal & Öner, 2020; Zapata-Ros, 2019), which may integrate with other STEM disciplines and take a truly integrated disciplinary approach.

The effectiveness of informal summer camps may be measured qualitatively or quantitatively. Learning gains, interest, attitude, motivation, and other effects have been measured in campers (Decker & McGill, 2019; Hammack et al., 2015; Kong et al., 2014) using interviews, field note observations, and other methods that yield qualitative, interpretive results and usually describe a process of change. On the other hand, researchers may choose to quantify or quantitatively analyze the overall change from the impact of the camp experience itself, which is typically done by a pretest of initial status, followed by the camp experience (treatment), followed by a posttest. The pre-and post- are examined for meaningful change in participants, as evidenced through their responses. One of the tools often used in evaluations of educational methods is a questionnaire, either a previously developed instrument or one created for the camp, typically consisting of a small number of categories to choose from, like the 3- or 5-point Likert scale example described in more detail in later sections. Responses from the questionnaires are analyzed quantitatively to measure change from pre- to post-, and whether the change was meaningful (e.g., statistically significant) to support the claim that the treatment (e.g., camp) was effective.

Pearson's Chi-Squared Model

Specifically, the goal of pre/post survey quantitative analysis is to determine whether the observed numbers in each category of the Likert scale survey before the treatment is applied

ISSN 1822-7864 (Print) ISSN 2538-7111 (Online) https://doi.org/10.33225/pec/24.82.507

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 82, No. 4, 2024

(the class lesson or the camp activity) are or are not significantly different from the observed numbers afterward. This is an exercise in statistical inference. The most common method of statistical inference relies on the value of one or more statistics, generated by an experiment, as a measure of evidence. One such statistic is Pearson's Chi-squared, which is commonly used for survey research. It is simple to calculate, and its significance levels are widely available in table form.

Many textbooks written from 1950 to 2015 aimed at teachers or education researchers suggested tools for categorical data testing, such as counts per category of a Likert scale. The textbooks frequently took one of three approaches. First, textbooks meant to examine the operational aspects of testing and measurement in the classroom, administering quizzes, tests, and assignments did not include sections on statistical analysis. Even though the textbooks may have had the word *measurement* in the title, they lacked mention of quantitative analysis methods. Many did not use the word *statistics* at all. Other textbooks took an introductory approach intended to explain the most basic aspects of descriptive statistics to the working teacher. The textbooks were generally written from pre-1970 to 2009. The quantitative analysis techniques concentrated on the bell curve, the standardized *Z* statistic, its distribution, and so forth. The Chi-Squared statistic was mentioned in a cursory manner, if at all. Finally, some textbooks took a *modeling* approach, described as making future predictions based on the observations issuing from an educational treatment, e.g., a class lesson or a camp activity. Modeling requires advanced statistical tools such as least sequence regression or logistic regression that are beyond introductory textbooks in statistical inference. Thus, teachers and education researchers may consider taking a modeling approach as beyond the scope of evaluating various aspects of camp, including overall effectiveness.

The focus of many graduate level education research textbooks was to teach the use of statistics using a software package such as SPSS and applying this to numerous examples (Field, 2013). Avoiding the phrase *statistical inference*, the textbook instead discusses *statistical significance,* which Field defined as a process that is a blend of Ronald Fisher's idea of computing probability to assign weight of evidence with the Neyman-Pearson weighing of competing hypotheses (Field, 2013). Other textbooks used concepts such as confidence intervals, p-values, or significance levels.

Only a few education research textbooks identified the multi-category Likert as an instance of a multinomial distribution. The only simple statistical test, other than modeling, suggested for testing multinomial distributions is Pearson's Chi-Squared in a *goodness of fit* mode. The usual path recommended in education research textbooks to follow is thus:

1. Decide on a sampling statistic; Pearson's Chi-Squared in this case.
2. Somehow determine a set of probabilities per category that define the null hypothesis: ($p_i \in (0,1);\ i = 1 \ldots k;\ \Sigma_i^k = 1$). These may be available from some theoretical consideration or from the before treatment survey itself.
3. From everything known ***a priori***, decide on a level of significance for the *test* of the after-treatment survey being not consistent with the null hypothesis. This value of alpha (α) is typically .01 to .05 and describes the fraction of the probability distribution of the test statistic equal to a significant event or an event "more extreme". One might use the 5% level if there is already some evidence supporting efficacy.
4. Calculate the test statistic and decide if it reaches or exceeds the significant value. If it does, conclude the treatment having an effect and evidence for rejecting the null hypothesis. If it falls short of the significant value, decide instead that there is no evidence to reject the null hypothesis that before and after treatments are the same (Meyer, 1970).

https://doi.org/10.33225/pec/24.82.507 ISSN 1822-7864 (Print) ISSN 2538-7111 (Online)

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 82, No. 4, 2024

Limitations of Pearson's Chi-Squared Method

From the previously mentioned procedure, one might conclude that the Chi-Squared tool is appropriate to analyze or test Likert survey results scientifically. There are some complications to taking the Chi-Squared approach, however. First, a distinction needs to be drawn between Chi-Squared, which is a family of probability distributions depending on two parameters (α and β) and for which probability versus deviation can be calculated exactly, and Pearson's Chi-Squared, which is the sampling statistic in play. Pearson's Chi-Squared is a calculation from observations. Chi-Squared is a theoretical distribution that can be found in a table. They become one and the same with a large enough sample size. Pearson's Chi-Squared is defined as $D^2 = \sum_i^k (O_i - E_i)^2/E_i$; where O_i is the observed number of responses per Likert category (i runs from 1 to 5 in a typical case) and E_i is the expected number from the null hypothesis. How to determine E_i is set aside until further in the article when an example is shown.

As scholars describe the situation, the true probability distribution of D^2 is very complicated, but when the number of observations (*N*) becomes large enough, the continuous Chi-Squared distribution provides a good approximation for D^2 (Meyer, 1970; Papoulis, 1990; Snedecor & Cochran, 1967). The usual advice on this matter is that if the expected frequency per category is at least 5 for all categories, then Chi-Squared is an adequate approximation. This advice is legion, nearly a tradition, and the advice appears to refer directly or indirectly to a 1952 paper by W.G. Cochran. What Cochran had to say on this matter is not so simple.

Cochran explained that the rule of thumb at the time was 5 or maybe even 10, but these values were pulled from a hat as it were. He then set out to "appraise the performance of the tabular approximation in the borderline region between statistical significance and non-significance" (Cochran, 1952, p. 328). That is, he set out to determine the disturbance or departure between the true distribution of D^2 and a table of Chi-Squared values used for convenience. As he stated, "A disturbance is regarded as unimportant if when the P is 0.05 in the Chi-Squared table, the exact P lies between 0.04 and 0.06, …" (Cochran, 1952, p. 328). One can see a difficulty with the Chi-Squared approach immediately. Suppose one sets a 5% significance level ahead of time, as statistics books instruct, and subsequently finds a Chi-Squared approximation for D^2 near 4%. This person now rejects the null hypothesis and publishes the statistically significant results. Later, someone else fetches data from the supplementary materials for this paper and using an exact calculation for the multinomial distribution, or a Monte Carlo calculation, they find the actual value of D^2 is undoubtedly 6%, thus establishing an inconsistency with the interpretation of the data. A full reading of Cochran's paper reveals that the circumstances under which a table of Chi-Squared can substitute for a knowledge of the true percentile of D^2 is complex and depends on several parameter values and factors, not just the sample size.

Multinomial Probability Model

The multinomial probability distribution makes a reasonable model of analyzing a questionnaire with Likert scale responses. It is a joint probability distribution with as many variables as there are response categories. It is a generalization of the binomial model for sets of two-choice outcomes such as pass/fail, yes/no, and so forth, but allows for more than two outcomes. The Likert scale responses in this study involved five possible outcomes. Each possible outcome is, itself, a binomial process. For example, the extreme category *completely true of me* is seen as a binomial outcome when paired against its alternative *less than completely true of me*.

In the simplest possible view, regard a participant responding to a question as an extension of a Bernoulli trial, an extension resulting in a count for one of k categories rather than a dichotomous outcome like pass/fail. Let $x_1,...,x_k$ be the observed counts per category in a

ISSN 1822-7864 (Print) ISSN 2538-7111 (Online) https://doi.org/10.33225/pec/24.82.507

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 82, No. 4, 2024

questionnaire, and $p_1,...,p_k$ as the associated fixed probability of each. Then, the probability that N respondents result in data of $(x_1,...,x_k)$ is:

$$P(x_1,x_2,...,x_k|p_1,p_2,...,p_k)=\left(\frac{N!}{x_1!.x_2!...x_k!}\right)p_1^{x1}\,p_2^{x2}\cdots p_k^{xk}$$

While each participant may choose independently and may choose any of the k options for any reason at all, or no reason, without the intervention of the treatment it is assumed they would choose consistently. There are potential problems with this assumption. Participants may conspire to choose similarly, thus hampering independence, and they may choose two options, or skip the question completely, or even supply a hand-written fractional choice. These censored responses can be dealt with in a variety of ways, such as deleting incomplete records entirely, imputing values, or predicting what the censored values would likely be (Edwards, 1992).

While the multinomial is a generalized binomial distribution, more insight is gained by viewing it as a collection of *k* Poisson processes each with its own rate p_k of being chosen in each trial. This provides a simple explanation about the relationship of Pearson's Chi-Squared to the Chi-Squared distribution. After *N* trials the *kth* category is expected to have $\mu_k = Np_k$. If *N* is large enough, then the difference $(x_k-\mu_k)$ follows a normal distribution. The variance of the Poisson distribution is Np_k itself. Thus $(x_k-\mu_k)^2/\mu_k$ follows a Chi-Squared distribution if *N* is large enough. According to Engels (n.d.), however, *N* is often not large enough, or the circumstances under which *N* is large enough to depend on more than just *N* (Cochran, 1952; Taroni et al., 2010). What would work better is to have a robust test method for small *N*.

Likelihood

Having chosen a probability model, a random process, to analyze responses to a questionnaire with Likert scale responses, a statistic drawn from it to use for significance testing is needed. Likelihood is a conditional probability, P(A|H); the probability that proposition A is true given the conditions in the hypothesis being tested, H. The conditions making up H are the probabilities of an individual respondent choosing each of the Likert scale categories. Likelihood differs from probability in the sense that probability focuses on the probability of A conditioned upon the truth of H, whereas likelihood focuses on the observation A as evidence for the hypothesis H (Royall, 1997). One does not use P(A|H) directly as evidence, but rather the ratio of this conditional probability as applied to two different propositions A, the probability of the observed frequency of responses per category after treatment and B, the probability of the observed frequency of responses before. One may also use the logarithm of the Likelihood Ratio (LLR).

An attractive aspect of the likelihood ratio as a measure of evidence is that it depends only on what was observed or measured during an experiment or study. Royall (1997) presented a strong case for likelihood ratio being the only logical and consistent measure of evidence among any of the competitive measures such as p-value, confidence intervals, significance level, and so forth. In the case of this study, it is additionally attractive that likelihood may be calculated exactly from the multinomial distribution for arbitrarily small cohorts.

Sample

In 2022, ten campers participated in camp, and nine were able to complete both pre- and post- questionnaires. The camp was funded again during the summer of 2023, and eight campers participated and completed both pre- and post-questionnaires. Thus, $N = 9$ in 2022, N

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 82, No. 4, 2024

= 8 in 2023, and total for both years of camp $N = 17$. Attitude (affect) towards computer science and a 21st Century skill of teamwork, composed of trust and consensus building of campers in small groups, were measured through administering two questionnaires. The questionnaires were intended to measure high interaction in small groups and an appreciation for blockchain technology and computational thinking within the umbrella of computer science discipline (Kim et al., 2019; Wiebe et al., 2020). The research was approved by the Institutional Review Board. Campers provided assent and parents/guardians provided consent for the study. A pre/post approach was used to measure each issue and quantitatively analyze if notable change had occurred.

Instrument and Procedures

Two questionnaires were administered to campers, both measuring the effect and both administered on the first hour of the first day of camp (pre-test) and on the last hour of the last day of camp (post-test). The questionnaires were used with permission and were designed for a middle school audience (Cantrell & Andrews, 2002; Rachmatullah et al., 2020). One questionnaire measured attitude towards computer science in a series of 12 questions with 5-scale Likert responses from strongly disagree to strongly agree including a neutral point (Rachmatullah et al., 2020). The other questionnaire focused on the camper's opinions toward group work. It posed 30 questions in five categories with 5-point Likert responses from *not at all true of me* to *very true of me* with a neutral point (Cantrell & Andrews, 2002).

Data Analysis

Campers completed the questionnaires. Each pre- and post-test were paired. The non-paired questionnaires were removed from the data analysis. Because of the low number of campers and correspondingly low number of paired tests (N), the analysis of the questionnaires was complicated by two problems. First, some categories in the questionnaire were empty – they contained zero response. Second, even when some categories were not empty, all the categories fell far short of the rule of thumb for using Chi-Squared to test the uniformity of pre/post responses. This caused a search for alternatives to Chi-Squared. It was determined the solution lay in considering alternatives to the Chi-Squared approach, namely using exact calculations of the multinomial probability model.

Research Results

A question from the Feelings Toward Group Work (GW) survey used in this study provides an example of using statistical reasoning to arrive at conclusions when the cohort is too small to use Chi-Squared.

GW27: "I rarely feel relaxed within a group."

First, one must establish the assumptions in which the inference is conditioned. There is some reason for believing that the treatment, e.g., the camp experience, will have some effect on responses to this question from pre to post. Assume that there is no difference between Jack, who chooses response 3 and Jill who chooses 4. They are both simply trials that result in a count in one of the survey categories according to a probability for that category (p_i); assume that before the treatment each of the responses has some probability of being chosen and while it is unknown what these probabilities are (they will be estimated from the observations), assume that in the absence of the treatment (camp), these would remain fixed. The counts themselves are a sample which could change from one group of respondents to another, i.e., the counts are random variables.

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 82, No. 4, 2024

Next, estimate the probabilities of students choosing one or another of the five Likert categories per question. The obvious path is to use the before treatment response frequencies themselves as these probabilities. However, this leads to the possibility of a zero count in one or more categories leading to a zero probability; an issue made more acute with small cohort numbers.

A Bayesian approach was used to update probability with observations. An uninformative prior guess as to the probabilities is a guess that quickly fades into irrelevance with increasing new information from the pretreatment survey. The Dirichlet distribution provides an uninformative prior. It consists of a frequency per category of a uniform value of one (Johnson & Christensen, 2012). The value is updated with observed frequencies from the pretreatment survey. The probability found per category (i) is $p_i = (1+f_i)/(k+N)$ where f is the frequency of response in the i^{th} category, k is the number of categories, and N is the total number of respondents. Probability so found produces a maximum likelihood close to the observed frequencies and solves the problem of avoiding zero probability.

Table 1 shows the data collected as pretreatment during summer camp for this statement in 2022 and 2023. The per category probabilities based on 2022 responses alone, and then on pooled responses for both years are included. Finally, the likelihood ratios of year 2023 responses based on either set of estimated probabilities are shown in the final row. The LR values outside the range of the null hypothesis (⅛ to 8) would be representative of strong evidence of a difference. Any LR values falling within this range are too weak to reject the null hypothesis that the year 2022 and year 2023 responses are the same. Note in Table 1 that neither estimated probability results in an LR smaller than or equal to ⅛ nor larger than or equal to 8. Thus, there is no evidence that the belief of students entering camp in the year 2023 is different from those who entered in the year 2022.

Table 1

Pretreatment Survey Results Testing Responses to Statement GW27- "I rarely feel relaxed within a group." Comparing Year 2022 to Year 2023.

Category	2022 Responses	2023 Responses	P_i using only 2022 frequencies	P_i using pooled frequencies
Not at all true of me	1	3	0.143	0.227
Partly not true of me	3	1	0.286	0.227
Neutral	3	2	0.286	0.273
Partly true of me	1	1	0.143	0.136
Very true of me	1	1	0.143	0.136
LR (2022/2023)	N/A	N/A	2.6	0.819

What has been established at this point is that there is no evidence of changing pre-treatment views of campers between the year 2022 cohort and the 2023 cohort. Thus, it makes sense to use pooled data to estimate probabilities. If the pre-treatment views of campers differed between years, then there is an indication that the campers are not homogeneous, e.g., they may have participated in an event that altered their initial perceptions about group work and computer science.

Using pooled estimates of P_i, analysis proceeded to examine evidence for or against the proposition GW27 having changed pre- to post-treatment. Table 2 shows the results. The probabilities in the far-right column are calculated using the pretreatment frequencies and result in the LR of 12.1. The LR value of 12.1 shows strong evidence that post-treatment frequencies

https://doi.org/10.33225/pec/24.82.507 ISSN 1822-7864 (Print) ISSN 2538-7111 (Online)

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 82, No. 4, 2024

are very unlikely to arise from the null hypothesis; that is, shows strong evidence that the treatment has altered the categories to be more heavily weighted toward categories 3, 4 and 5. After having engaged in the camp activities, participants perceive themselves as less relaxed within a group.

Table 2
Pooled (2022+2023) Frequencies Before and After Treatment.

Category	Pooled Pre-treatment Responses	Pooled Post treatment Responses	P_i using pooled frequencies pretreatment
Not at all true of me	4	1	0.227
Partly not true of me	4	4	0.227
Neutral	5	7	0.273
Partly true of me	2	1	0.136
Very true of me	2	4	0.136
LR (Before/After)	N/A	N/A	12.1

It is interesting at this point to calculate how such results should alter one's beliefs about students' view of the treatment (camp). How one should rationally update beliefs based on Bayes theorem is:

$$P(H|A) = \left(\frac{P(A|H)}{P(A)} * P(H)\right)$$

H could stand for either H_0, the null hypothesis that the probabilities per category have remained at the pretreatment values, or H_1, the alternative that probabilities have shifted to values represented by the posttreatment results. Event A is the observed frequencies posttreatment. The total probability of event A, P(A), is composed of only two observations because that is all that the study contains. Thus, $P(A) = P(A|H_0)P(H_0)+P(A|H_1)P(H_1)$, and rearranging produces

$$P(H_0|A) = \left(\frac{P(A|H0)*P(H0)}{P(A)}\right) = \left(\frac{P(A|H0)P(H0)}{P(A|H0)(P(H0)+(P(H1)P(A|H1)}\right)$$

Assume that one judged a priori even odds that the treatment would have an effect. Thus, the a priori belief in the probability of H_0, which is to say $P(H_0)$, is 50%. But $P(A|H_1)/P(A|H_0))$ is the LR as calculated in the fourth column of Table 2. And this modifies the a priori belief that $P(H_0)=P(H_1)=.5$. Thus, in this case:

$$P(H_0|A) = \left(\frac{0.5}{0.5+0.5*12.1}\right) = 7.6\%$$

The observation of frequencies posttreatment (event A) has decreased belief in H_0 from 50% to 7.6%. It is almost a certainty (92.4%) that the probabilities of the null hypothesis (H_0) no longer apply after the camp activities. The observation of event A is far more likely (12.1 times more likely) with the probabilities of H_1 than with the probabilities of H_0.

Discussion

The ability to quantify differences using likelihood ratio applied to the multinomial distribution, enabled one to make a few determinations about year 2022 and 2023 camp

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 82, No. 4, 2024

experiences. First, regarding pre-camp attitudes toward group work, the 2022 and 2023 cohorts looked to have come from a homogenous group with no significant differences between them. This made it possible to combine the two into a larger group for analysis, which improved the resolution of pre-camp to post-camp changes. Regarding attitudes toward computer science, the second of the two questionnaire surveys, the only significant differences pre-camp appeared to be related to females, being more numerous in the 2023 cohort, also being somewhat less sure initially about the usefulness of computer science to their future careers.

Second, quantitative results substantiated only a few significant changes in student attitudes regarding group work or computer science pre-camp to post-camp. Specifically, by the end of camp, campers felt more neutral to negative about comfort working in a group, and communicating with other group members, from their perceived comfort pre-camp. Campers felt less positive about enjoying group work and its effectiveness. The responses to these items indicate a tempering of views that group work runs smoothly and that all members contribute.

Of the nine-item questionnaire, two items showed differences between 2022 and 2023 regarding initial attitudes toward computer science. These differences may possibly result from female campers being more reserved about their programming abilities at the beginning of camp. Campers may also be more cautious about the usefulness of programming for their future work pre-camp.

The results of the pooled responses from campers pre- to post- in 2022 and 2023 showed a change in campers' attitudes toward computer science regarding one item. The item stated *I would like to use creativity and innovation in my future*. The change in attitude to *somewhat true of me* or *very true of me* may pertain to activities done in camp that emphasized creativity.

Taking a step back to reflect on using an alternative approach to quantitative data analysis, the authors suggest that using this approach strengthened the conclusions that campers tempered their initially high enthusiasm with the realism of challenges associated with computer programming and teamwork skills. Such insights may not have had the support if purely qualitative evidence had been used such as field note observations, or they may not have been noted at all. From the authors' viewpoint, adding the quantitative analysis helped steer overall impressions of camp toward pinpointing how future camps emphasizing group work may be better structured. Overall, quantitative analysis provided an additional line of evidence to help *triangulate* the otherwise qualitative data.

Other Exact Methods

Recognition that the use of Chi-square seems both archaic and limited to large samples has led people to suggest alternative exact tests. Fisher's (1934) exact method, for example, is appropriate to 2x2 contingency tables. It is often called the *exact calculation of Chi-Squared* and is based on the hypergeometric probability model of sampling a finite number of objects without replacement. It's mentioned, often only in passing, as a calculation of the Chi-Squared test when n is a small number but never in circumstances beyond 2x2 contingency tables or independent samples with dichotomous outcomes (Field, 2013; Huck, 2012; Siegrist, 2022). However, the existence of more than two categories in the case of a Likert scale produces a contingency table larger than 2x2. In this case, an exact calculation of Chi-Squared requires a joint distribution generalized from the hypergeometric distribution.

Other exact methods exist. Basic combinatorial arguments can be used to derive the probability density function of the multivariate hypergeometric distribution. Details about this distribution are beyond the scope of this article but are available elsewhere (Engels, n.d., Siegrist, 2022).

Engels (n.d.), in another example of using exact methods, outlined two alternative exact tests, one of which is the log likelihood ratio without a particular null hypothesis in mind,

https://doi.org/10.33225/pec/24.82.507 ISSN 1822-7864 (Print) ISSN 2538-7111 (Online)

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 82, No. 4, 2024

the other being just the multinomial probability itself. Both can be calculated from routines available in the statistical package *R*. Resin (2023) illustrated an algorithm for exact multinomial calculation equivalent to the significance tests such as significance level, confidence intervals and so forth which uses an algorithm implemented in the package *R* and delivers a probability equal to the observed frequency counts and observations that would be more extreme were they to be observed. Some scholars mention using the log likelihood ratio (Resin, 2023). This is not an alternative analysis method but simply another way to measure the likelihood ratio.

Conclusions and Implications

Differences between pre-camp and post-camp surveys were analyzed using a 5-level Likert scale questionnaire modeled with a multinomial distribution and a likelihood ratio statistic. The comparison of a pre-camp survey to a post-camp survey revealed significant changes in campers' attitudes after completion of camp. This provided quantitative measures of significant changes in response in a setting where the more well-known Pearson's Chi-Squared could not be used because of the small sample size.

Using an exact calculation of the multinomial probability model with the use of likelihood ratio (LR) as a statistical test of the strength of evidence has numerous advantages. It is an exact calculation that can be used on any size cohort or sample. It depends only on what was observed during a study and not on any characteristics of a sample space. It does not require any advanced calculation. Practically all modern spreadsheets and statistical package programs (e.g., *Microsoft Excel* spreadsheet, *R* statistical package) have the capability to calculate what the analysis demands, providing only that packages can handle small-number factorials. Finally, the likelihood ratio, which is the measure of the strength of evidence, can be employed immediately in Bayes' theorem to see how prior beliefs about a hypothesis should be updated in the face of evidence.

Although this article outlines a summer camp example, the key point rests with the power of using quantitative analysis with a small population sample. It is posited that utilizing quantitative methods with small sample sizes, which would normally not be analyzed statistically, will strengthen insights into data trends and reveal implications for modifications that could have gone unnoticed. Multiple audiences could benefit from garnering the potential of a small size quantitative investigation.

Note

This work has been approved by the University of Wyoming Institutional Review Board (IRB) on August 19, 2022. Code: #20220819TK03373.

References

Amo, D., Fox, P., Fonseca, D., & Poyatos, C. (2021). Systematic review on which analytics and learning methodologies are applied in primary and secondary education in the learning of robotics sensors. *Sensors, 21*(1), Article 153. https://doi.org/10.3390/s21010153

Anwar, S., Bascou, N. A., Menekse, M., & Kardgar, A. (2019). A systematic review of studies on educational robotics. *Journal of Pre-College Engineering Education Research (J-PEER), 9*(2), Article 2. https://doi.org/10.7771/2157-9288.1223

Baek, Y., & Touati, A. (2020). Comparing collaborative and cooperative gameplay for academic and gaming achievements, *Journal of Educational Computing Research, 57*(8), 2110–2140. https://doi.org/10.1177/0735633118825385

Boddy, C. R. (2016). Sample size for qualitative research. *Qualitative Market Research: An International Journal, 19*(4), 426–432. https://doi.org/10.1108/QMR-06-2016-0053

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 82, No. 4, 2024

Cantwell, R. H., & Andrews, B. (2002). Cognitive and psychological factors underlying secondary school students' feelings towards group work. *Educational Psychology, 22*(1), 75–91. https://doi.org/10.1080/01443410120101260

Chen, P., Yang, D., Metwally, A. H. S., Lavonen, J., & Wang, X. (2023). Fostering computational thinking through unplugged activities: A systematic literature review and meta-analysis. *International Journal of STEM Education, 10*. https://doi.org/10.1186/s40594-023-00434-7

Cheng, M., Su, C.-Y., & Kinshuk (2021). Integrating smartphone-controlled paper airplane into gamified science inquiry for junior high school students. *Journal of Educational Computing Research, 59*(1), 71–94. https://doi.org/10.1177/0735633120953598

Chingos, M.M., & Whitehurst, G. J. (2011, May 11). *Class size: What research says and what it means for state policy*. Brookings. https://www.brookings.edu/articles/class-size-what-research-says-and-what-it-means-for-state-policy

Chou, P.-N. (2020). Using ScratchJr to foster young children's computational thinking competence: A case study in a third-grade computer class. *Journal of Educational Computing Research, 58*(3), 570–595. https://doi.org/10.1177/0735633119872908

Cochran, W. G. (1952). The Chi-Squared test of goodness of fit. *Annals of Mathematical Statistics, 23*(3), 315-345.

Cui, Z., & Ng, O.-L. (2021). The interplay between mathematical and computational thinking in primary school students' mathematical problem-solving within a programming environment. *Journal of Educational Computing Research, 59*(5), 988–1012. https://doi.org/10.1177/0735633120979930

Darmawansah, D., Hwang, G.-J., Chen, M.-R. A., & Liang, J.-C. (2023). Trends and research foci of robotics-based STEM education: A systematic review from diverse angles based on the technology-based learning model. *International Journal of STEM Education, 10*. https://doi.org/10.1186/s40594-023-00400-3

Decker, A., & McGill, M. M. (2019). A systematic review exploring the differences in reported data for pre-college educational activities for computer science, engineering, and other STEM disciplines. *Education Sciences, 9*. https://doi.org/10.3390/educsci9020069

Delal, H., & Öner, D. (2020). Developing middle school students' computational thinking skills using unplugged computing activities. *Informatics in Education, 19*(1), 1–13. https://doi.org/10.15388/infedu.2020.01

Delice, A. (2010). The sampling issues in quantitative research. *Educational Sciences: Theory and Practice, 10*(4), 2001-2018.

Edwards, A. E. F. (1992). *Likelihood*. Johns Hopkins University Press.

Engels, W. R. (n.d.). *XNomial—Exact test for multinomial*. https://cran.r-project.org/web/packages/XNomial/vignettes/XNomial.html

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). Sage Publications.

Fiorella, L., Kuhlmann, S., & Vogel-Walcutt, J. J. (2019). Effects of playing an educational math game that incorporates learning by teaching. *Journal of Educational Computing Research, 57*(6), 1495–1512. https://doi.org/10.1177/0735633118797133

Fisher, R. A. (1934). *Statistical methods for research workers* (5th ed.). Oliver and Boyd.

Fugard, A. J., & Potts, H. W. (2015). Supporting thinking on sample sizes for thematic analyses: A quantitative tool. *International Journal of Social Research Methodology, 18*(6), 669–684. https://doi.org/10.1080/13645579.2015.1005453

Hammack, R., Ivey, T. A., Utley, J., & High, K. A. (2015). Effect of an engineering camp on students' perceptions of engineering and technology. *Journal of Pre-College Engineering Education Research (J-PEER), 5*(2). https://doi.org/10.7771/2159-9288.1102

Huck, S. W. (2012). *Reading statistics and research* (6th ed.). Pearson.

Johnson, B., & Christensen, L. (2012). *Educational research* (4th ed.). Sage Publications.

Kim, N. J., Belland, B. R., Lefler, M., Andreasen, L., Walker, A., & Axelrod, D. (2019). Computer-Based scaffolding targeting individual versus groups in problem-centered instruction for STEM education: Meta-Analysis. *Educational Psychology Review, 32*, 415–461. https://doi.org/10.1007/s10648-019-09502-3

Kong, X., Dabney, K. P., & Tai, R. H. (2014). The association between science summer camps and career interest in science and engineering. *International Journal of Science Education, Part B, 4*(1), 54–65. https://doi.org/10.1080/21548455.2012.760856

Meyer. (1970). *Introductory probability and statistical applications* (2nd ed.). Addison Wesley.

National Center for Education Statistics (NCES). (n.d.) *NTPS tables*. https://nces.ed.gov/surveys/ntps/tables/ntps1718_fltable06_t1s.asp

Northrup, A. K., & Burrows Borowczak, A. C. (2023). Integrating computer science across Wyoming's K-12 curriculum from inception to implementation: Analysis using systems theory. *Computers in Education Journal*, *13*(2).

NRC, National Research Council (2015). *Identifying and supporting productive stem programs in out-of-school settings*. National Academies Press.

Papoulis, A. (1990). *Probability and statistics*. Prentice Hall.

Rachmatullah, A., Wiebe, E., Boulden, D., Mott, B., Boyer, K., & Lester, J. (2020). Development and validation of the Computer Science Attitudes Scale for middle school students. *Computers in Human Behavior Reports 2*, Article 100018. https://doi.org/10.1016/j.chbr.2020.100018

Resin, J. (2023). A simple algorithm for exact multinomial tests. *Journal of Computational and Graphical Statistics, 32*(2), 539–550.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. Chapman & Hall.

Siegrist, K. (2022, April 24). *The multivariate hypergeometric distribution*. LibreTexts Statistics. https://stats.libretexts.org/Bookshelves/Probability_Theory/Probability_Mathematical_Statistics_and_Stochastic_Processes_(Siegrist)/12%3A_Finite_Sampling_Models/12.03%3A_The_Multivariate_Hypergeometric_Distribution#:~:text=The%20Joint%20Distribution,-Basic%20combinatorial%20arguments&text=Recall%20that%20since%20the%20sampling,size%20n%20chosen%20from%20D.&text=The%20distribution%20of%20(Y1,mk)%2C%20and%20n

Snedecor, G. W., & Cochran, W. G. (1967). *Statistical methods* (6th ed.). The Iowa State University Press.

Talafian, H., Moy, M. K., Woodard, M. A., & Foster, A. N. (2019). STEM identity exploration through an immersive learning environment. *Journal for STEM Education Research, 2*, 105–127. https://doi.org/10.1007/s41979-019-00018-7

Taroni, F., Bozza, S., Biedermann, A., Garbolino, P., & Aitken, C. (2010). *Data analysis in forensic science: A Bayesian decision perspective*. John Wiley and Sons.

Yilmaz Ince, E., & Koc, M. (2021). The consequences of robotics programming education on computational thinking skills: An intervention of the Young Engineer's Workshop (YEW). *Computer Applications in Engineering Education, 29*(1), 191–208. https://doi.org/10.1002/cae.22321

Wiebe, E., Kite, V., & Park, S. (2020). Integrating computational thinking in STEM. In Johnson, C. C., Mohr-Shroeder, M. J., Moore, T. J., & English, L. D. (Eds.), *Handbook of Research on STEM Education*, (pp. 196–209). Routledge.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM, 49*(3), 33–35.

Zapata-Ros, M. (2019). Computational thinking unplugged. *Education in the Knowledge Society, 20*, 1–29. https://doi.org/10.14201/eks2019_20_a18

Received: *July 05, 2024* Revised: *July 23, 2024* Accepted: *August 03, 2024*

PROBLEMS
OF EDUCATION
IN THE 21st CENTURY
Vol. 82, No. 4, 2024

Cite as: Kilty, T. J., Kilty, K. T., Burrows Borowczak, A. C., & Borowczak, M. (2024). Quantitative techniques with small sample sizes: An educational summer camp example. *Problems of Education in the 21st Century*, *82*(4), 507–520. https://doi.org/10.33225/pec/24.82.507

Trina Johnson Kilty *(Corresponding author)*	PhD, Post Doctoral Researcher, Department of Electrical Engineering and Computer Science, University of Wyoming, Laramie, WY 82071, USA. E-mail: trina.j.kilty@gmail.com ORCID: https://orcid.org/0000-0002-8713-8234
Kevin T. Kilty	PhD, Lecturer, Department of Mechanical Engineering, University of Wyoming, USA. E-mail: kkilty1@uwyo.edu ORCID: https://orcid.org/0000-0002-9768-0676
Andrea C Burrows Borowczak	EdD, PhD, Professor and Director of the School of Teacher Education (STE), College of Community Innovation and Education (CCIE), University of Central Florida, USA. E-mail: Andrea.Borowczak@ucf.edu ORCID: https://orcid.org/0000-0001-5925-3596
Mike Borowczak	PhD, Associate Professor, Department of Electrical and Computer Engineering, University of Central Florida, USA. E-mail: mike.borowczak@ucf.edu ORCID: https://orcid.org/0000-0001-9409-8245