

Ib extended essay statistics



**ASSIGN
BUSTER**

I. Abstract In the study of mathematical statistics, there are different ways that a set of data can be proved to be a significant set. However, the ways that I am going to prove if the data is significant can be done in two different ways by using the z-proportion test and the Chisquare test (χ^2). These two significant tests can prove if a set of data is either significant or not significant. What is interesting about the three tests is that they are done in a very similar way yet end up with different results in regards to significance.

In order to prove this thesis, I did an experimental test on the students at my high school. This experiment was done in order to create a sample problem. I will continually use this problem and it will be rephrased, reused, and integrated to represent the test being used. In this paper, the three tests' formulas and development will be shown to gain an overall understanding of how the tests are similar. II. Creation of the Sample Problem A self conducted statistical experiment was done twice on same populations and data was collected and will be used as a sample question and integrated through this paper to their respective significance test. The question was conducted with a double-blind experiment in randomly selected classrooms. Double blind represents that both the experimenter and those receiving the soda have no idea on what soda they had received, therefore, eliminating the bias in the conducted experiment help write my essay . Each building in Charter Oak High School has a letter on them (A, B, C, etc.). Firstly, all building letters on the Charter Oak High School was placed into a hat, then picked out and recorded.

Then, a number from 1-9 was placed into the hat, representing the classroom number. After this process, the room in one of the buildings was

randomly selected to have an experiment conducted in their class for example, D-4. Two selected volunteers in the class would give the students cups with letters representing the soda brand (A and B). After the experiment was done, the volunteers revealed that soda A was the branded Coca Cola [®] and that soda B was the non branded Safeway Cola [®]. The research question derived from the self conducted and collected experiment is of the simple question of ??? Do you prefer a soda A or soda B??? The sample was collected from a data of 150 randomly selected students. The second data was "collected in the same format but however, had a population of 157 randomly selected students. I tested students in two periods combined (IE: Period one and two made up sample one). The data is represented in Table 1.

1 (below) and will be used through out the paper. Table 1. 1 Coca Cola (A) Safeway[®]'s Cola (B) Data Set 1 68 82 Data Set 2 73 84 III. Nature of a Significance Test What is a significance test The term significance can imply many different meanings. In statistics, the definition of significant means that what one is researching is probably true and not due to a chance (Creative Research Systems).

In other words, there are many things that can cause a research to be true??" a lurking variable. Lurking variables are hidden factors that could possibly contribute to the significant outcome of the question. The purpose of a significance test is to test the implications in which the researcher desires to explore (Cox 30). In statistics, there are many different tests that can prove if something is significant or not. There are many different ways to prove if something is significant, however the two branches that I am using <https://assignbuster.com/ib-extended-essay-statistics/>

in this paper is the z-"proportion test and the Chi-Square test. Of course, there are many different tests that could be used to see if something is significant.

However, these two branches are the most closely related in regards of proportional testing, interestingly, are done in different mathematical procedures. The statistical inference provides methods for drawing conclusions about a population from the sample data provided in table 1. 1. It provides the researcher with a mathematical understanding about a sample population through the probability of the event occurring.

IV. The Test Hypotheses The most commonly used ??? test percentage??? in the two tests is: α (theta) = 5%. ??? This means that the finding has a 95% chance to become true??? (Creative Research Systems). The ??? hypothesis is nothing more than a guess or an assumption??? (Megeath 181). People in everyday lives make their own hypotheses and test them out by doing so one comes to the conclusion of either acceptance or rejection of the hypothesis (Megeath 181).

The most common representation for the null and alternative for proportional hypotheses is represented in the next page. Hypothesis Set 1. 1 Null: $H_0: p = p_0$ Alternative: $H_a: p < p_0$ Or $H_a: p \neq p_0$ The null hypothesis is the statement being tested in a test of significance and is designed to assess the strength of the evidence against the null hypothesis.

The null hypothesis is usually in the statement of, ??? there is no difference??? or, ??? there is no effect??? (Yates 582). In contrast, the alternative hypothesis states that there is no difference or change that is

significant (Yates 582-3) and can be represented in three different ways of which it can be written according to if it is a one sample or two sample population test (sided). There are a number of ways in which these two hypotheses could be phrased and written with the greater than and less than symbols as shown above in formula set 1.

1. It is also used for the various types of significance tests that statistics incorporates. The conclusion of the hypothesis depends on the P-value computed by the formulas for the z-proportion and chi-square test. The P-value is computed by assuming that the null hypothesis is true and that the observed will take a value as extreme as or more extreme than observed (Yates 581-2). The smaller the P-Value is, the stronger the evidence is against the null hypothesis. Example 1 (From using the data): Coca Cola [®] was determined to prove that their soda was better than any other soda out in the market. The results proved that an average of 68 people preferred the Coca Cola [®] brand above those that preferred Safeway[™]'s Cola [®] brand in Orange County, CA. Questioners still believe that there are more people who like Coca Cola than that of the general population.

. $H_0 = p = 0.50$: The mean value (average) of people who like Coca Cola is 68 per county. $H_a = p > 0.50$.

50: The mean value (average) of people who like Coca Cola is greater than 68 people per county. Note the set up for the hypothesis. V. Requirements and Nature of the Z-Proportional Test Formulas and Requirements for a One Sample Z-Proportional Test The z-proportional tests have only three

conditions to fulfill for it to be the test. This test is supposed to find a population's favor in a specific interest, in this paper's case, soda.

There are three conditions that are needed to meet the interests of this test.

These conditions will be exemplified later when the nature of the proportion test is done."Conditions 1. 1 a) The sampling design is close to a simple random sample (easily done). b) The number of the people tested in the population is much larger than the sample size (n).

1. $n(10) = x$ amount c) The counts of answers to the question is larger than 10. ? 1. $(n)(p) > 10$? 2. $n(1-p) > 10$ (Yates 688) Formulas 1.

1 ? a) Finding the p . ? $p = \frac{\text{Count of successes in the sample}}{\text{Count of observations in the sample}}$ b) Computation of the P-Value to test the hypothesis Standard Error (SE) $SE = \sqrt{\frac{p(1-p)}{n}}$ We need a simple random sample because it does not allow a bias in the collection of the data or process of the experiment. We also need the number of people in the population so we are sure that it is a sample of the population and not of the whole population. The counts of the numbers or sample of that p of the population are able to be calculated with a calculator. If the sample is not equal to ten, then it is most likely not a proportional test. The inferences for a population proportion can also be seen on a graph in which shows the distributions own approximate P-value against the H_0 as show in graphs 1. 1 to 1.

3. Graph 1. 1 $H_a: p > p_0$ is $P(Z > z)$ Graph 1. 2 $H_a: p < p_0$ is $P(Z < z)$ Graph 1. 3 $H_a: p \neq p_0$ is $2P(Z > |z|)$ These graphs represent the limit of which the p is being tested in which they have a normal distribution.

When the hypothesis is a "less than" p test, then it is in the higher 5% of the inference in order for it to be true and vice-versa for the "greater than" p test. We usually use the 5% rejection area because it is the most common significance percent testing level. However, the test for the two sides is multiplied by two because you are testing if it falls into the unshaded region, less than 10%. Nature of the One Sample Z-Proportional Test

In showing of how the test is done, I chose to use the problem that I personally created from my self-experiment at school, the interest in either soda A or soda B. The following example is shown with data set 1. To start off the problem, we have to first integrate the problem for the Z-Proportional test.

The sample problem being used will be Sample Question 1. 1. Sample Question 1. 1 Students in Charter Oak High School decided that branded sodas are much better than non-branded sodas.

However, I did not agree with their conclusion and decided to create my own experiment (explained in section 2 of the paper) to prove their hypothesis wrong. Out of 150 students, only 82 people liked the un-branded soda, Safeway's Cola. Is this data significant when $\alpha = .10$ that more than a half of people actually do like non-branded sodas over branded sodas After reading the problem, we need to form the hypothesis in which would prove that there is a greater amount of people that like non-branded over non-branded. Let's calculate the p first and then create our null and alternate hypothesis.

The formula to figure out the p is shown in formula 1.1 in page 4 of this paper. $p = 85 / 150 = 0.57$

Hypotheses $H_0: p = 0.50$, there is only a 50% chance that people like non-branded sodas. $H_a: p > 0.50$.

50, there are more people that like non-branded sodas than branded sodas and has a greater chance. After setting up the hypotheses, we have to now create the conditions to make sure it is usable for the one-sample Z-Proportional test. Conditions a) The sample is a simple random sample (SRS) because we are only picking numbers and letters from a box.

Also, we are only testing with two variables in an easy question experiment.

b) The sample size of students in Charter Oak High School is greater than 1,500 students. a. $150(0.50) < \text{Total Amount of Students in Charter Oak High School}$ c) Counts of the population is larger than 10 (the minimum requirement for a zproportional test).

a. $150(0.50) > 10$ b.

$150(1-0.50) > 10$ Yes Yes Since all the conditions meet the rules for a Z-proportional test, we can now do the test itself to see if the data is significant enough to prove that the alternate hypothesis is correct or not. Look at formula 1.

1 in order to figure out the correct setup for this test. In order to make things easier, let's calculate the standard error first (SE_x) as circled in formula 1.1.

$SE_x = 0.5 \sqrt{1 - 0.5} / \sqrt{100} = 0.041$ Now, we plug it into formula 1.1 B. $Z = 0.57 / 0.041 = 13.9$

$$57 \cdot .050 = 1.7070.$$

041 After calculating the Z value, we plug it into a calculator in order to find the correct P Value to prove if the hypothesis is significant or not. Since I personally use a TI-83 Plus, the process will be explained in regards to this calculator model. TI-83: ??? Choose 2nd + Vars ? 2: Normalcdf(??? Place in the right values into the equation of Normalcdf(1.707, 100) ??? Then hit enter ??? The value should be 0.0439 We choose 100 as our value after normalcdf because it is the maximum value in the inference chart in which we place in. After placing the value into the calculator, we calculated that the P-value for this problem is approximately 0.0439. Since the P-value is lower than .

10, we can conclude that there is more than half of the school's population that likes un-branded sodas (Safeway's cola). We use the 10% test interval because it is the most commonly accepted test level. If anything is lower than 10%, then it is considered significant" the lower the value, the more significant" therefore, we can come to the conclusion above. If any value is lower than .10, then we conclude that it is significant. However, if the value was over .10 (IE: 0.

471), then we would conclude that there is no significant data that concludes that more than half of the students like un-branded sodas. It is the probability of ??? getting a difference as large as ??? the sample p-hat is ??? no greater than [.10] ??? or the probability of " ??? getting a difference as large as this is no greater than .10 ??? (Phillips 120). Therefore, it is necessary for us to use a small interval percent such as a 10% or 5%. VI.

Properties and Nature of the Chi-Square Test (χ^2) Formulas and

Requirements for a Chi-Square Test There are many ways that chi-square tests are differentiated. Some characteristics that describe a chi-square test: Continuous distribution Distribution of the sums of squared normal random variables. (only takes positive values) Has a right-skewed distribution. The graphical shape of the chi-square test is dependent on the degrees of freedom (df).

o Refer to graph 1. The mean of a chi-square distribution is equal to its df. Note: the graphs shift to the right as the degrees of freedom increases. To find the standard deviation, the formula is: $\sqrt{2df}$ (Mulekar 245) (Formula 2. 1C) "Chi-square tests are usually done in tables such as the one in table 1. 1 (page 2). They are commonly called the chi-square test for the goodness of fit.

Goodness of fit is defined as a test that can be applied to see if the observed sample distribution is significantly different from the hypothesized population distribution (Yates 728). It is almost similar to the z-proportional test in the sense that it also has something to do with proportional testing. On top of this common similarity, the hypothesis set up is also similar to the z-proportional test in which you would find the p through proportional division or given a percent in the problem. Just a reminder the null hypothesis is to draw a conclusion that there is no difference among the proportions and the alternative is to test that there is some difference but not all (Yates 747). Usually, chi-square tests are done on two way tables in which columns and rows are added together and divided to find expected counts for each.

A two way table is ??? a table that gives counts for both successes and failures??? (Yates 746). The most commonly used formulas are provided in the next page. Probability Density Function Graph 1. 4 Examples of the degrees of freedom??™s curve on the chi-square graph."Hypothesis 2. 1 For the chi-square, the hypothesis does not need any numbers and only needs a statement of the question being observed. Example 2 Scientists in the Coca Cola Company believes that more people liked the Safeway??™s Cola more than theirs because there was not enough carbon dioxide in their soda. Identify the null and alternative hypothesis for this problem.

Ho: There is no relation between the amount of carbon dioxide and if a person likes the soda more. Ha: There is a relation between the amount of carbon dioxide and if a person likes the soda more. Conditions 2. 1 a) Random sample of size n is taken and is a simple random sample. b) Sample size n is large enough to get an approximate chi-square distribution. a.

All expected counts are at least 5 Formulas 2. 1 a) Finding the expected count Row total x Column total 1. Expected Count = Table total b) Chi-Square Statistic 2 1. $\chi^2 = \sum (O - E)^2 / E$ c) Degrees of Freedom 1.

n ??“ 1 degrees of freedom (Mulekar 255) (Yates 743) Nature of the Chi-Square Test In the chi-square test, the table totals of all the columns and rows are found by adding them together and finding the expected count (Refer to table 1. 2 on this page). Table 1. 2 Observed Data Set 1 Data Set 2

Row Total	Coca Cola (A)	Safeway??™s Cola (B)	Column Total
68	68	82	(68 + 82)
82	73	84	(73 + 84) = 157
150	141	166	(141 + 166) = 307

307 (Total subjects) The columns and rows of each are added together so the observed

counts of people who liked each the soda in the sample populations can be counted. Table 1. 2 represents a 2x2 table because it has 2 columns and 2 rows total. The difference between observed and Table 1.

21 Expected Coca Cola (A) Data Set 1 $(141 \times 150) / 307 = 68.9$ Safeway??™s
 Cola (B) Data Set 2 $(141 \times 157) / 307 = 72.1$? 69? 72 People
 $(166 \times 150) / 307 (166 \times 157) / 307 = 81.1$? 81 people = 79.3 ? 79 People
 expected is that observed is the actual sample proportions where as
 expected is our predicted proportions/percentages.

Sample Question 2. 2 Carry out an appropriate chi-square test to determine if the distribution of whether people liked the soda brand for Safeway??™s Cola is significantly different from the distribution of the likeliness towards the Coca Cola brand."Note: The way to find the expected values are shown on formula 2.

1A. All numbers found here were used from table 1. 2??™s observed values. We can also confirm that it is able to be tested on because all expected values are greater than five.

Hypotheses Ho: There is no significant difference in distribution of likeliness towards the Coca Cola brand. Ha: There is a significant difference. TI-83 Plus After finding the table??™s observed and expected values (tables 1.

2 and 1. 21) and ??? Stat ? Edit??! o we can now find the degrees of into your determining the hypotheses, Plug in all your observed values freedom for the test (df). list one (68, 82, 73, 84) There is a need for degrees of freedom because it is a proportional test??" a key difference o Plug in all the expected

values into the test between the z-proportional test and the chi-square test.

(69, 81, 73, 79) ??? Highlight the L3 and insert this of freedom Degrees of freedom = 2-1 = degrees in it o + Enter (Refer to formula set 2. 1C for the formula) L2 ??? You would receive your χ^2 values in which are After finding the degrees of p-values. But are able to look find the chi-square values by your freedom, we wait, we are not done yet! ??? Stat ? Calc ? 1-Var Stats putting it into the calculator. Stats L3 + Enter ??? 1-Var o We look at the Ex2 because we tested two samples o The value should be 0. 100 ??? 2nd + Dist ? χ^2 cdf(o χ^2 cdf(0. 100, 100, 1) ? We put a one because it is our degrees of freedom! ??? P-Value = 0. 752(L1 ? L2) 2"Since our value exceeds 0.

10, we can conclude that there is no significant difference in likeliness between the two soda brands. VII. Common Similarities Between the Z-Proportional and Chi-Square Test As said earlier, there are some common things between a z-proportional test and a chi-square test.

For example, the conditions for both tests require a population value to meet and consist of a sample size n. In addition, the tests determine their answers through the finding of their proportions, for the z-proportional it is to find the \hat{p} and for the chi-square, it is to find the tables??™ expected counts.

However, there are still some major differences between the two tests. One thing that is contrasting is their way of solving for whether something is significant enough. Similarities between the two tests:"??? ??? ??? ??? They both require proportional testing They both use a sample size of n that must be greater than a set value They both test for the same conclusion (whether

something is significant) The hypotheses for both deals with the relations of the experiment (if something causes something else to be significantly different)??? Both are commonly tested at a 5% or 10% interval because it is the most common and represents the majority Of course there are differences between the two: ??? Z-proportional is still a part of the normal distribution testing category whereas the chi-square test is of its own branch thus causing the tests to be done in very different ways VIII. Conclusion In summary, the ??? nature of a simple significance test is set out and its implications explored??? (Cox 30).

The creation of tests of significances were created so that these implications of societal uses could be explored such as the asking if one prefers something over another thing. By doing so, we create frequentism, the study probability through the creation of tests. In this paper, there were two tests used to show the study of frequentism??” the z-proportional and chi-square test. However, these two tests are done in very different ways as outlined in section seven and the"previous sections of which the research question was integrated to fit the tests. These two tests are very good examples of frequentism over time because the two tests are used to provide a similar conclusion and hypothesis on whether something are significant or not. Over time, frequenters developed the z-proportional and chi-square to prove whether something provides a direct relation to the question being explored. In this case, does someone like soda because of its brand name After testing with the z-proportional, we determined that it was possible that people do not like a soda for its brand name because our p-value was below the 10% ($p= 0.04??!$).

However, in the chi-square test, when we tested for the same thing, we came up with a p-value that was exceedingly high, thus causing us to say that there was no difference in proportional significance in likeliness between the two sodas ($p = 0.752$). It is interesting how one test proves the null hypothesis to be false and another for it to be true. In conclusion, over time, statistical math changes by adding new methods of mathematical procedures in which may seem redundant due to its varying conclusions yet meant to provide the same conclusion. Works Cited Page 1.

Cox, D. R. Principles of Statistical Inference.

New York: Cambridge University Press, 2006. 2. Megeath, Joe D.

. How to use Statistics. New York, New York: Canfield Press & Harper and Row Publishers, Inc, 1975. 3. Mulekar, Madhuri S.. Cracking the AP Statistics Exam.

04-05. New York: Random House, Inc. 2004. 4. Phillips, John L.. How to Think About Statistics.

6. New York: W. H.

Freeman and Company, 2000."5. Creative Research Systems, " Significance in Statistics & Surveys – What is significance and the meaning of significance.

" Significance in Statistics & Surveys 2007 22 Oct 2008 . 6. Yates, Daniel S., David S. Moore, and Daren S. Starnes.

The Practice of Statistics: TI-83-89Graphing Calculator Enhanced. 2. New York: W. H Freeman and Company, 2002.