Ismael Leyva
April 17, 2019
Mathematical Studies SL

Tardiness, bedtime,and your gender

Introduction:

This project will determine if there is a correlation between your bedtime, tardiness, and gender. It all stems from an essential question of is there a correlation between bedtime,tardiness, and gender? To test my question I will need data so i will do some convenience sampling and voluntary response sampling to get what is necessary. Convenience sampling is when a person gets data from people who will give a response. Or another way of putting it is sending a survey to a group of people where you know you'll get responses. In my case everyone in my grade are sending out their own surveys to their peers for their own projects. So I will create a survey that consists of three questions. The questions are "Give an estimate of how many times you been late to school this year?", "what is your gender?", and "How late do you normally go to sleep on a school night?". So I sent my survey just like everyone else. However this is also voluntary response sampling because this is when one sends a survey to a group of people (like an email) and if they do not send any responses back then nothing.This means people who answered my survey were voluntarily doing so. This is the most efficient way to collect data to answer my question because it will be honest answers from students. Also, them answering my survey and me answering their surveys is like a fair trade so the answers should be usable. To form an analysis from the data I will use the Pearson's Correlation Coefficient formula and the Chi-Square Statistic Formula to test my essential question. The Pearson's Correlation Coefficient formula is extremely useful in statistics because it is used to determine the strength of two variables Where the chi-square test finds the relationship through categorical data. This will then see if my variables are in fact correlated or not. Then I will be able to determine a valid answer to my question.

Statement of Task:

The task of this project is to find the relationship if there is any between three variables with data that was personally collected for accurate testing. To determine the relationship using the Pearson Correlation Coefficient formula is to find the coefficient value. The coefficient value is always between -1.00 and 1.00 where if the value is negative then that shows the variables are negatively correlated to each other and vise versa when the value is positive showing how they are positively correlated. Where to determine the relationship using the chi-square statistic formula it finds how likely the observed data is from chance or the correlation of the two variables. There is no need to change the chi-square value that is found to the p-value. This is the value that will tell me if the variables are correlated or not. I can use a table that tells me the chi-square values found in tables with degrees of freedom.

Data:

After one week of when I sent out my survey I got a total of 36 responses. I used google forms to create my table that I sent out. So to be able to analyze my data and use the pearson correlation coefficient formula and the chi-square formula. So the table that I labeled “table 1” is all of my data that I will be using. In other words it’s my original data table in this project. Everything is coming from this table allowing me to see all of my data to organize however will be the most useful to me.

Table 1					
Time	**Lateness**	**Gender**	**Time**	**Lateness**	**Gender**
11.5	10	male	11.5	5	Female
10.5	0	female	12.5	60	Male
1	3	female	1	5	Male
11.5	7	female	10	0	Male
11.5	0	Female	11.5	15	Female
10.5	1	Male	10.5	0	Female
10.5	10	Female	10.5	0	Male
11.5	0	Male	11.5	0	Female
10.5	1	Male	12.5	30	Female
11.5	2	Male	10.5	2	Female
10.5	0	Female	10.5	0	Female
10.5	1	Female	11.5	2	Female
11.5	0	Male	1	2	Female
12.5	3	Female	10.5	1.5	Female
10.5	5	Female	10.5	1	Male
11.5	10	Female	10.5	4	Male
12.5	2	Female	11.5	3	Female
12.5	50	Female	12.5	3	Female

Now I need to create a data table that will help me find the Chi-square value. As explained before chi-square value only takes in categorical data. That means the data may be divided into groups or when numbers are collected into groups or categories. So I need to create another table from table 1 that follows those parameters. Then i will be able to use the chi-square test. Otherwise If i use table I then my calculations would be unsuccessful because this is showing individual answers not grouped or categorical. Table 2a and 2b will fix that by showing the average number for male student’s or female student’s bedtime, and lateness to school. Finding the average in a column of a data table is when all of the data elements undergo a calculation of addition finding the total sum of the data elements in the column of the data table. Once the sum is found it is divided by the total number of elements in that row. I understand that completing this process and finding the average will give me meaningful data that when I do my calculations with it will be correct. This data is correct because the average is categorical data following the parameters of the chi-square value formula of having categorical data.

Chi-square Data tables:

Table 2a		
Male Students	Time	Lateness
1	11.5	10
2	10.5	1
3	11.5	0
4	10.5	1
5	11.5	2
6	11.5	0
7	12.5	60
8	1	5
9	10	0
10	10.5	0
11	10.5	1
12	10.5	4
SUM	122	84
average	10.16666667	7

As explained this data table shows all of my male student responses and when they go to bed and their tardiness. This identifies all of my variables that I want to address but my female student responses. So that is the reason for table 2b showing the missing female student responses value. Also both tables gives me the average that I explained is categorical data.

Table 2b								
Female students	Time	Lateness	Female students	Time	Lateness	Female students	Time	Lateness
1	10.5	0	14	11.5	15	12	12.5	50
2	1	3	15	10.5	0	13	11.5	5
3	11.5	7	16	11.5	0			
4	11.5	0	17	12.5	30			
5	10.5	10	18	10.5	2			
6	10.5	0	19	10.5	0			

7	10.5	1	20	11.5	2
8	12.5	3	21	1	2
9	10.5	5	22	10.5	1.5
10	11.5	10	23	11.5	3
11	12.5	2	24	12.5	3

Now I need to organize data so I will be able to use the Pearson's Correlation coefficient formula right. The parameters for this formula is that you need an x and y variable. The data table 2a and 2b are within these parameters so I do not need to create anything knew. But I could only use the average values but i fear it would give me insufficient data because I believe more the better. I will also explain later in the project why.

Data analysis:

I have organized the data now I need to use the equations so I can get my answer for my question. These equations I got from two websites that are cited below in MLA format and the textbook *The Mathematics for the international student Mathematical studies SL third edition by Mal coad, Glen Whiffen, Sandra Haese, Michael Haese, Mark Humphries.* This textbook is used for the IB diploma Programme and the websites just have a dulled down equation from the textbook.They are essentially the same equation but the ones from the websites have numbered steps.

To find the coefficient value from the pearson's correlation coefficient formula I will use the data from data table 2a to create data table 3 that holds only male student responses. And I will use the data from data table 2b to create data table 4 that holds only female student responses. The actual pearson's correlation coefficient formula that I am using looks like this and I got this equation from the websites and textbook that is sited below.

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

The way I solved this equation is break up the specific variables by finding them one by one in the data table 3. Once I find every variable I can plug them into the equation to find the coefficient value. In data table 3 the first step to find the coefficient value for the male student responses will be to label the bedtime values x and the Lateness to school values y. Correlation Coefficient Data tables: Then I solved for xy first my simply finding the product of x and y values. Then I found the sum of the xy values giving me the first variable in the equation. Then I solved for the x^2 value and finding the sum of that. Then y^2 values and I found the sum of that. I simply plugged in my equation and solved it.

Table 3					
For Male Responses	Time (x)	Lateness (y)	xy	x squared	y squared
1	11.5	10	115	132.25	100
2	10.5	1	10.5	110.25	1
3	11.5	0	0	132.25	0
4	10.5	1	10.5	110.25	1
5	11.5	2	23	132.25	4
6	11.5	0	0	132.25	0
7	12.5	60	750	156.25	3600
8	1	5	5	1	25
9	10	0	0	100	0
10	10.5	0	0	110.25	0
11	10.5	1	10.5	110.25	1
12	10.5	4	42	110.25	16
SUM = (Σ)	122	84	966.5	1337.5	3748

$$r = \frac{12(966.5) - (122)(84)}{\sqrt{\left[12 \times 1337.5^{2} - (122)^{2}\right]\left[12 \times 3748^{2} - (84)^{2}\right]}}$$

First I found the 1337.5^2 value did all of the multiplication.

$$r = \frac{11598 - 10248}{\sqrt{[16050 - 14884][44976 - 7056]}}$$

Then I did all of the subtraction in the equation.

$$r = \frac{1350}{\sqrt{(1166)(37920)}}$$

Now I found the square root in the denominator (that is the bottom of the fraction bar).

$r = \frac{1350}{6649.415012} = 0.2030253786$ This is the coefficient value for males I did this same process in table 4 but with the female student responses.

Table 4

Female responses	Time (x)	Lateness (y)	xy	x squared	y squared
1	10.5	0	0	110.25	0
2	1	3	3	1	9
3	11.5	7	80.5	132.25	49
4	11.5	0	0	132.25	0
5	10.5	10	105	110.25	100
6	10.5	0	0	110.25	0
7	10.5	1	10.5	110.25	1
8	12.5	3	37.5	156.25	9
9	10.5	5	52.5	110.25	25
10	11.5	10	115	132.25	100
11	12.5	2	25	156.25	4
12	12.5	50	625	156.25	2500
13	11.5	5	57.5	132.25	25
14	11.5	15	172.5	132.25	225
15	10.5	0	0	110.25	0
16	11.5	0	0	132.25	0
17	12.5	30	375	156.25	900
18	10.5	2	21	110.25	4
19	10.5	0	0	110.25	0
20	11.5	2	23	132.25	4
21	1	2	2	1	4
22	10.5	1.5	15.75	110.25	2.25
23	11.5	3	34.5	132.25	9
24	12.5	3	37.5	156.25	9
Sum	251	154.5	1792.75	2833.5	3979.25

$$r = \frac{24(1792.75) - (251)(154.5)}{\sqrt{\left[24\times2833.5^{2} - (251)^{2}\right]\left[24\times3979.25^{2} - (154.5)^{2}\right]}}$$

$r = 0.2243174691$ This is the coefficient value for female students. I just followed the steps I layed out when I solved for the coefficient value for the male students.

The meaning of the Pearson’s correlation coefficient formula is to find the coefficient value. Since the two coefficient values are between -1 and 1 I did a valid equation using valid data. Now from the equations I have two valid coefficient values and that will tell me how strong the correlation is between the three variables I am testing. Therefore It will answer my essential question for this project. And My answer is that there is a weak correlation with these variables for both the male and female responses. The values were almost identical. Now to get further proof I will now find the chi-square value.

Now I can solve for the chi-square value because I have already found the coefficient values. to find the chi-square value as explained earlier is the probability if these variables have any correlation with each other or are related by chance. That means I need to find an expected value from my observed value The observed value are the data that comes from data table 1 and is organized to fit a chi-square data table to be able to solve for the chi-square value. The expected value is found in the process. The reason why I am using the average values that were calculated from data table 2a for the male responses and from data table 2b the female responses as my observed data. The average is categorical data because it groups all of the response to be one response. This is also my observed data because this is what comes from the original data tab
le 1. So the equation for the chi-square value that I got from the textbook and websites that I cited at the end.

$$x^2 = \sum \frac{(O - E)^2}{E}$$

Step 1) I need to set up a new data table with organized data to I am able to solve for the chi-square value. This new data table is called data table 5 and it has the average values that comes from data table 2a and data table 2b. But in order to find the expected values I need to find the total. To find the total on the right side of data table 5 you first find the sum of the average male values to get a total of 17.16666667 and then find the sum of the average female values with a total of 16.89583333. Then the find the sum of the sums just found to get a total of 34.0625. To find the total value for time column you find the sum of the average male value for time and the average female value for time to get a total of 20.625. Then do the same but to find the total for the lateness value. The total for the lateness value is 13.4375.

Data Table 5	Time	lateness	Total
Male	10.16666667	7	17.16666667
Female	10.45833333	6.4375	16.89583333
Total	20.625	13.4375	34.0625

Step 2) Now the next step is to multiply the total time value and the average male response for time to get $209.6875 = (10.16666667 \times 20.625)$. That is the data that goes in row 2 and column 2 then in row 3 and column 2 you get $215.703125 = (10.45833333 \times 20.625)$. Then you do the same in column 3 but the total value you are multiplying with is 13.4375 the total for lateness not time. So the data that goes in row 2 column 3 is $94.0625 = (7 \times 13.4375)$. then the data in row 3 column 3 is $86.50390625 = (6.4375 \times 13.4375)$. This data will be put into data table 6 below.

Data Table 6	Time	lateness
Male	209.6875	94.0625
Female	215.703125	86.50390625

Step 3) Now the final step to find the expected values I need to divide the values in data table 6 by the total number found in data Table 5. So row 2 column 2 will look like $6.155963303 = (209.6876 \div 34.0625)$ and in row 3 column 2 is $6.332568807 = (215.703125 \div 34.0625)$. where in row 2 column 3 is $2.76146789 = (94.0625 \div 34.0625)$ and row 3 column 3 is $2.53956422 = (86.50390625 \div 34.0625)$ This data will make up Data table 7 right below.

Data table 7	Time	lateness
Male	6.155963303	2.76146789
Female	6.332568807	2.53956422

Step 4) Now I need to solve for the numerator portion of the chi-square formula by finding the difference between the observed data and the expected data. So in row 2 column 2 the equation is simply $4.010703367 = 10.16666667 - 6.155963303$ and row 3 column 2 is $4.125764526 = 10.45833333 - 4.125764526$. Finally the equation for the data in row 2 column 3 is $4.23853211 = 7 - 2.76146789$. And in row 3 column 3 is $3.89793578 = 6.4375 - 2.53956422$. This data is used to create data table 7 below.

Data Table 7	Time	lateness
Male	4.010703367	4.23853211
Female	4.125764526	3.89793578

Step 5) Now to find the value to put in the numerator for the chi-square value is to square the data in data table 7 to create data table 8.

Data Table 8	Time	lateness
Male	16.08574147	17.96515445
Female	17.02193292	15.19390334

5)Now I know what goes into the numerator and denominator in the chi-square formula. So the equation used to find the data in row 2 column 2 is $2.61303401 = \frac{16.08574147}{6.155963303}$ where in row 3 column 2 is $2.6879981 = \frac{17.02193292}{6.332568807}$. Now the equations for the data in row 2 column 3 is $6.505653936 = \frac{17.96515445}{2.76146789}$ and in row 3 column 3 is $5.982878174 = \frac{15.19390334}{2.53956422}$. This data will be used to create data table 9 that is right below.

Data Table 9	Time	lateness
Male	2.61303401	6.505653936
Female	2.6879981	5.982878174

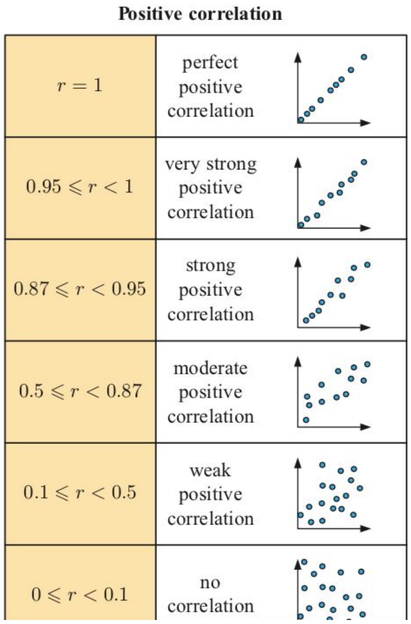
6)Now all that is left is to find the sum of the values in data table 9. So the equation will be a simple addition equation that looks like this

$$17.78956422 = 2.61303401 + 2.6879981 + 6.505653936 + 5.982878174$$

So my chi-square value is 17.78956422 and that came from data table 9 that came from data table 1.

Conclusion:

I used two processes in my project. I used the pearson’s correlation coefficient formula and the chi-square formula. I then broke each formula into steps making it a process. I got both the equations from two websites and an IB textbook. However, this project is trying to determine if there is a relationship for students going to bed too late and being late for school. I only sent my survey out to students who are in the Dwight school. So I learned that there is a weak positive correlation according to this chart that I got out of the IB textbook.

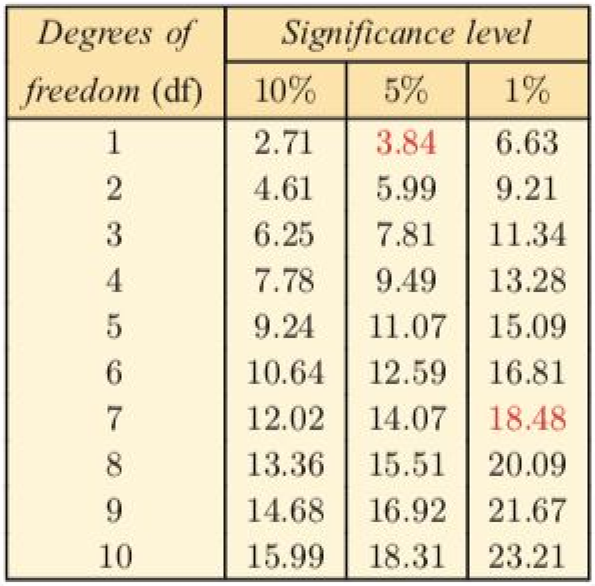


Positive correlation

$r = 1$	perfect positive correlation	
$0.95 \leqslant r < 1$	very strong positive correlation	
$0.87 \leqslant r < 0.95$	strong positive correlation	
$0.5 \leqslant r < 0.87$	moderate positive correlation	
$0.1 \leqslant r < 0.5$	weak positive correlation	
$0 \leqslant r < 0.1$	no correlation	

So what this chart is telling me that the variables that I am testing are not very related according to the math. This means when somebody says that going to sleep late effects when you get to school they are correct but not very. More or less one can argue that it really doesn’t and their is another variable that affects your tardiness whether you are a male or female. So any question of bedtime and getting to school on time if you are a male or female are related but not by much meaning their is another variable in effect. This is proved by my math and how my calculations proves this conclusion. I have further checked my basic math in case of little mistakes with the help of a TI-84 calculator. This is a calculator that has a function to calculate the coefficient value using the pearson’s correlation coefficient formula. When I plugged in my data using the appropriate table (I used data table 5) to calculate the coefficient value I got the same answer. So my conclusion still stands. So when your parents or teacher says that going to sleep early will help you get to school on time they are only a little right.

So for a conclusion of the chi-square formula process. My chi-square value was 17.78956422. And as explained the chi-square formula tests if the variables are independent or not. When using this process I created two hypotheses: if the chi-square value is greater than the critical value than the variables are not independent, and if the chi-square value is less than the critical value than the variables are independent. According to the chart below that I got from the IB textbook the chi-square value is greater than then all of the critical values. So this means that the variables that I am testing are not independent. This means that going to sleep late and being tardy have a relationship with one another. So when a teacher or parent tells a student that going to sleep late will make the student late for school. They are correct but as seen with the coefficient value the variables are not strongly correlated. So this means a valid answer to try and go to sleep later is telling the one asking the question that going to sleep late will not automatically make me late for school. I just simply do not want to go to school. That answer identifies another variable in effect and that variable is wanting to go to school. To this variable correlates with tardiness more then going to sleep early.



Degrees of freedom (df)	*Significance level*		
	10%	5%	1%
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21

The numbers in this chart are the critical values from the IB textbook

Validity:

One limitation that may affect my overall conclusion is that there were responses that were unidentifiable. In one of the responses from my survey that I sent out to my peers was blank. The question the student did not answer was how many latenesses they had. To complete my table I inferred that the student ment zero. But since the survey was anonymous there was no way to track down the student for clarification. However something this small should not take effect in my overall conclusion.

Works Cited

Chi-Square Test, www.mathsisfun.com/data/chi-square-test.html.

Study.com, Study.com,

study.com/academy/lesson/pearson-correlation-coefficient-formula-example-significance.html.

Coad, Mal. *Mathematics for the International Student: Mathematical Studies*. Haese And Harris Pub, 2010.

“Correlation Coefficient: Simple Definition, Formula, Easy Calculation Steps.” *Statistics How To*, www.statisticshowto.datasciencecentral.com/probability-and-statistics/correlation-coefficient-formula/.