SEMESTER PROJECT

MBA 506 | Team 13

Reema Bhattacharya | Sten Swenson | Alanna Hurley | Colin Donahue

JUNE 16, 2020

NCSU JENKINS MBA PROGRAM
SUMMER 2020

Executive Summary

Team 13 of MBA 506 was tasked with analyzing a multitude of data gathered from a survey given to students in Math and Portuguese courses. In this report, there were several questions that needed to be answered using confidence intervals and hypothesis testing.

We analyzed the following confidence intervals: 90% confidence interval for the age of the students; 99% confidence interval for the final grade; and a 95% confidence interval for number of absences and workday alcohol consumption. We found the confidence intervals to be (16.3914, 16.8086), (10.1609, 12.1314), (2.7569, 6.2957), and (1.1848, 1.5309), respectively.

We also performed hypothesis testing on the following questions: do students whose parents live together get better grades (alpha = 0.05)?; do students with internet access get better grades (alpha=0.01)?; are the math grades between the two schools different (alpha=0.1)?; do students who consume higher levels of alcohol on work days get worse grades?; and do students who attended nursery school get better grades?

To answer all of these hypotheses, we analyzed the data using a t-test using a sample of 95 data points out of the entire population. For the first hypothesis, we found that there is not sufficient evidence to suggest that students whose parents live together get better grades. For the second hypothesis there is a discrepancy between the sample and population t-tests-- the sample t-test indicated we fail to reject the null hypothesis and the population t-test indicates we reject the null hypothesis. Based on the population having more data points, we indicate that there is a significant difference and we conclude that those with internet access do indeed get better grades. This is a Type I error. For the third hypothesis, there is not sufficient evidence to suggest that the math grades between the two schools are different.

For the fourth hypothesis, there is a discrepancy between the sample and population t-tests-- the sample t-test shows that we fail to reject the null hypothesis but the population t-test indicates that we should reject the null hypothesis. Due to the population being larger than the sample, we conclude that we should reject the null hypothesis, confirming that those who drink more on workdays get worse grades. This is a Type II error. For the fifth hypothesis there is also a discrepancy-- the sample t-test indicates we should reject the null hypothesis, but the population t-test indicated we should fail to reject the null hypothesis. Again, due to the higher amount of data points, we will conclude that we fail to reject the null hypothesis and that there is no indication that students that attend nursery school get better grades. This is a Type I error.

Team 13 hopes that management finds this information, as well as the attached graphs, useful in their forthcoming assessment.

DATA ANALYTICS

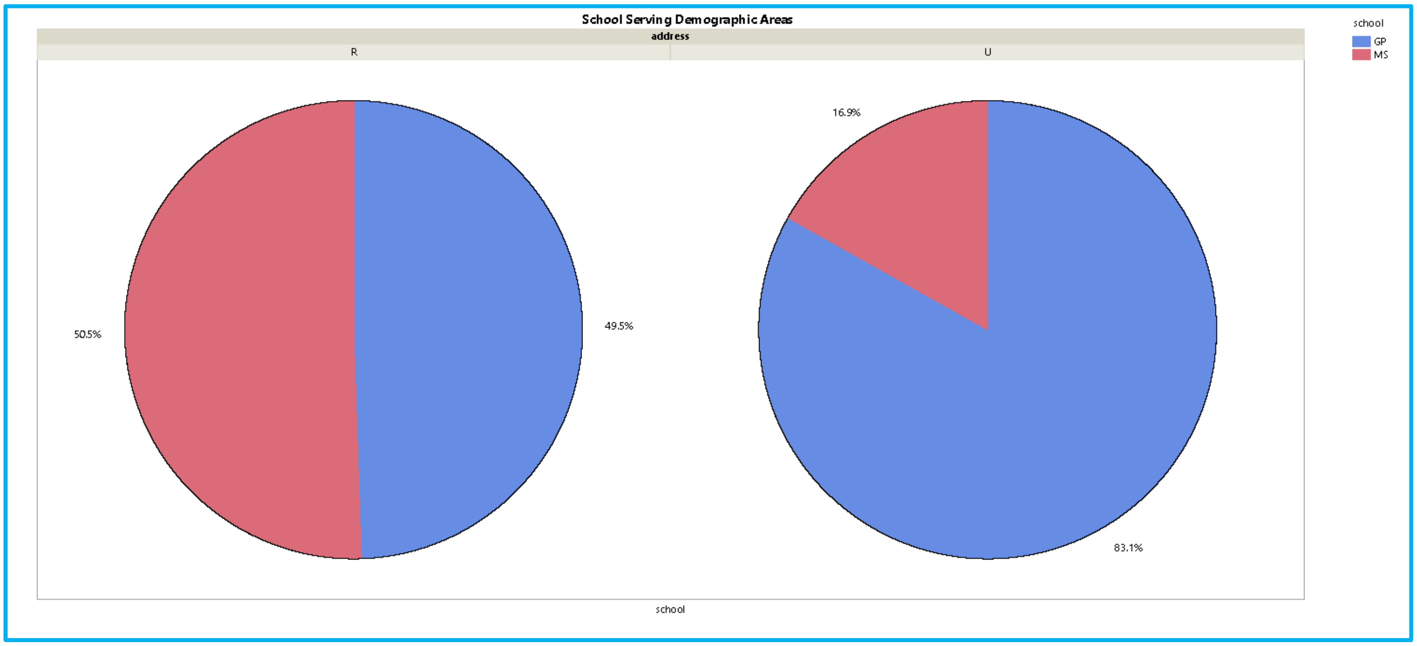
IN THIS SECTION WE HAVE CREATED TEN GRAPHS OF AT LEAST FOUR DIFFERENT TYPES THAT ILLUSTRATE USEFUL INFORMATION ABOUT THE DATA.

Graph details

#	Graph Title	Graph Type
1	*Schools serving urban and rural areas*	*Pie Chart*
2	*Student failures to study time grouped by gender*	*Bar graph*
3	*Effect of internet availability on student grades*	*Histogram*
4	*Effect of Nursery School Education on Student' Final Grades*	*Scatter plot*
5	*Effect of Weekday Alcohol Consumption on Mean Final Grade*	*Line Graph*
6	*Distribution of Grades for Math and Portuguese grouped by School*	*Box plot*
7	*Effect of Absences over Final Grades(G3)*	*Scatter plot*
8	*Distribution of Grades by Student Age*	*Bar Graph*
9	*Frequency Distribution of Grades and Relationship VS No Relationship*	*Histogram*
10	*Percentage of Internet Users*	*Bar graph*

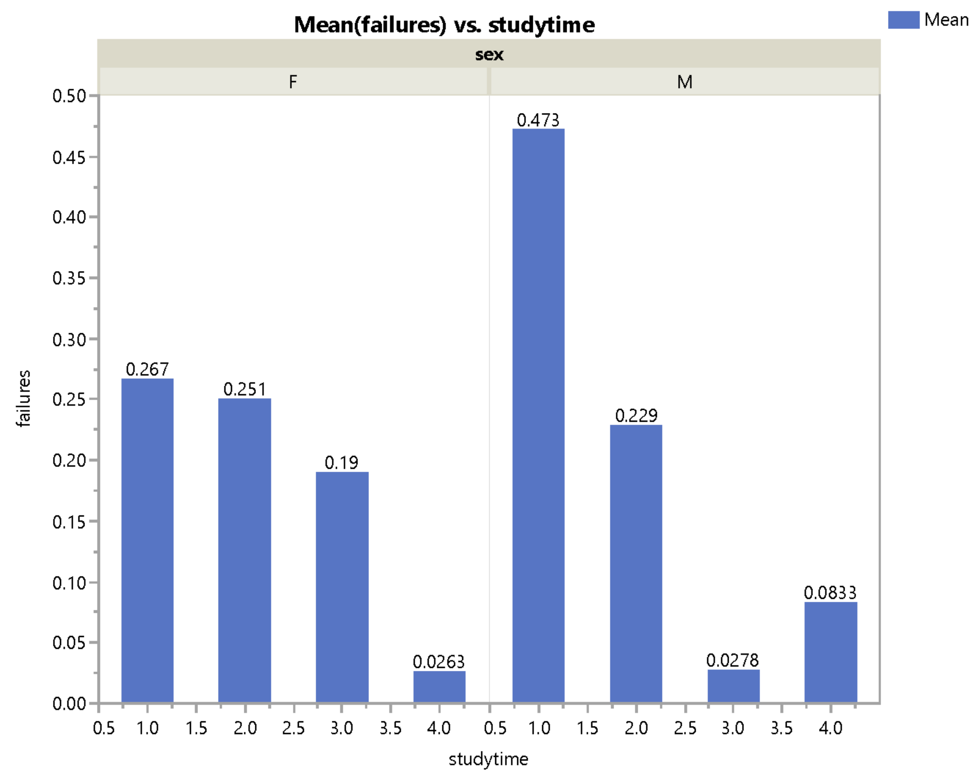
SCHOOL SERVING DEMOGRAHIC AREAS

RURAL VS. URBAN



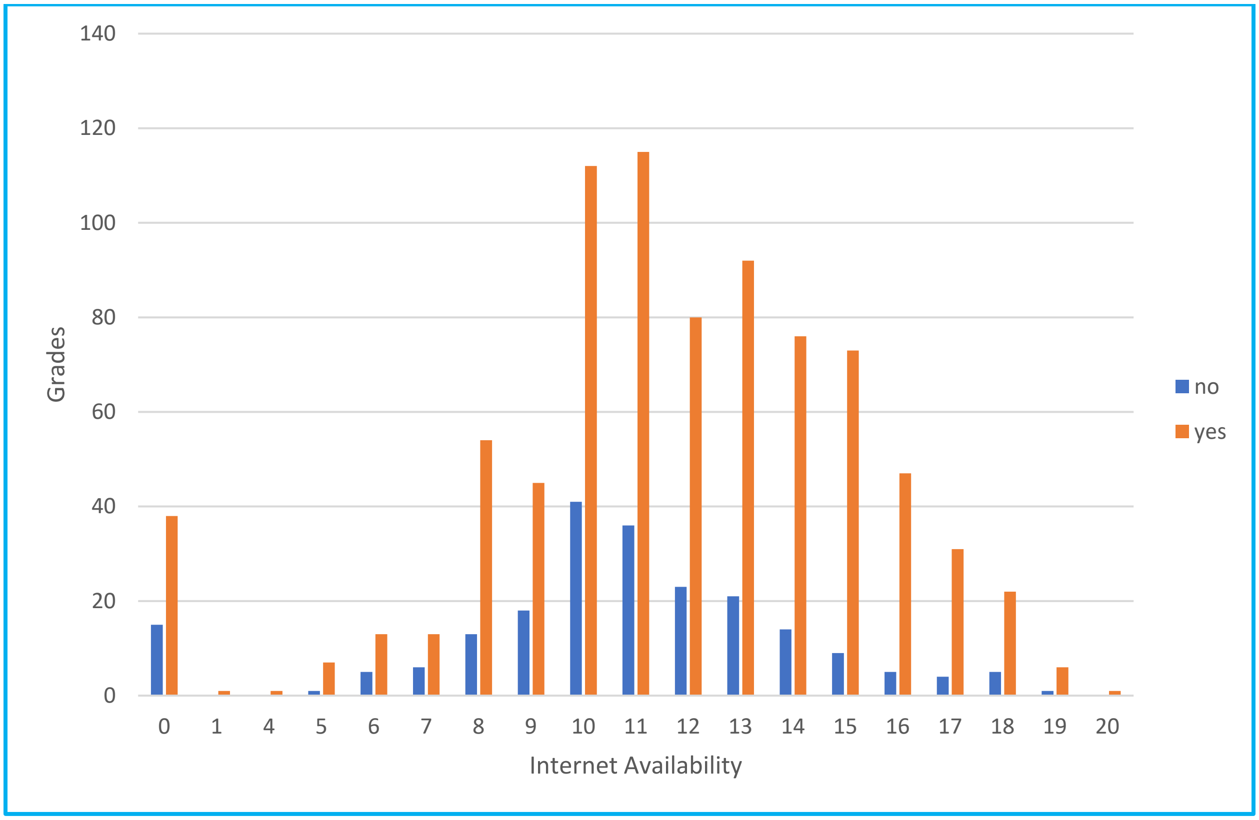
In this graph we show how the two schools, 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira are serving the rural and the urband areas. Gabriel Pereira almost has a 50 – 50 distribution of students for rural and urban areas where as Mousinho da Silveira largely serves the urban area.

STUDENT FAILURES TO STUDYTIME GROUPED BY GENDER



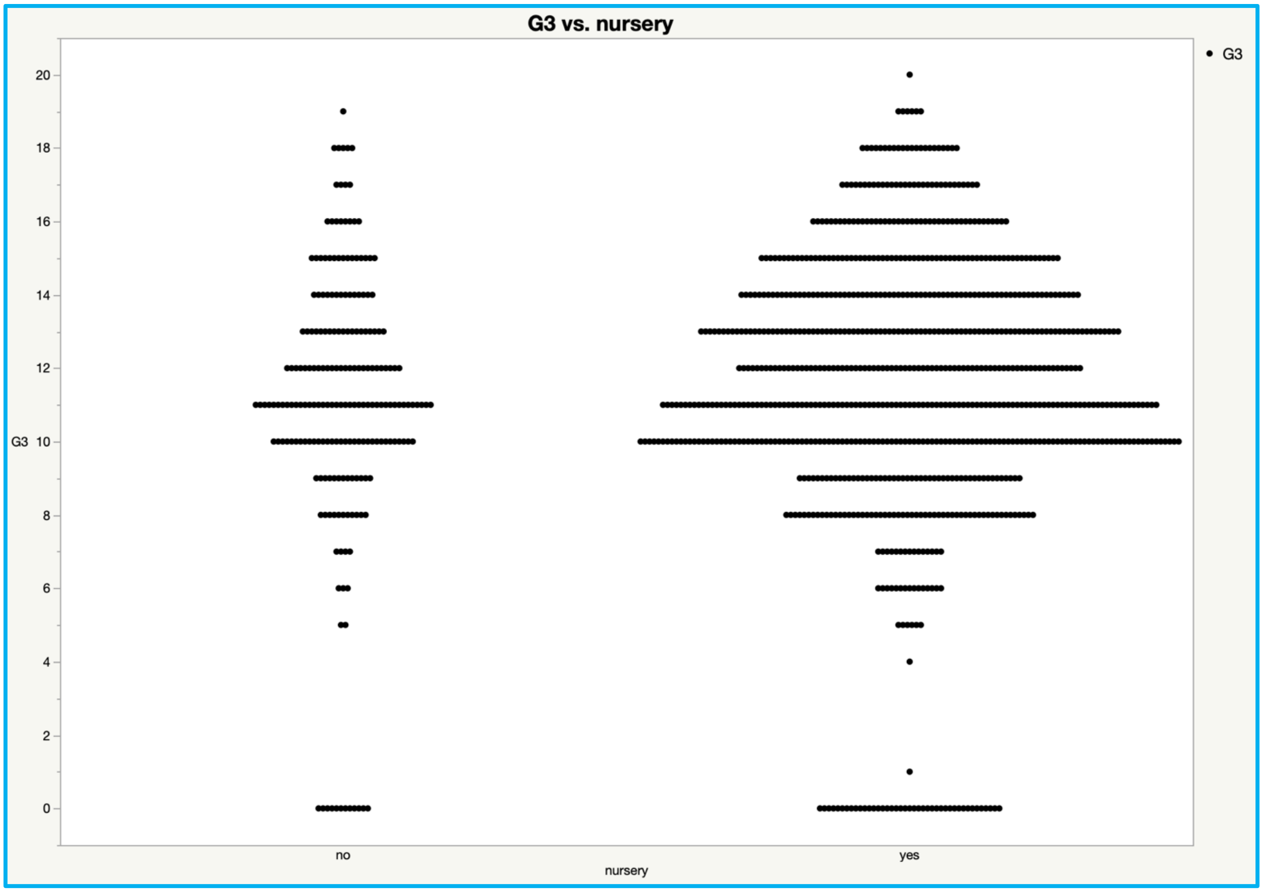
This graph shows the relationship with study times and failures and how it effects the male and female demographics. Even though it shows more you study, less your chances of failing for both the groups but the rate of failure in males are higher than females.

Effect of internet availability on student grades



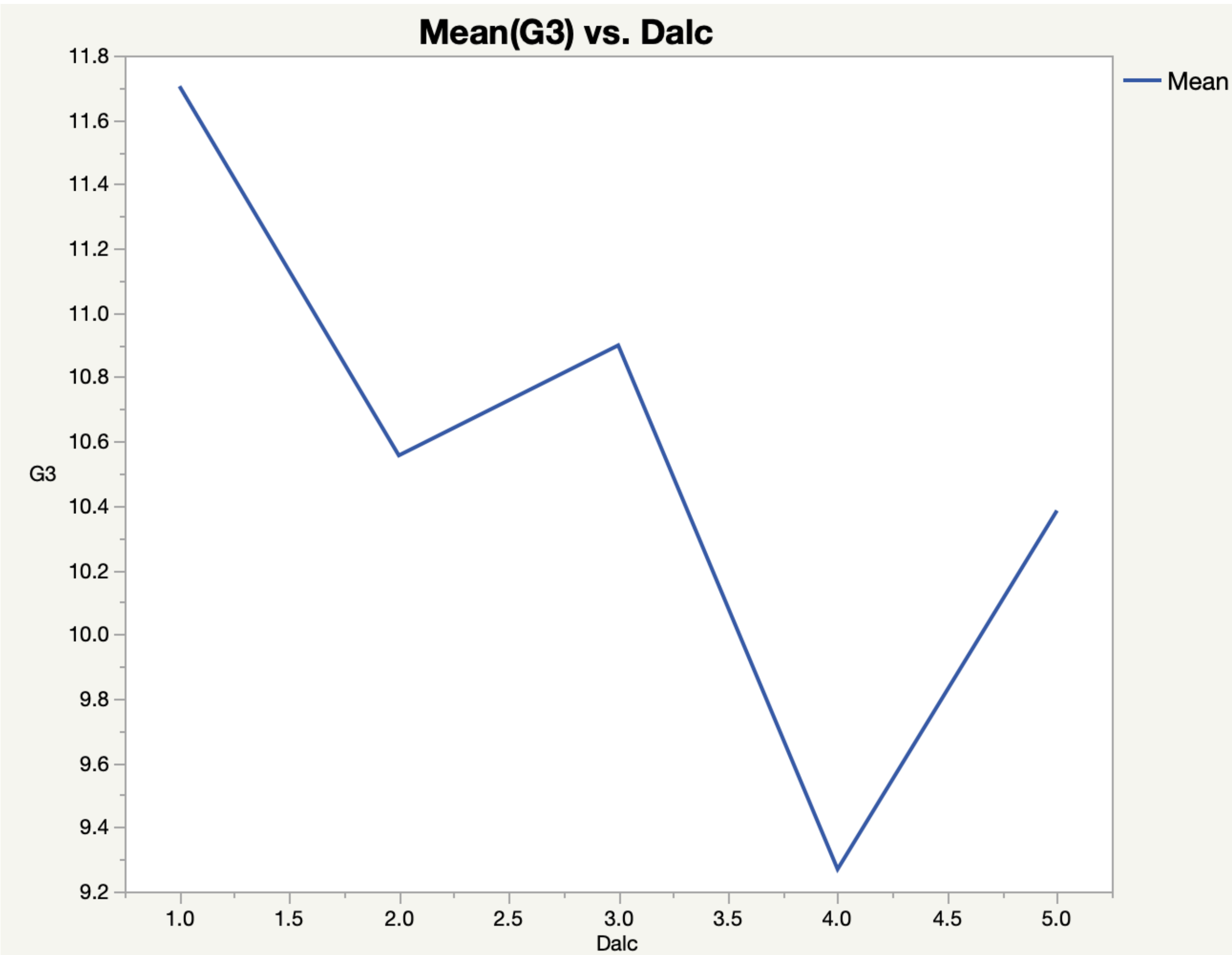
The Histogram compares effects of internet availability on student grades. They are skewed left with outliers.

Effect of Nursery School Education on Student' Final Grades



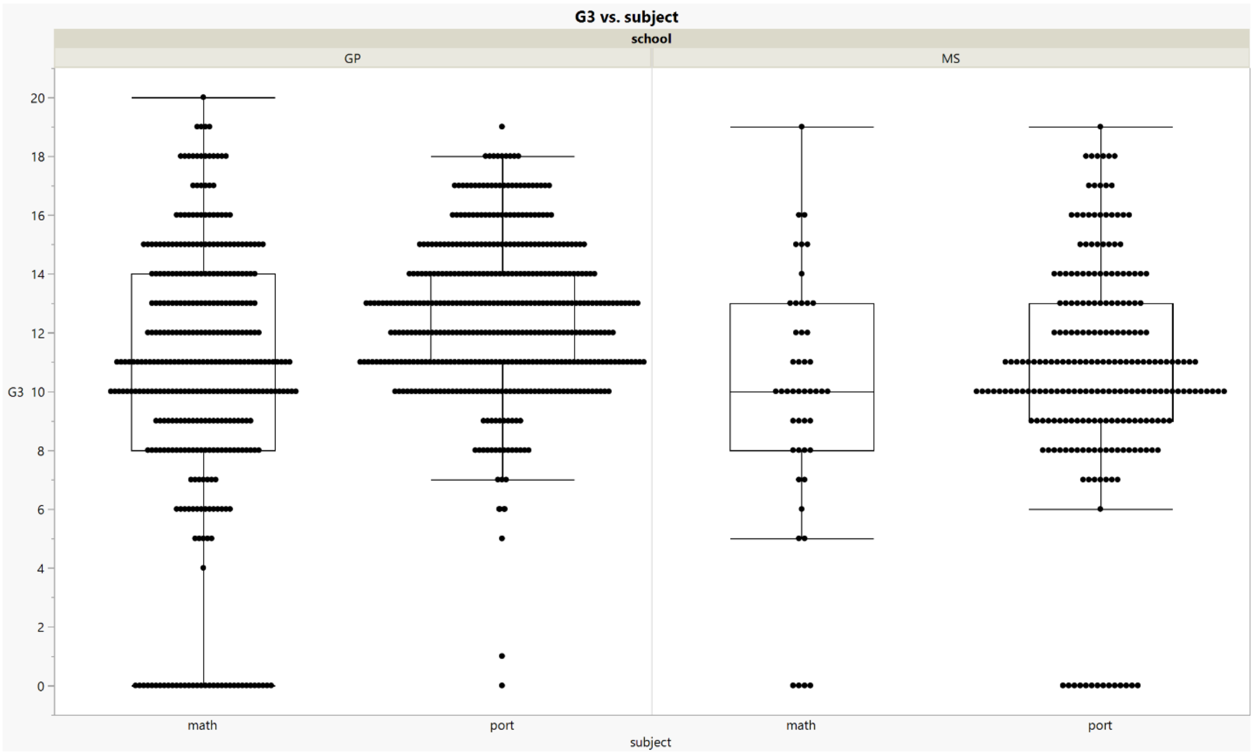
This Scatter Plot above represents students' final grades, separating them into those that did and did not attend nursery school. At a glance you can see that nursery school appears to have no impact on the final grade distribution.

Effect of Weekday Alcohol Consumption on Mean Final Grade



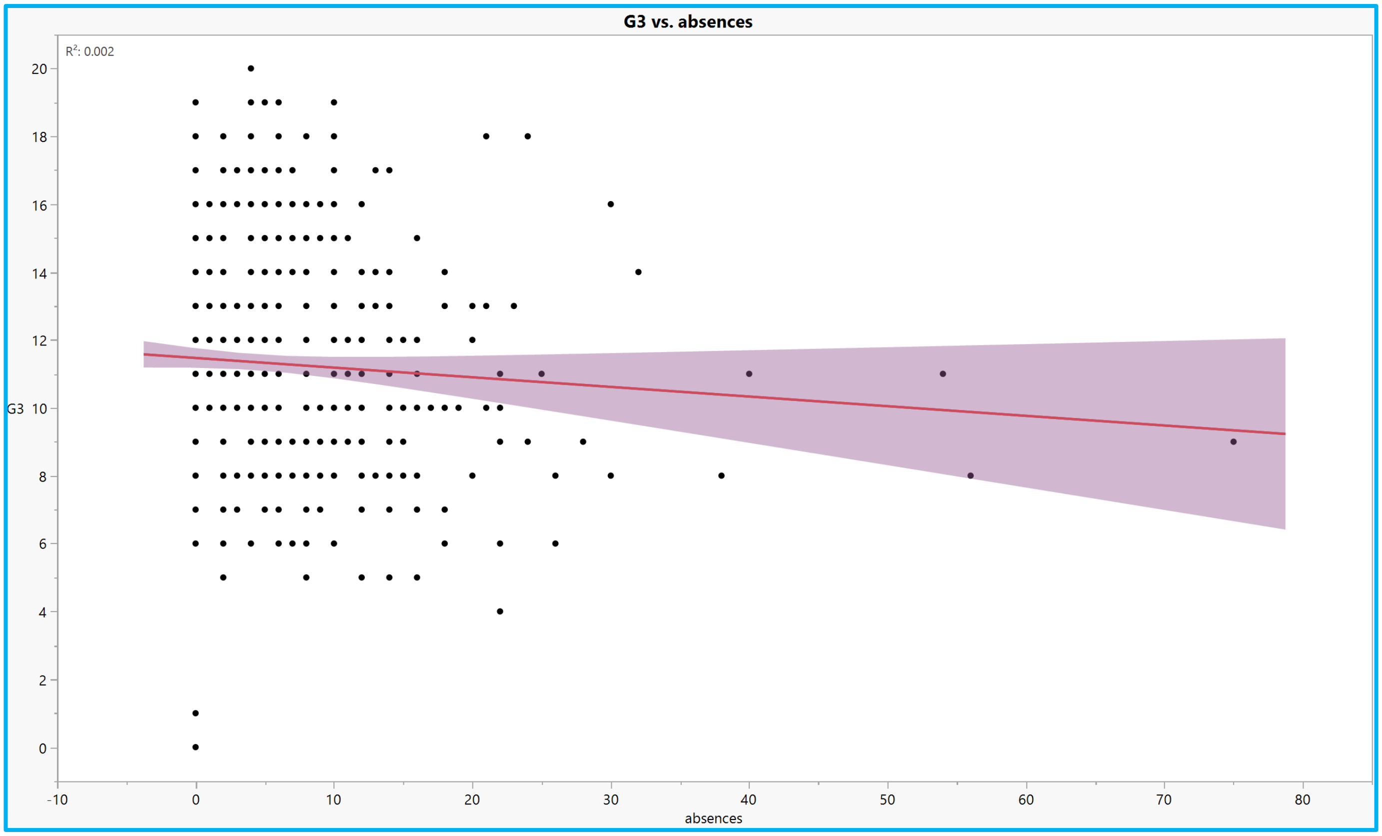
The Line Graph above represents students' average final grades, based on their workday alcohol consumption, 1 being very low and 5 being very high. Judging from this visualization, it appears that higher levels of workday alcohol consumption are correlated with lower average grades.

Distribution of Grades for Math and Portuguese grouped by School



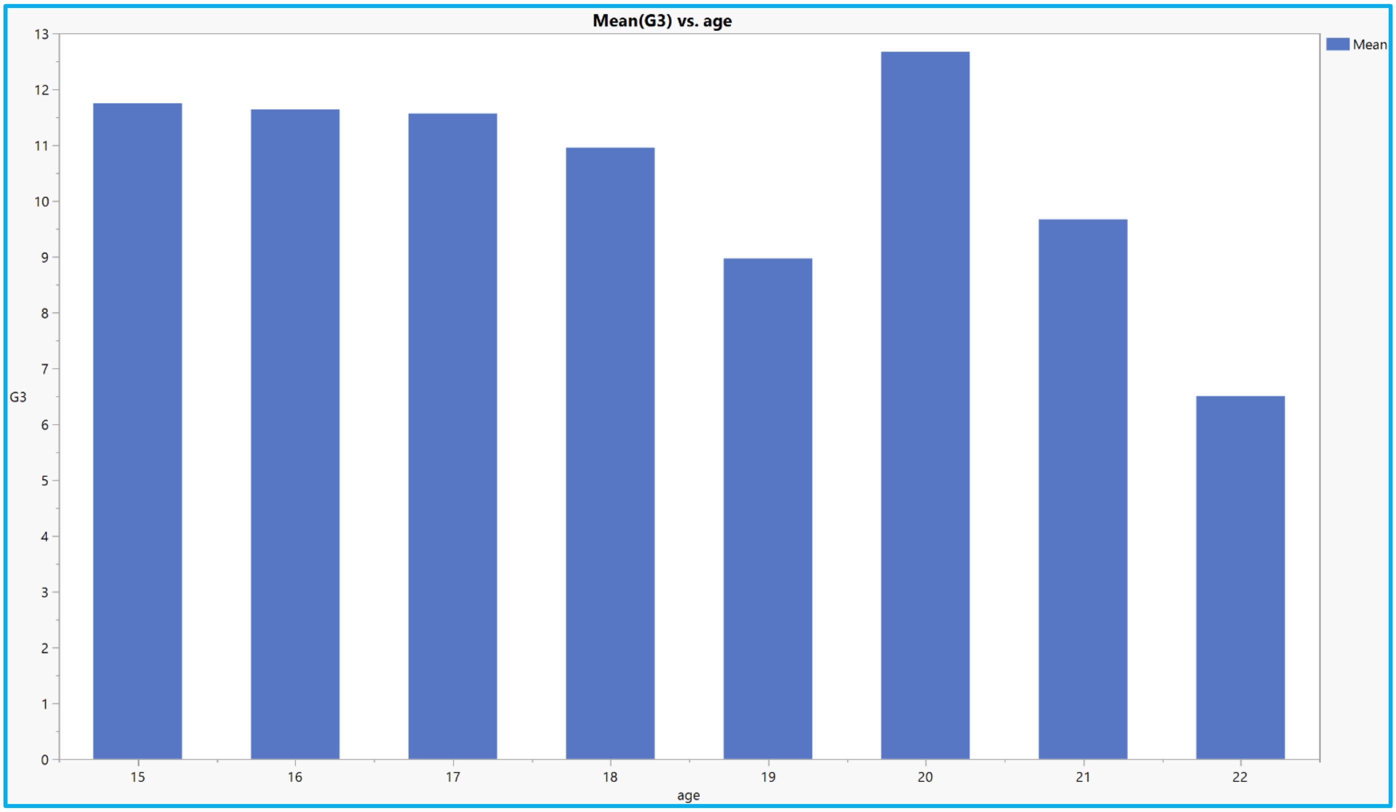
The above box plot shows the distribution of grades between the two school for each subject. The distributions are relatively similar, with the means hovering around 10-13 for each school in each subject.

Effect of Absences over Final Grades(G3)



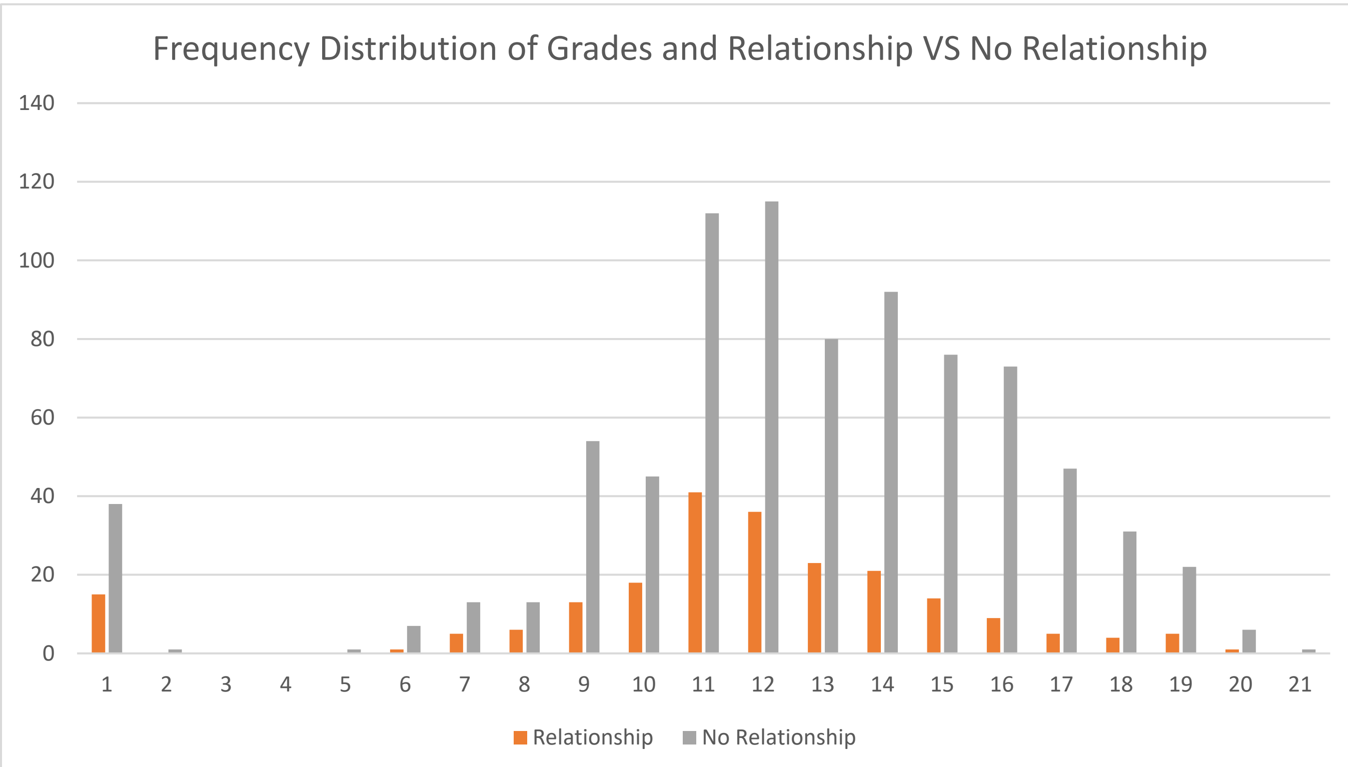
The scatter plot above shows the relationship between the number of absences a student has and the final grade they receive. The data is generally random but there is a there is a correlation between grades and the absences. The correlation is negative as the absences increases the grades decreases. R^2*: 0.0002*

Distribution of Grades by Student Age



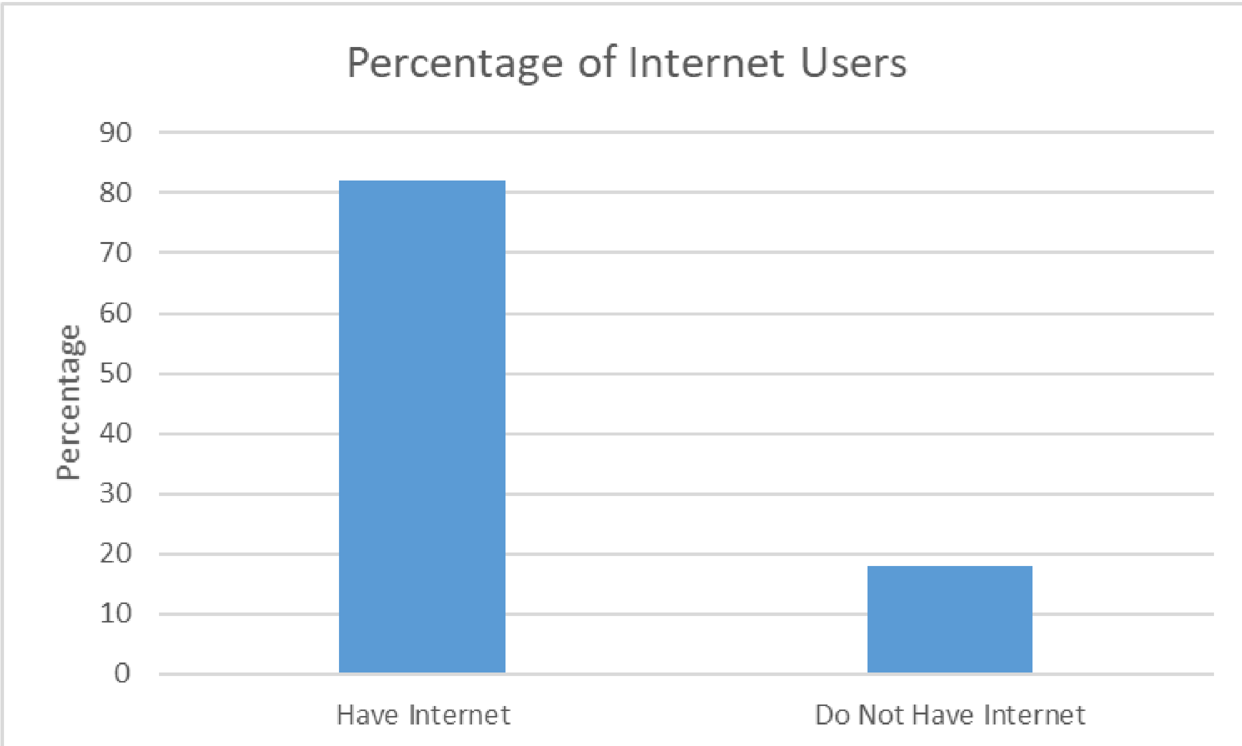
The above bar chart depicts the mean grade of each age group. As shown, 20-year-olds performed the best on average while 22-year-olds performed the worst.

Frequency Distribution of Grades and Relationship VS No Relationship



The Histogram compares the frequency distributions of Relationship and grade distribution and No Relationship and grade distribution. They are skewed left with outliers.

Percentage of Internet Users



The box plot shows a little over 80% students does have internet and about 20% do not have access to internet

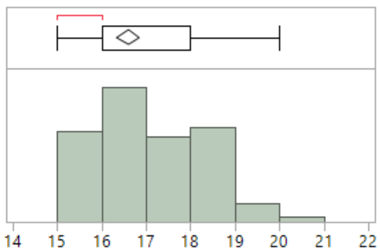
CONFIDENCE INTERVAL

THE CONFIDENCE INTERVALS BELOW WERE CALCULATED BY FIRST CREATING A RANDOM SAMPLE IN JMP PRO 14.1.0. THE SAMPLE SIZE IS 95 OBSERVATIONS.

A 90% confidence interval for the age of the students

Distributions

age



Quantiles

100.0%	maximum	20
99.5%		20
97.5%		19
90.0%		18
75.0%	quartile	18
50.0%	median	16
25.0%	quartile	16
10.0%		15
2.5%		15
0.5%		15
0.0%	minimum	15

Summary Statistics

Mean	16.6
Std Dev	1.2238759
Std Err Mean	0.125567
Upper 95% Mean	16.849316
Lower 95% Mean	16.350684
N	95

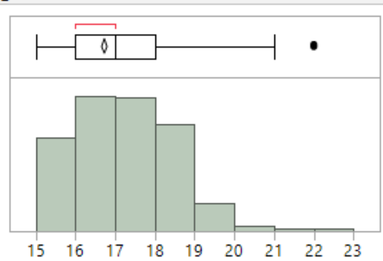
Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	16.6	16.3914	16.8086	0.900
Std Dev	1.223876	1.094055	1.39224	0.900

The 90% confidence interval for a sample size of 95 is (16.3914, 16.8086).

Distributions

age



Quantiles

100.0%	maximum	22
99.5%		20.775
97.5%		19
90.0%		18
75.0%	quartile	18
50.0%	median	17
25.0%	quartile	16
10.0%		15
2.5%		15
0.5%		15
0.0%	minimum	15

Summary Statistics

Mean	16.726054
Std Dev	1.2399747
Std Err Mean	0.0383763
Upper 95% Mean	16.801357
Lower 95% Mean	16.65075
N	1044

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	16.72605	16.66287	16.78923	0.900
Std Dev	1.239975	1.196995	1.286431	0.900

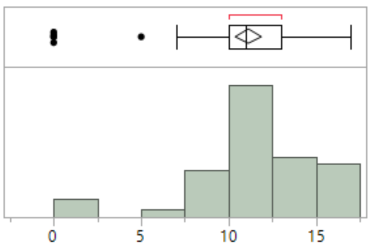
Population mean = 16.72 Sample mean = 16.66

Sampling error = 16.66 – 16.72 = - 0.06

A 99% confidence interval for the final grade (G3)

Distributions

G3



Quantiles

100.0%	maximum	17
99.5%		17
97.5%		17
90.0%		15
75.0%	quartile	13
50.0%	median	11
25.0%	quartile	10
10.0%		8
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	11.147368
Std Dev	3.6581835
Std Err Mean	0.3753217
Upper 95% Mean	11.892578
Lower 95% Mean	10.402158
N	95

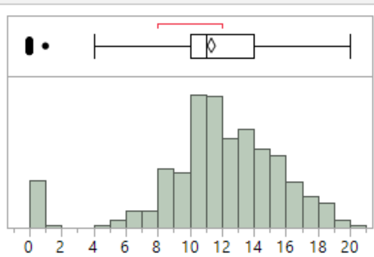
Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	11.14737	10.16059	12.13414	0.990
Std Dev	3.658184	3.074731	4.488576	0.990

The 99% confidence interval for a sample size of 95 is (10.1609, 12.1314).

Distributions

G3



Quantiles

100.0%	maximum	20
99.5%		19
97.5%		18
90.0%		16
75.0%	quartile	14
50.0%	median	11
25.0%	quartile	10
10.0%		8
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	11.341954
Std Dev	3.8647958
Std Err Mean	0.1196124
Upper 95% Mean	11.576662
Lower 95% Mean	11.107246
N	1044

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	11.34195	11.03329	11.65062	0.990
Std Dev	3.864796	3.657782	4.094817	0.990

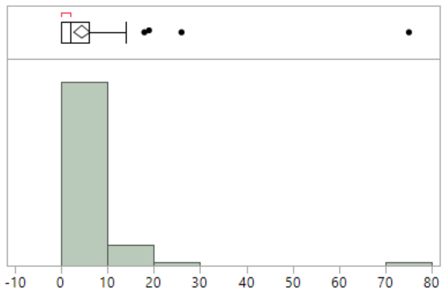
Population mean = 11.34 Sampling mean = 11.14

Sampling Error = 11.14 – 11.34 = -0.2

A 95% confidence interval for the number of absences students have.

Distributions

absences



Quantiles

100.0%	maximum	75
99.5%		75
97.5%		23.2
90.0%		10
75.0%	quartile	6
50.0%	median	2
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	4.5263158
Std Dev	8.6858188
Std Err Mean	0.8911462
Upper 95% Mean	6.2957075
Lower 95% Mean	2.7569241
N	95

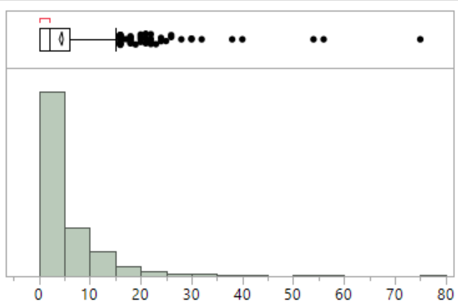
Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	4.526316	2.756924	6.295707	0.950
Std Dev	8.685819	7.601962	10.13298	0.950

The 95% confidence interval for a sample size of 95 is (2.7569, 6.2957).

Distributions

absences



Quantiles

100.0%	maximum	75
99.5%		36.650000000001
97.5%		20
90.0%		12
75.0%	quartile	6
50.0%	median	2
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	4.4348659
Std Dev	6.2100166
Std Err Mean	0.1921952
Upper 95% Mean	4.8119992
Lower 95% Mean	4.0577326
N	1044

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	4.434866	4.057733	4.811999	0.950
Std Dev	6.210017	5.9546	6.488497	0.950

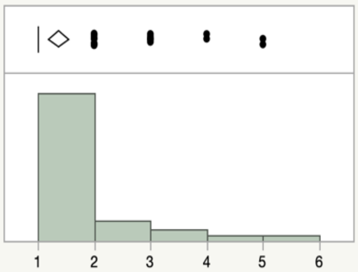
Population mean = 4.43 Sample Mean = 4.52

Sampling error = 4.52 – 4.43 = 0.09

A 95% Confidence interval for the Workday Alcohol Consumption of Students (Dalc)

Distributions

Dalc



Quantiles

100.0%	maximum	5
99.5%		5
97.5%		4.6
90.0%		2.4
75.0%	quartile	1
50.0%	median	1
25.0%	quartile	1
10.0%		1
2.5%		1
0.5%		1
0.0%	minimum	1

Summary Statistics

Mean	1.3578947
Std Dev	0.8494778
Std Err Mean	0.0871546
Upper 95% Mean	1.5309422
Lower 95% Mean	1.1848473
N	95

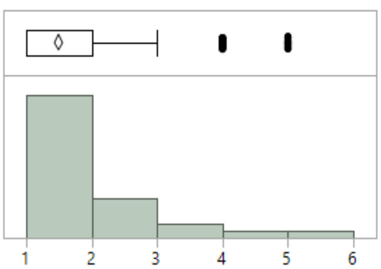
Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	1.357895	1.184847	1.530942	0.950
Std Dev	0.849478	0.743476	0.991011	0.950

The confidence interval for a sample size 95 at 95% is (1.1848, 1.5309)

Distributions

Dalc



Quantiles

100.0%	maximum	5
99.5%		5
97.5%		4.875
90.0%		3
75.0%	quartile	2
50.0%	median	1
25.0%	quartile	1
10.0%		1
2.5%		1
0.5%		1
0.0%	minimum	1

Summary Statistics

Mean	1.4942529
Std Dev	0.9117143
Std Err Mean	0.0282168
Upper 95% Mean	1.5496211
Lower 95% Mean	1.4388846
N	1044

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	1.494253	1.438885	1.549621	0.950
Std Dev	0.911714	0.874216	0.952599	0.950

Population mean = 1.49 Sample mean = 1.36

Sampling Error = 1.36 – 1.49 = - 0.13

HYPOTHESIS TESTING

1	DO STUDENTS WHOSE PARENTS LIVE TOGETHER (PSTATUS) GET BETTER GRADES?

Grades based on Pstatus: u1 = Parents Together (T); u2 = Parents Apart (A)		**Alpha = 0.05**
Ho: u1 = u2		
Ha: u1 > u2		
Alpha = 0.05		
Random sample size= 95		
t-Test: Two-Sample Assuming Unequal Variances		
	Pstatus = T	***Pstatus = A***
Mean	11.09638554	11.5
Variance	14.5271819	5.90909091
Observations	83	12
Hypothesized Mean Difference	0	
df	20	
t Stat	-0.494034381	
P(T<=t) one-tail	0.313332065	
t Critical one-tail	1.724718243	
P(T<=t) two-tail	0.62666413	
t Critical two-tail	2.085963447	
Decision: Failed to Reject the Null Hypothesis		
Entire Population		
t-Test: Two-Sample Assuming Unequal Variances		
	Pstatus = T	***Pstatus = A***
Mean	11.2990	11.6694
Variance	15.1773	13.0898
Observations	923.0000	121.0000
Hypothesized Mean Difference	0.0000	
df	159.0000	
t Stat	-1.0492	
P(T<=t) one-tail	0.1478	
t Critical one-tail	1.6545	
P(T<=t) two-tail	0.2957	
t Critical two-tail	1.9750	
Decision: Failed to Reject the Null Hypothesis		

The T Test for the sample data and the entire population indicates that there is not enough evidence to prove students whose parents live together performed better on their final grades than students who did not.

2 DO STUDENTS WITH INTERNET ACCESS GET BETTER GRADES?

Grades based on internet access: u1 = Yes; u2 = No		**Alpha = 0.01**
Ho: u1 = u2		
Ha: u1 > u2		
Random sample size= 95		
t-Test: Two-Sample Assuming Unequal Variances		
	Internet = Yes	***Internet = No***
Mean	11.26760563	10.79167
Variance	14.1416499	11.47645
Observations	71	24
Hypothesized Mean Difference	0	
df	44	
t Stat	0.578282901	
P(T<=t) one-tail	0.283010696	
t Critical one-tail	2.414134368	
P(T<=t) two-tail	0.566021392	
t Critical two-tail	2.692278266	
Decision: Failed to Reject the Null Hypothesis		
Entire Population		
t-Test: Two-Sample Assuming Unequal Variances		
	Internet = Yes	***Internet = No***
Mean	11.55380895	10.53456
Variance	14.75829964	14.86107
Observations	827	217
Hypothesized Mean Difference	0	
df	337	
t Stat	3.468957918	
P(T<=t) one-tail	0.0003	
t Critical one-tail	2.337463916	
P(T<=t) two-tail	0.0006	
t Critical two-tail	2.590496576	
Decision: Reject the Null Hypothesis		

The T-test for the sample data fails to reject the null whereas the T-test on the entire the population shows suggested to reject the hypothesis that students with internet access get better grades than students without internet access. This is a Type I error or false positive.

3	ARE THE MATH GRADES BETWEEN THE TWO SCHOOLS DIFFERENT	
Math Grades for GP = u1 and MS = u2		**Alpha = 0.1**
Ho: u1 = u2		
Ha: u1 ≠ u2		
Random sample size= 95		
t-Test: Two-Sample Assuming Unequal Variances		
	Math Grade GP	***Math Grade MS***
Mean	10.28125	11
Variance	16.66028226	7
Observations	32	3
Hypothesized Mean Difference	0	
df	3	
t Stat	-0.425454746	
P(T<=t) one-tail	0.349590201	
t Critical one-tail	1.637744354	
P(T<=t) two-tail	0.699180402	
t Critical two-tail	2.353363435	
Decision: Failed to Reject the Null Hypothesis		
Entire Population		
t-Test: Two-Sample Assuming Unequal Variances		
	Math Grade GP	***Math Grade MS***
Mean	10.48997135	9.847826
Variance	21.39429569	17.95411
Observations	349	46
Hypothesized Mean Difference	0	
df	60	
t Stat	0.955547525	
P(T<=t) one-tail	0.171567561	
t Critical one-tail	1.295821094	
P(T<=t) two-tail	0.343135123	
t Critical two-tail	1.670648865	
Decision: Failed Reject the Null Hypothesis		
The T-test for the sample data and the entire population shows there is not enough evidence to suggest that students in the two schools have different math scores.		

4	DO STUDENTS, WHO CONSUME HIGHER LEVELS OF ALCOHOL ON WORK DAYS (DALC) GET WORSE GRADES?	
Grades based on Dalc: u1 = Levels 1-2; u2 = Levels 3-5		**Alpha = 0.05**
Ho: u1 = u2		
Ha: u1 > u2		
Random sample size= 95		
t-Test: Two-Sample Assuming Unequal Variances		
	Levels 1-2	***Levels 3-5***
Mean	11.1627907	11
Variance	14.5378933	2.75
Observations	86	9
Hypothesized Mean Difference	0	
df	19	
t Stat	0.23630088	
P(T<=t) one-tail	0.407863956	
t Critical one-tail	1.729132812	
P(T<=t) two-tail	0.815727913	
t Critical two-tail	2.093024054	
Decision: Fail to Reject the Null Hypothesis		
Entire Population		
t-Test: Two-Sample Assuming Unequal Variances		
	Levels 1-2	***Levels 3-5***
Mean	11.46045504	10.43801653
Variance	15.42441064	10.3815427
Observations	923	121
Hypothesized Mean Difference	0	
df	170	
t Stat	3.193420478	
P(T<=t) one-tail	0.000837779	
t Critical one-tail	1.653866317	
P(T<=t) two-tail	0.001675559	
t Critical two-tail	1.974016708	
Decision: Reject the Null Hypothesis		

The T-Test for the sample data indicates that students who consume high levels of alcohol as defined by levels 3-5 performed the same on their final grades as students who did not. Interestingly, when the test is run for the entire population, a student who consumes high levels of alcohol as defined by levels 3-5 did perform worse on their final grades than students who did not. We get a Type II error or false negative.

5 DO STUDENTS, WHO ATTENDED NURSERY SCHOOL (NURSERY) GET BETTER GRADES?

Grades based on Nursery:
u1 = Yes; u2 = No

Alpha = 0.05

Ho: u1 = u2

Ha: u1 > u2

Random sample size= 95

t-Test: Two-Sample Assuming Unequal Variances

	Nursery = Yes	***Nursery = No***
Mean	11.6901	9.5417
Variance	11.4455	16.2591
Observations	71	24
Hypothesized Mean Difference	0	
df	35	
t Stat	2.3460	
P(T<=t) one-tail	0.0124	
t Critical one-tail	1.6896	
P(T<=t) two-tail	0.0248	
t Critical two-tail	2.0301	

Decision: Reject the Null Hypothesis

Entire Population

t-Test: Two-Sample Assuming Unequal Variances

	Nursery = Yes	***Nursery = No***
Mean	11.4192	11.0335
Variance	15.0927	14.2633
Observations	835	209
Hypothesized Mean Difference	0	
df	327	
t Stat	1.3127	
P(T<=t) one-tail	0.0951	
t Critical one-tail	1.6495	
P(T<=t) two-tail	0.1902	
t Critical two-tail	1.9672	

Decision: Fail to Reject the Null Hypothesis

The T-Test for the sample data indicates that students who attended Nursery schools performed better on their final grades than students who did not. Interestingly, when the test is run for the entire population, there is no significant difference in mean scores. The sample HT gave us a Type I error also called a false positive.