SESUG Paper 065-2021

Moving beyond Frequency and Percentage to Chi Square, t-Tests, and Correlation Analysis

Smith, Kelly D., Central Piedmont Community College

ABSTRACT

When institutional research offices are overwhelmed with data requests, outgoing reports may be limited to frequencies and percentages. While an important first step, additional analysis and a deeper understanding of the data are just a few steps away.

In this presentation, a typical data request for frequencies and percentages will be taken to the next level through the use of PROC FREQ, PROC CORR, and PROC TTEST. Data sets obtained from the UCI Machine Learning Repository are used for analysis. The UCI Student Performance data sets represent actual academic and demographic data of students from two Portuguese secondary schools. Participants will receive SAS® code for analysis and visualization.

INTRODUCTION

Offices of Institutional Research or Institutional Effectiveness are often more properly titled Institutional Reporting as the amount of time spent locating, collecting, preparing, analyzing, and summarizing data for internal and external reports can be substantial. As a consequence, the ability of these offices to respond to other data requests may be limited to providing descriptive analytics such as frequency and percentage.

With the increasing use of dashboards and a push for data democratization, analysts have the ability to move beyond surface level reporting and to dig deeper with inferential statistical tests. SAS offers multiple options for evaluating data’s suitability for inferential analysis and for then performing inferential analysis. In this discussion, the PROC UNIVARIATE, PROC CORR, PROC TTEST, and PROC FREQ procedures are applied to the Student Performance Data Set (Cortez & Silva 2008) obtained from the University of California Irvine Machine Learning Repository (2014).

The Student Performance Data Set was used to create a new data set, MathPort, consisting of 366 distinct records and 37 variables including a research ID for each observation. Table 1 lists key variables and their attributes from the MathPort data set. Relevant SAS code and selected SAS output is presented for data preparation and inferential analysis.

Name	Type	Definition	Notes
School	Character	Student’s school	GP, MS
Sex	Character	Student’s gender	F, M
Age	Numeric	Student’s age	Integer, 15 – 22
Activities	Character	Extracurricular activities	Yes, No
Absences	Numeric	Number of absences	Integer, 0 – 93
StudyTime	Ordinal	Weekly study time	Low (1) to high (4)
WALC	Ordinal	Weekend alcohol consumption	Very low (1) to very high (5)
G1_Port	Numeric	First grade, Portuguese	Integer, 0 – 20
G2_Port	Numeric	Second grade, Portuguese	Integer, 0 – 20
G3_Port	Numeric	Final grade, Portuguese	Integer, 0 – 20
G1_Math	Numeric	First grade, Math	Integer, 0 – 20
G2_Math	Numeric	Second grade, Math	Integer, 0 – 20
G3_Math	Numeric	Final grade, Math	Integer, 0 – 20

Table 1. Key Variables from Student Performance Data Set (UCI).

DATA PRE-CHECK

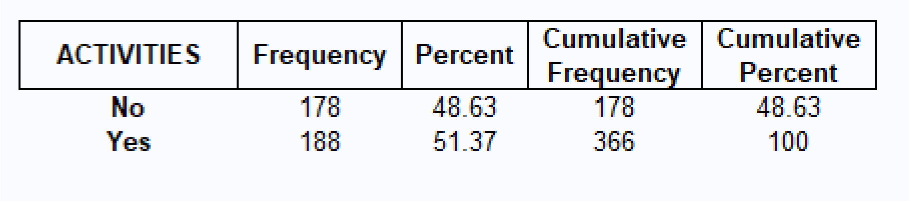
Evaluating data suitability for analysis is key when moving beyond frequency and percentage. As a first step, PROC FREQ can be used to check numeric and character variables for missing values. By including the chi-square option, the PROC FREQ can also check goodness of fit for categorical variables (as here, determining if females and males are equally represented within the dataset, or within each school):

```
PROC FREQ DATA=MathPort;
TABLES School Sex Activities StudyTime WALC;
RUN;


PROC FREQ DATA=MathPort;
TABLES Sex / CHISQ TESTP=(50 50);
RUN;

PROC FREQ DATA=MathPort;
BY School;
TABLES Sex / CHISQ TESTP=(50 50);
RUN;
```

Figure 1 presents the results from the PROC FREQ analysis of Activities. As with the other variables analyzed, Activities has no missing values.



ACTIVITIES	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	178	48.63	178	48.63
Yes	188	51.37	366	100

Figure 1. PROC FREQ analysis showing no missing values of Activities variable.

Once missing data has been analyzed and dealt with as appropriate, the next step in data preparation is often checking numeric variables for centrality measurements (mean, median, mode, standard deviation) and normality. Pre-analysis of numeric variables is key because parametric statistical tests require analytical data to meet requirements of normal distribution. Non-parametric tests do not require normality and provide a better option for robust analysis when data does not meet the normal distribution standard.

Key assumptions of the statistical tests applied to MathPort are included in Table 2. Parametric tests can be identified by the highlighted "***normal distribution***" assumption. Chi square analysis is the only non-parametric test in Table 2.

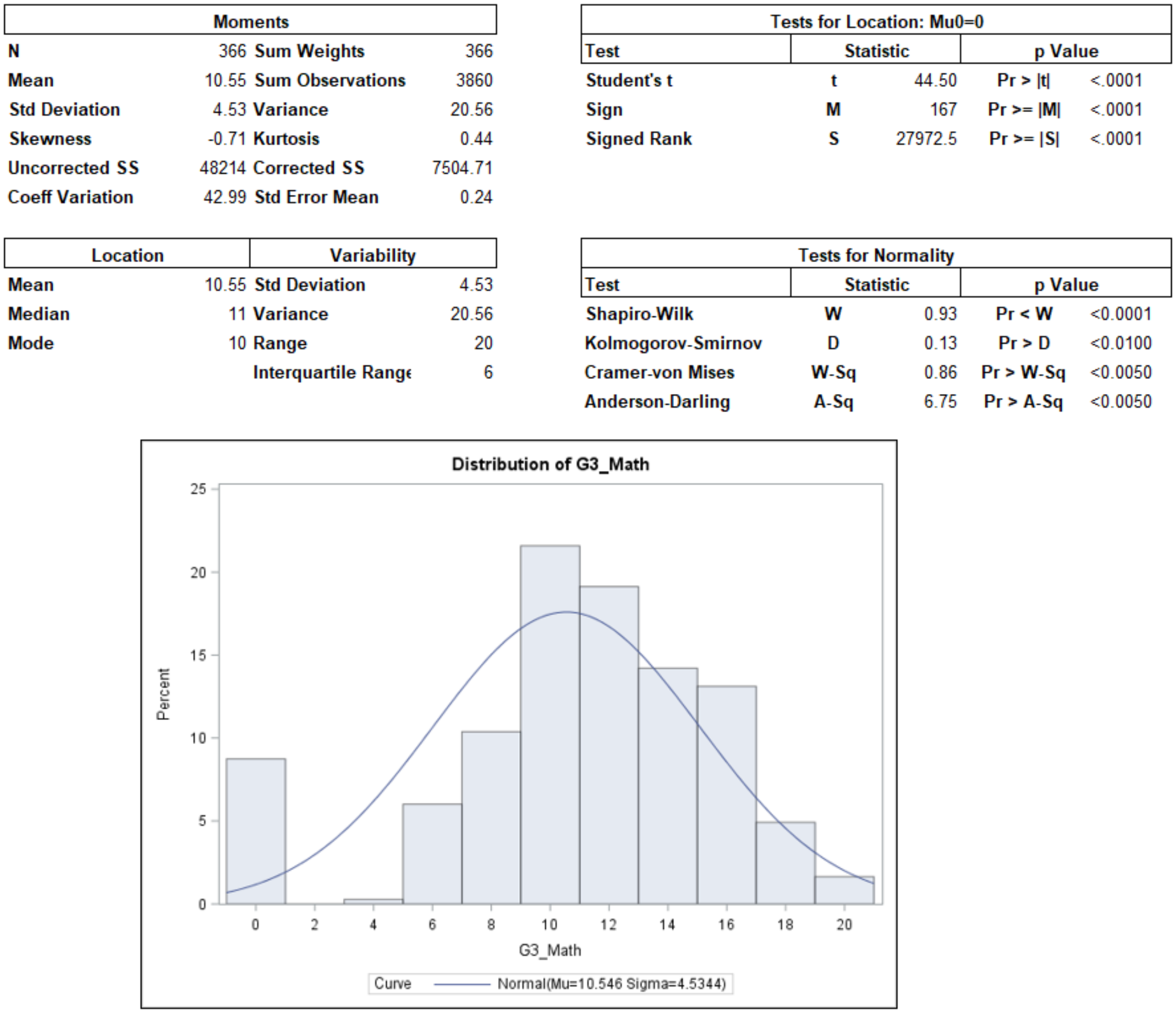
Test	**Assumptions**
Pearson Correlation	Random sample of population, independent scores
	Bivariate ***normal distribution*** of variables
Chi Square	Random sample of population, independent scores
	Test statistic approximates chi square distribution (when *N* is large)
One Sample t-Test	Random sample of population, independent scores
Two Sample t-Test	***Normal distribution*** of variables
Paired Samples t-Test	***Normal distribution*** of test variable or difference scores (as appropriate)

Table 2. Underlying assumptions of selected statistical tests.

PROC UNIVARIATE can be used to determine measures of centrality and to assess normality of numeric variables. Including the Normal option after Histogram adds an overlay of a normal distribution curve on the histogram display.

```
PROC UNIVARIATE NORMAL PLOT DATA=MathPort;
VAR G3_Math G3_Port Absences;
HISTOGRAM / NORMAL;
RUN;
```

A portion of the SAS output for the analysis of G3_Math is presented in Figure 2. The presence of extreme outliers (value = 0) distorts the G3_Math variable from a normal distribution, as reflected in skewness and kurtosis values distinct from zero and in the normality tests. The G3_Port variable also had extreme outliers with a value equal to zero which resulted in a skewed distribution. In a real-life scenario, the analyst would reassess the use of these variables or consider variable transformations in an attempt to obtain a normal distribution. For our purposes however, we note that G3_Math and G3_Port do not have normal distributions and proceed with further testing.



```
                         Moments                                            Tests for Location: Mu0=0
N                       366  Sum Weights          366     Test              Statistic            p Value
Mean                  10.55  Sum Observations    3860     Student's t    t      44.50      Pr > |t|     <.0001
Std Deviation          4.53  Variance           20.56     Sign           M        167      Pr >= |M|    <.0001
Skewness              -0.71  Kurtosis            0.44     Signed Rank    S    27972.5      Pr >= |S|    <.0001
Uncorrected SS        48214  Corrected SS     7504.71
Coeff Variation       42.99  Std Error Mean      0.24

           Location               Variability                       Tests for Normality
Mean                  10.55  Std Deviation       4.53     Test                  Statistic           p Value
Median                   11  Variance           20.56     Shapiro-Wilk          W      0.93    Pr < W      <0.0001
Mode                     10  Range                 20     Kolmogorov-Smirnov    D      0.13    Pr > D      <0.0100
                             Interquartile Range    6     Cramer-von Mises      W-Sq   0.86    Pr > W-Sq   <0.0050
                                                          Anderson-Darling      A-Sq   6.75    Pr > A-Sq   <0.0050
```

Figure 2. Portion of PROC UNIVARIATE output for analysis of G3_Math variable.

After analytical data has been checked for completeness and suitability, the proper inferential test can be selected based on the analytical hypothesis or question under consideration. Table 3 lists potential analytical questions and the appropriate test.

Question	**Statistical Test**
Is there a linear relationship between Math and Language (Portuguese) scores?	Pearson Correlation
Is there an association (relationship) between student gender (sex) and extracurricular activities?	Chi Square
Is the mean final Math grade at GP significantly different from the state standard (score of 11)?	One Sample t-Test
Do students at GP and MS schools have similar final Math grades?	Two Sample t-Test
Do GP students have similar final Math and Language grades?	Paired Samples t-Test

Table 3. Connecting Analytical Questions and Statistical Tests.

CORRELATION

Two numeric variables are correlated if a linear association or relationship can be defined between them. The strength and direction of the relationship is reflected in the sign and magnitude of the correlation coefficient ($|r| \leq 1$). To determine if a linear relationship exists between Math and Language (Portuguese) scores, the following code can be used:

```
PROC CORR DATA=MathPort FISHER PLOTS=ALL;
VAR  G3_Math;
WITH G3_Port;
RUN;
```

The PLOTS=ALL option will produce a scatter plot and a scatter plot matrix. Since the FISHER option is included, SAS will calculate a confidence interval for the correlation coefficient. Figure 3 presents a portion of the SAS output, including the scatter plot with a 95% prediction ellipse overlay. The results indicate G3_Math and G3_Port have a positive, moderately strong correlation (r = 0.49).

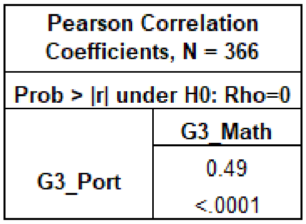
PROC CORR is not restricted to one set of pairwise analysis. If multiple variables are listed after WITH, the VAR variable will be analyzed with each of the listed variables:

```
PROC CORR DATA=MathPort FISHER PLOTS=ALL;
VAR  G3_Math;
WITH G1_Math G2_Math;
RUN;
```

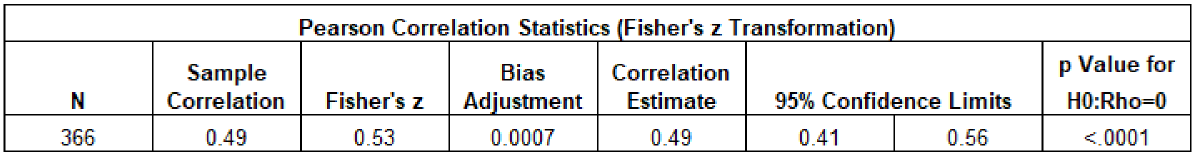
This second analysis will analyze the G3_Math / G1_Math and G3_Math / G2_Math variable pairings. If multiple variables are listed after VAR and WITH is omitted, SAS will determine the correlation coefficient for each possible variable pair:

```
PROC CORR DATA=MathPort FISHER PLOTS=ALL;
VAR  G1_Math G2_Math G3_Math;
RUN;
```

In this final analysis, three variable pairs will be analyzed: G1_Math / G2_Math, G1_Math / G3_Math, and G2_Math / G3_Math.



Pearson Correlation Coefficients, N = 366 Prob > |r| under H0: Rho=0	G3_Math
G3_Port	0.49 <.0001



Pearson Correlation Statistics (Fisher's z Transformation)							
N	Sample Correlation	Fisher's z	Bias Adjustment	Correlation Estimate	95% Confidence Limits		p Value for H0:Rho=0
366	0.49	0.53	0.0007	0.49	0.41	0.56	<.0001

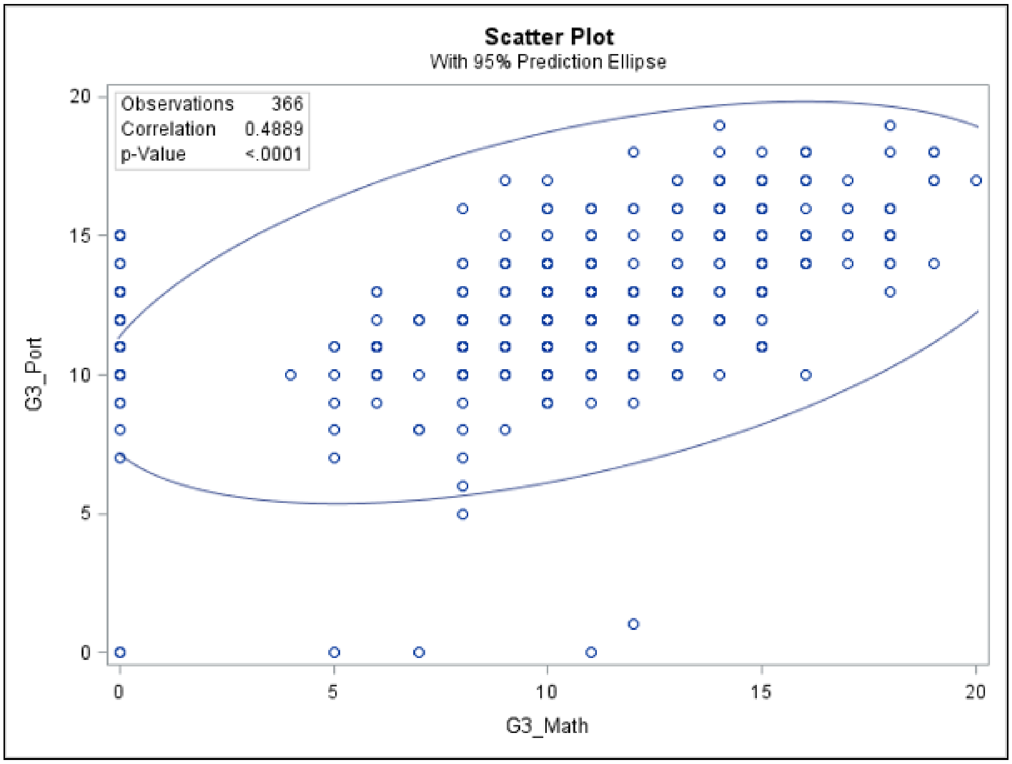


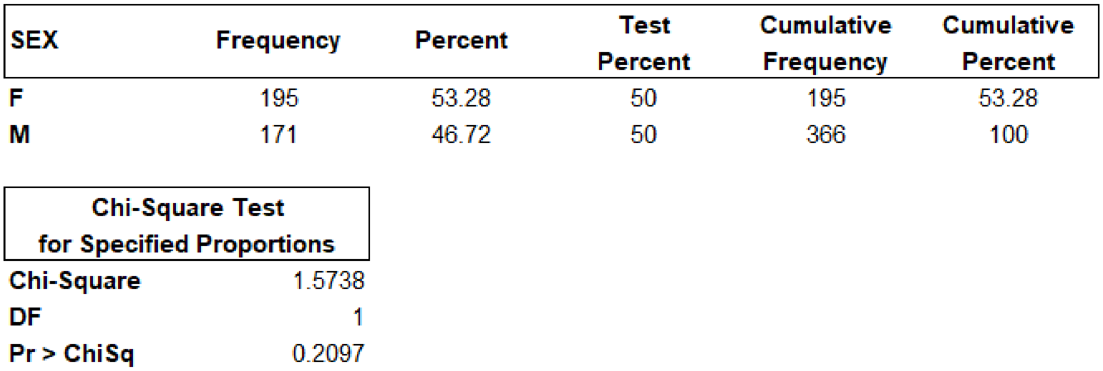
Figure 3. Partial PROC CORR output for analysis G3_Math and G3_Port.

CHI SQUARE

Chi square analysis is used with categorical variables and can answer several different types of questions. Chi square analysis is an option available in the PROC FREQ procedure. As mentioned earlier, a goodness of fit analysis for one variable can be conducted when the expected proportions are included after CHISQ. The following code examines the gender distribution of MathPort and compares it to the expected 1:1 ratio:

```
PROC FREQ DATA=MathPort;
TABLES Sex / CHISQ TESTP=(50 50);
RUN;
```

Figure 4 presents key SAS output for the goodness of fit analysis of student gender. Since the *p* value for the analysis is greater than 0.05, the gender distribution of students is not significantly different from the expected 1:1 ratio.



SEX	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
F	195	53.28	50	195	53.28
M	171	46.72	50	366	100

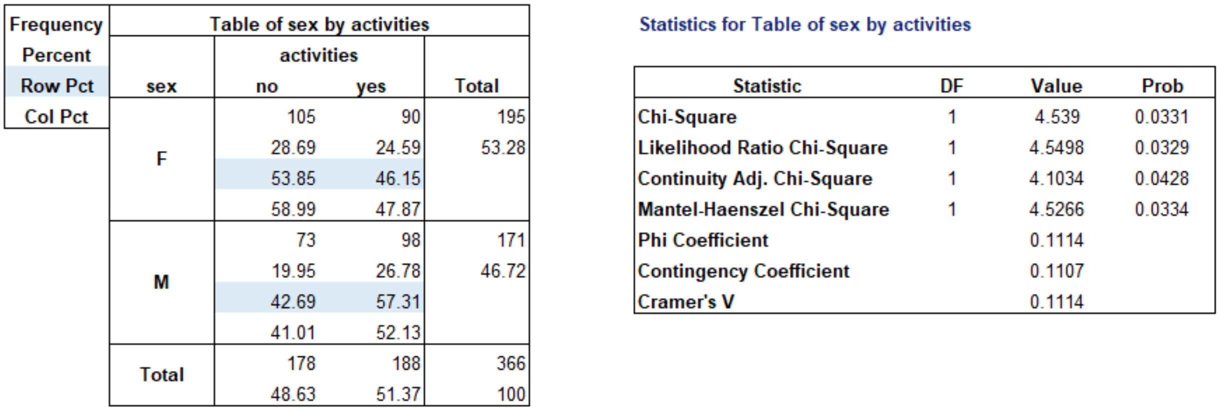
Chi-Square Test for Specified Proportions	
Chi-Square	1.5738
DF	1
Pr > ChiSq	0.2097

Figure 4. Partial SAS output from goodness of fit analysis of student gender,

Chi square testing can also be used to determine if two categorical variables are associated with each other. The variables being analyzed do not have to have the same number of levels. To determine if a relationship exists between student gender (sex) and extracurricular activities, the following code was used:

```
PROC FREQ DATA=MathPort;
TABLES Sex*Activities / CHISQ;
RUN;
```

Figure 5 displays all the SAS output from the analysis except the Fisher's Exact Test results. Those results are not needed to draw a conclusion because no warning was given above the table (Waller, 2015). Since we are looking for a potential association between the two variables, we look at the row percentages (third line) of each cell (highlighted in blue). The *p* value for the chi square statistic (0.0331) is less than 0.05, indicating a relationship does exist between student gender and extracurricular activity involvement. Male students are more likely to be involved in extracurricular activities (57.31%) than female students (46.15%).



Frequency
Percent
Row Pct
Col Pct

Table of sex by activities

sex	activities: no	yes	Total
F	105	90	195
	28.69	24.59	53.28
	53.85	46.15	
	58.99	47.87	
M	73	98	171
	19.95	26.78	46.72
	42.69	57.31	
	41.01	52.13	
Total	178	188	366
	48.63	51.37	100

Statistics for Table of sex by activities

Statistic	DF	Value	Prob
Chi-Square	1	4.539	0.0331
Likelihood Ratio Chi-Square	1	4.5498	0.0329
Continuity Adj. Chi-Square	1	4.1034	0.0428
Mantel-Haenszel Chi-Square	1	4.5266	0.0334
Phi Coefficient		0.1114	
Contingency Coefficient		0.1107	
Cramer's V		0.1114	

Figure 5. SAS output from Chi Square test of association between student gender (sex) and extracurricular activity.

T-TESTS

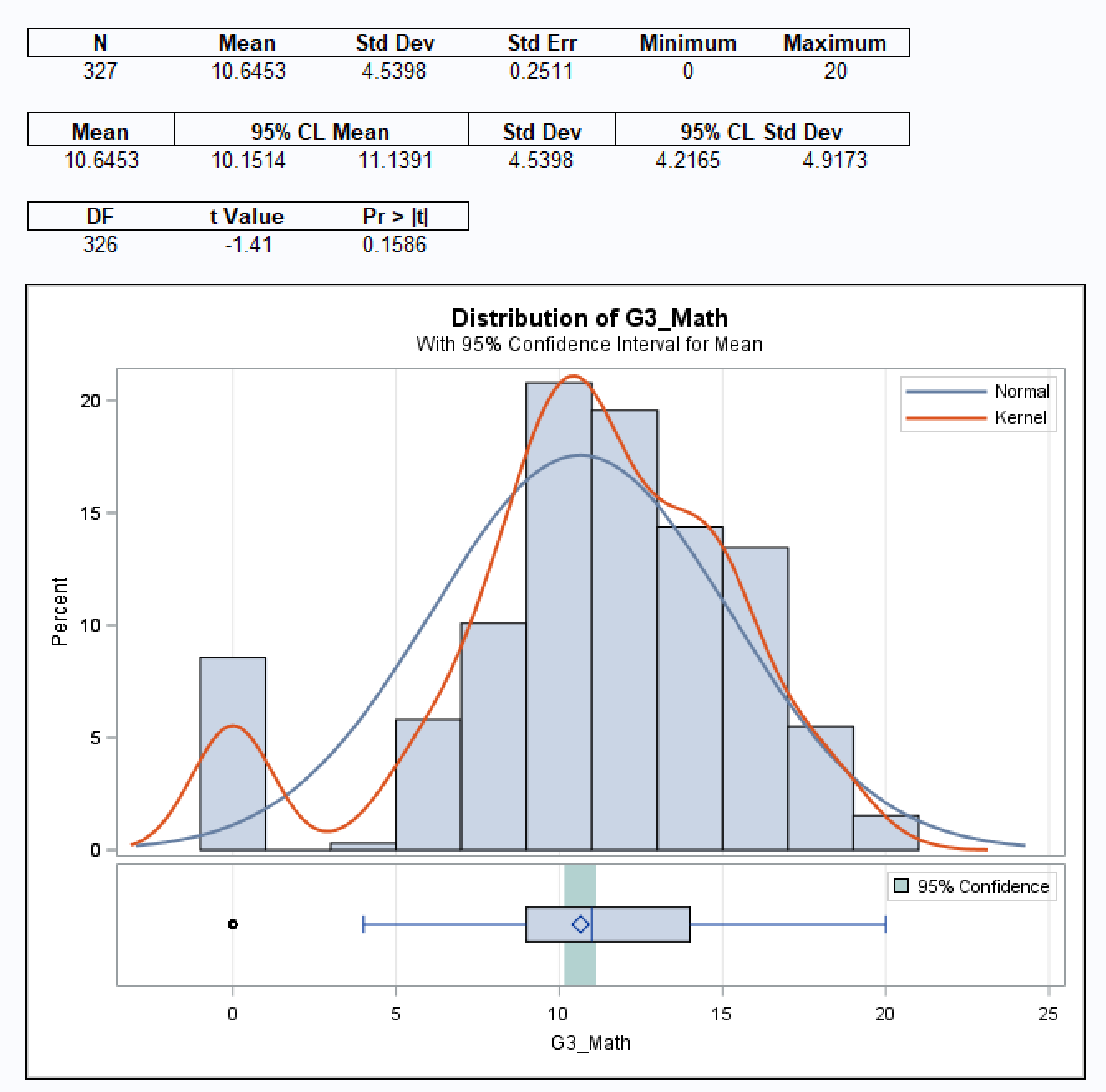
The inferential t-test involves continuous numeric variables and is used to examine differences between groups, between one group and a standard value, or between two separate observations of one group. PROC TTEST can be used for all types of t-test analysis.

ONE SAMPLE T-TEST

A one-sample t-test compares the mean of a test value to the expected (standard) value. The following code compares the mean final math grade for students in GP school to the state standard (11).

```
PROC TTEST DATA=MathPort H0=11 PLOTS=Summary;
WHERE School EQ 'GP';
VAR   G3_Math;
RUN;
```

The PLOTS=SUMMARY option produces a histogram and a box plot with a visualized 95% confidence interval for the G3_Mean (Figure 6).



```
   N        Mean      Std Dev    Std Err    Minimum    Maximum
  327     10.6453     4.5398     0.2511        0          20

  Mean         95% CL Mean       Std Dev     95% CL Std Dev
10.6453    10.1514    11.1391    4.5398     4.2165    4.9173

  DF       t Value    Pr > |t|
  326       -1.41      0.1586
```

Figure 6. SAS output from One Sample t-Test analysis of GP student final math grades (G3_Math).

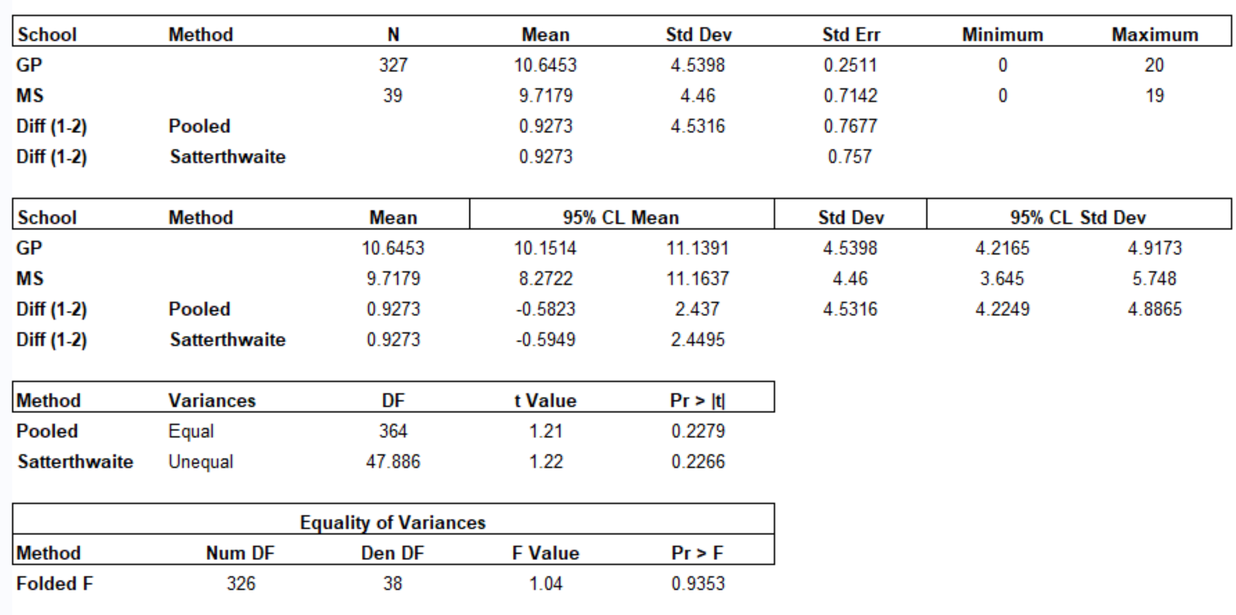
As Figure 6 shows, the G3_Math variable does not have a normal distribution so a t-test is not the best option for this analysis in a real-world scenario. If we work from the presumption that G3_Math does have a normal distribution, the results of the analysis show that the mean G3_Math (10.6453) is not significantly different from the state standard score of 11 because *p* > 0.05.

TWO SAMPLE T-TEST

The two-sample t-test is also known as the independent samples t-test. The two-sample t-test compares the mean of two groups with the same continuous test variable. The following code compares the mean final math grade for GP and MS students.

```
PROC TTEST DATA=MathPort PLOTS=Summary;
CLASS   School;
VAR     G3_Math;
RUN;
```

In this case, a histogram and box plot (with 95% confidence interval) for each level of School will be produced by PLOTS=SUMMARY. Figure 7 displays the analytical tables from the analysis of G3_Math.



School	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
GP		327	10.6453	4.5398	0.2511	0	20
MS		39	9.7179	4.46	0.7142	0	19
Diff (1-2)	Pooled		0.9273	4.5316	0.7677		
Diff (1-2)	Satterthwaite		0.9273		0.757		

School	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
GP		10.6453	10.1514	11.1391	4.5398	4.2165	4.9173
MS		9.7179	8.2722	11.1637	4.46	3.645	5.748
Diff (1-2)	Pooled	0.9273	-0.5823	2.437	4.5316	4.2249	4.8865
Diff (1-2)	Satterthwaite	0.9273	-0.5949	2.4495			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	364	1.21	0.2279
Satterthwaite	Unequal	47.886	1.22	0.2266

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	326	38	1.04	0.9353

Figure 7. Partial SAS output from two sample t-test analysis of G3_Math.

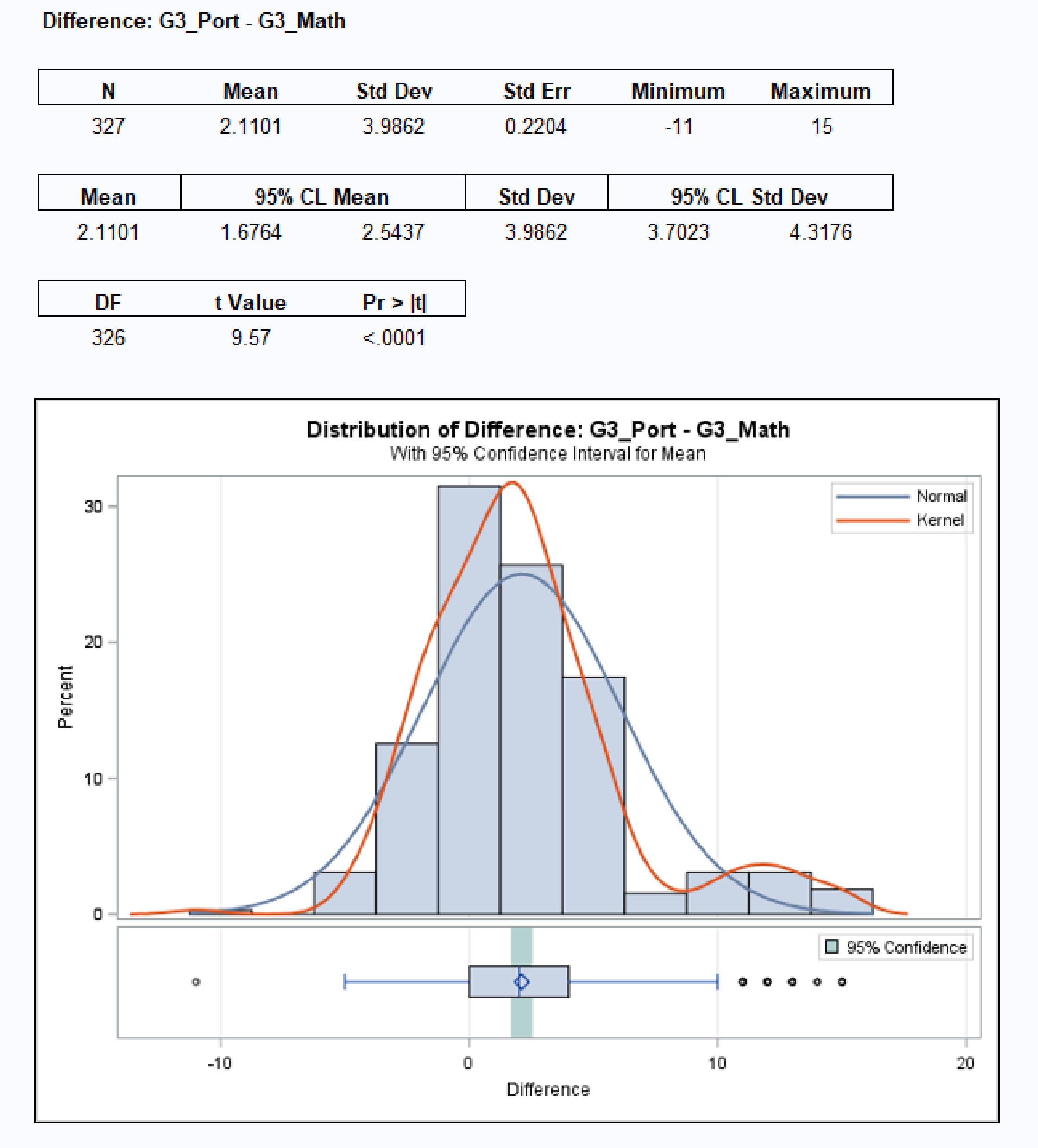
Since the *p* value for the equality of variances test (0.9353) is greater than 0.05, we conclude the variances of the two groups (GP, MS) are equal. The Pooled test can be used because the group variances are equal. The Satterthwaite test, which assumes unequal variances, is more conservative and many analysts report out those results even when the equality of variances test indicate the Pooled test may be used (Waller, 2015). In this instance, both the Pooled and Satterthwaite test have *p* values greater than 0.05, indicating there is not a significant difference in the mean G3_Math values for GP and MS students.

PAIRED SAMPLES T-TEST

The paired sample t-test requires two continuous variables and one categorical variable. Each value of the categorical value must have a non-missing value in each continuous variable. Rather than comparing the two means, the paired samples t-test examines the difference between paired readings of the continuous variables. The mean of the difference scores is compared to a defined standard value. In the following code, the difference between each student's final math and language (Portuguese) grade is compared to the expected difference of zero.

```
PROC TTEST DATA=MathPort SIDES=2 ALPHA=0.05 H0=0;
WHERE School EQ 'GP';
PAIRED G3_Port*G3_Math;
RUN;
```

As Figure 8 shows, the mean difference in final math and language grades is 2.1101. With a *p* value less than 0.05, we conclude there is a significant difference between students' final math and language scores.



Difference: G3_Port - G3_Math

N	Mean	Std Dev	Std Err	Minimum	Maximum
327	2.1101	3.9862	0.2204	-11	15

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
2.1101	1.6764	2.5437	3.9862	3.7023	4.3176

DF	t Value	Pr > |t|
326	9.57	<.0001

Figure 8. Partial SAS output from paired samples t-test analysis of G3_Math and G3_Port.

CONCLUSION

Colleges are in the process of developing data cultures through data literacy training, dashboard implementation, and data democratization. These changes are permitting Institutional Research offices to move beyond surface level analysis and to incorporate inferential analytical testing. Successful inferential analysis requires checking data suitability against test assumptions (see Table 2) by such steps as addressing missing values and checking variable distributions. Once data suitability has been established, inferential analysis provides actionable information and results, as demonstrated by the questions and answers in Table 4.

Question	**Answer**	**Statistical Test**
Are male and female students equally represented in the data set (MathPort)?	Yes	Chi Square, Goodness of Fit
Is there a linear relationship between Math and Language (Portuguese) scores?	Yes, $r = 0.49$	Pearson Correlation
Is there an association (relationship) between student gender (sex) and extracurricular activities?	Yes, male students are more likely to participate in extracurricular activities than female students	Chi Square
Is the mean final Math grade at GP significantly different from the state standard (score of 11)?	No, the GP mean final math grade is not significantly different from the state standard	One Sample t-Test
Do students at GP and MS schools have similar final Math grades?	Yes, there is no significant difference in GP and MS final Math grades	Two Sample t-Test
Do GP students have similar final Math and Language grades?	No, there is a significant difference between GP students' Math and Language final grades	Paired Samples t-Test

Table 4. Results from inferential analysis of MathPort data set.

REFERENCES

Chen, Z. (2021). *Reporting correlation coefficient results and plots – a SAS® macro that does it all.* Retrieved from https://communities.sas.com/t5/SAS-Global-Forum-Proceedings/

Cody, R. (2017). *Cody's Data Cleaning Techniques Using SAS®* (3rd Ed.). Cary, NC: SAS Institute, Inc.

Cortex, P., & Silva, A. (2008, April). Using Data Mining to Predict Secondary School Student Performance. In Brito, A., & Teixeira, J. (Eds*.), Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)* (pp. 5-12). Porto, Portugal: EUROSIS, ISBN 978-9077381-39-7.

DePuy, V., & Pappas, P. A. (2017*). Perusing, choosing, and not mis-using: non-parametric vs. parametric tests in SAS®*. Retrieved from https://www.lexjansen.com/nesug/nesug04/an/an10.pdf

University of California Irvine Machine Learning Repository (2014). Student Performance Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/Student+Performance

Waller, J. L. (2015). *Chi-Square and T-Tests using SAS®: Performance and Interpretations.* Retrieved from https://www.lexjansen.com/wuss/2015/122_Final_Paper_PDF.pdf

RECOMMENDED READING

- *Cody's Data Cleaning Techniques Using SAS® (3rd Ed.)*
- *Chi-Square and T-Tests using SAS®: Performance and Interpretations*
- *Perusing, Choosing, and Not Mis-Using: Non-parametric vs. Parametric Tests in SAS®*
- *Reporting Correlation Coefficient Results and plots – a SAS® macro that does it all*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Kelly D. Smith
Central Piedmont Community College
kds.aewas@gmail.com
kelly.smith@cpcc.edu
www.linkedin.com/in/kelly-d-smith