A Modified Marascuilo Procedure for Data Mining in R

Corresponding Author: Victor L. Landry, PhD**,** College of Doctoral Studies, Grand Canyon University, 3300 West Camelback Road, Phoenix, Arizona 85607 USA

Email: victor.landry@my.gcu.edu.

Abstract

This paper discusses the Marascuilo Procedure and introduces the Modified Marascuilo Procedure as an education data mining tool. An application for data exploration using data from the Arizona Department of Education Instrument for Measuring Science (AIMS) proficiency is given. This modified tool runs in R and is not currently available in CRAN. The code for both the standard and the modified Marascuilo Procedure, along with information regarding acquiring RStudio, is attached in the appendices.

Key Words

Marascuilo Procedure, Modified Marascuilo Procedure, post hoc, data exploration, RStudio, aggregate, Chi Square, R, proportions, iteration, Z-score, public domain data, education, research, science proficiency, Arizona, primary education, secondary education, AIMS.

INTRODUCTION

The Standard Marascuilo Procedure and Its Limitation

The Marascuilo Procedure is a familiar iterative *post hoc* statistical test of significance for comparing multiple *k* associated group pairings within a test matrix (Marascuilo, 1966, Marascuilo and Serlin, 1998). It is limited, however, to six Chi Square rows.

By program iteration it evaluates, for *k(k-1)/2* unique and non-repeated values, the absolute difference between p_i and p_j proportions and tests each against the corresponding critical range margin of error. This is obtained by converting $\chi^2_{1-\alpha,k-1}$ to a Z-score and multiplying that by the standard deviation of p_i and p_j proportions. If $|p_i - p_j|$ > the critical range, then that comparison is statistically significant.

An alpha level of significance is typically set at 0.05 and the routine programmatically calculates the Chi Square statistic for the required degrees of freedom. A Z-score is obtained from the square root of $\chi^2_{1-\alpha,k-1}$.

$$Z = \sqrt{X^2_{1-\alpha,k-1}}\ .$$

The standard deviation of two unique population proportions, $p_i - p_j$ (where $i \neq j$) is calculated as

$$\sqrt{\frac{p_i(1-p_i)}{n_i} + \frac{p_j(1-p_j)}{n_j}}.$$

In combination, therefore, the margin of error (r_{ij}) is

$$r_{ij} = \sqrt{\chi^2_{1-a,k-1}}\sqrt{\frac{p_i(1-p_i)}{n_i} + \frac{p_j(1-p_j)}{n_j}}.$$

Finally, this is compared against the absolute difference between proportions $|p_i - p_j|$ for significance determination

$$|p_i - p_j| > r_{ij}\ .$$

If the absolute difference is greater than the margin of error, then the proportional difference is determined to be statistically significant.

The Marascuilo Procedure was meant to be a Chi Square *post hoc* test, not a data mining tool. As number of groups (N) increase, the associated Z-score becomes increasingly restrictive. Within a matrix of 6 rows, the margin of error for proportional difference evaluation equates to $p = 0.0001$. The relationship between the Chi Square statistics is shown in Table 1.

Table 1

The Relationship Between Sample (N), Degrees Of Freedom (df), Chi Square ($X^2_{0.05}$), Z-Score , Cumulative Distribution And Two-Tailed α.

N	*df*	$\chi^2_{0.05}$	*Z*	*P(Z ≤ z)*	*α*
2	1	3.841	1.960	0.9500	0.050
3	2	5.991	2.448	0.9928	0.007
4	3	7.815	2.796	0.9974	0.002
5	4	9.488	3.080	0.9990	0.001
6	5	11.070	3.327	0.9996	0.000
7	6	12.592	3.549	0.9998	0.000
8	7	14.067	3.751	0.9999	0.000
9	8	15.507	3.938	*	0.000
10	9	16.919	4.113	*	0.000
11	10	18.307	4.279	*	0.000
12	11	19.675	4.436	*	0.000
13	12	21.026	4.585	*	0.000

**Less than one with 5 or more decimals. The Z-distribution is two-tailed.*

The practical limit to a *post hoc* Marascuilo Procedure is a matrix with no more than six members. The R code for the standard Marascuilo Procedure is found in Appendix A.

As an illustration, summary data obtained from the Texas Department of Education reported the number of students who partook of the Character Education program in all of the state's high schools (Character Education, 2015). Six schools were randomly selected from a public domain dataset which had a highly significant omnibus Chi Square (χ^2 =39.754, df = 5, P(χ^2> 39.754) = 0.0000) (Table 2). The Marascuilo Procedure was then applied as a post hoc test to identify the school-school pair(s) that was contributory to the overall statistical difference.

Table 2

Summary Of Proportion Of High School Students Who Completed Character Education At 6 Selected Texas High Schools (Character Education, 2015).

High School	Count of Students Who Participated	Total High School Population	Proportion
Flour Bluffs	314	476	0.660
George West High School	53	73	0.726
Edna High School	60	130	0.462
Karnes City High School	42	66	0.636
Cleveland High School	95	198	0.480
Liberty High School	84	168	0.500

http://tea.texas.gov/Academics/Learning_Support_and_Programs/Character_Education/Character_Education/

The routine found that, within the 15 school-school pairings [(6-5)/2], six were significantly different. Because of the inherent size limitation to the standard Marascuilo Procedure, however, this routine is unsuitable for large data analysis.

The several proportions, absolute differences, critical range and significance determination are reported in Table 3. The practical importance of this finding is not at issue and this is presented solely for illustrative purposes. Nevertheless, the outcome is interesting and possibly worth further investigation.

Table 3

Results Of The Marascuilo Procedure Demonstrating Between Group Proportional Significance For15 Combinations Of 6 High Schools That Participated In Texas Character Education Program, 2015.

District[i]	District[j]	p_i	p_j	$\lvert p_i - p_j \rvert$	Critical Value	Decision
Flour Bluff	George West	0.660	0.726	0.066	0.188	NS
Flour Bluff	Edna High	0.660	0.462	0.198	0.162	Significant
Flour Bluff	Karnes City	0.660	0.636	0.023	0.210	NS
Flour Bluff	Cleveland High	0.660	0.480	0.180	0.138	Significant
Flour Bluff	Liberty High	0.660	0.500	0.160	0.147	Significant

George West	Edna High	0.726	0.462	0.264	0.227	Significant
George West	Karnes City	0.726	0.636	0.090	0.263	NS
George West	Cleveland High	0.726	0.480	0.246	0.210	Significant
George West	Liberty High	0.726	0.500	0.226	0.216	Significant
Edna High	Karnes City	0.462	0.636	0.175	0.245	NS
Edna High	Cleveland High	0.462	0.480	0.018	0.187	NS
Edna High	Liberty High	0.462	0.500	0.038	0.194	NS
Karnes City	Cleveland High	0.636	0.480	0.157	0.230	NS
Karnes City	Liberty High	0.636	0.500	0.136	0.235	NS
Cleveland High	Liberty High	0.480	0.500	0.02	0.174	NS

Note: For $\chi^2_{0.05,5}$ 'significant' equates to at least p=0.0009.

The Modified Marascuilo Procedure as a Data Mining Tool

A modified Marascuilo Procedure was developed which eliminates the row limitation by substituting the standard normal score of 1.96 in place of the square root of the Chi Square statistic to calculate the margin of error. In the modified version, the Z-score remains fixed, regardless *k*. The limitation for the researcher is not formulaic but computational. Because the unit of analysis is not the Chi Square matrix as a whole but multiple, single row-by-row comparisons, there will only be one degree of freedom, thus,

$$Z = \sqrt{X^2_{0.05,1}} = 1.96.$$

The modified formula becomes:

$$r_{ij} = 1.96\sqrt{\frac{p_i(1-p_i)}{n_i} + \frac{p_j(1-p_j)}{n_j}}.$$

There will be *k(k-1)/2* pairwise comparisons from a rectangular data matrix where k is the number of rows. Each row requires a nominal identifier and sufficient information to generate a proportion.

The R for the Modified Marascuilo Procedure works in conjunction with any standard software that can export data in comma separated value format (CSV). R can read a wide variety of database exports, including Excel, SPSS, SAS and others. This modified code, while not yet in any known repository, is shown in Appendix B.

DATA MINING ILLLUSTRATION

In the American state of Arizona, the Department of Education conducts a standards based assessment (Arizona Instrument to Measure Science--AIMS) that measures student proficiency of the Arizona Academic Content Standard in Science (Assessment Results. (n.d.)) . This test is given to all students within the state who are in Grades 4, 8 and high school. The mean scale

science score is reported on the Department's publically accessible website along with four performance levels descriptors. These were the percent of students who exceeds, meets, approaches, or falls far below the standard. The first two descriptors were summed as the variable, "science percent passing".

Additional collected school data was the county, school name, whether charter school or not, school district name, school name and grade cohort. For Fall 2016, 4577 records were compiled. This data was aggregated and did not show individual student performance. Nevertheless combined data can still be analyzed if the unit of analysis is at a higher level than the individual.

An intergroup comparison for 15 counties yielded 105 pairwise comparisons [$k(k-1)/2$]. Data from the Arizona Department of Education was converted into CSV format and was read by R.

Of the 105 pairs, 90 were found to have no statistically significant between-county differences. Fifteen pairings, however, did show important differences and these are reported in Table 4.

Arizona consists of urban, suburban, and rural/tribal counties. The Modified Marascuilo Procedure, done for illustration, nevertheless exposed differences in AIMS performance that might warrant further examination. Of particular interest is the comparison between Maricopa (Phoenix), Pinal (Casa Grande/Florence) and Pima Counties (Tucson) (Table 4).

Table 4

Results of the Modified Marascuilo Procedure Showing the County Proportions, Absolute Differences, Critical Range And Test Determination.

Counties		p_i	p_j	$\lvert p_i - p_j \rvert$	Critical Range	Determination
Apache*	Cochise*	0.346	0.502	0.157	0.135	Significant
Apache*	Coconino*	0.346	0.550	0.204	0.146	Significant
Apache*	Maricopa	0.346	0.559	0.213	0.109	Significant
Apache*	Mohave*	0.346	0.538	0.193	0.139	Significant
Apache*	Pima	0.346	0.513	0.167	0.114	Significant
Apache*	Yavapai*	0.346	0.593	0.247	0.128	Significant
Coconino*	Gila*	0.550	0.384	0.166	0.163	Significant
Gila*	Maricopa	0.384	0.559	0.175	0.131	Significant
Gila*	Yavapai*	0.384	0.593	0.209	0.147	Significant
Maricopa	Navajo*	0.559	0.429	0.130	0.100	Significant
Maricopa	Pima	0.559	0.513	0.046	0.042	Significant
Maricopa	Pinal	0.559	0.443	0.115	0.068	Significant
Maricopa	Yuma*	0.559	0.447	0.111	0.085	Significant
Navajo*	Yavapai*	0.429	0.593	0.164	0.120	Significant
Pima	Yavapai*	0.513	0.593	0.080	0.080	Significant

Note: Only significant results for the 2015 AIMS (4^{th} grade "Passing Percentage") are shown.

* Rural/tribal county.
http://www.azed.gov/research-evaluation/aims-assessment-results/aims-science-2016.xlsx

Appendix A

USING R

The following steps are necessary to run this procedure in R and some prior experience is helpful:

1. Obtain and install the RGui software (https://CRAN.R-project.org/) and/or the R-Studio program (https://www.rstudio.com/). Note that the basic R program needs to be installed before RStudio. Extensive help is available at the Comprehensive R Archive Network (https://CRAN.R-project.org/).
2. If the data matrix is small, it is easy to manually enter. But the read.csv(), or other R read functions, should be used for large files, both for ease and accuracy.
3. Modify as needed but there must be the same number of observations, totals and counties. Use Notepad® for code writing.
4. Block, copy and paste code at the R command prompt (“>”) in R-Studio or RGui.
5. Code will run upon entry.
6. All researchers write code differently and conciseness is a function of experience. Apologies are pre-given for areas where R could have been a little tighter.

Presently the Marascuilo Procedure is not archived in CRAN.

Appendix B

The R Code for The Marascuilo Procedure (*Post Hoc* Analysis Using Chi Square Degrees of Freedom For Critical Range Determination)

```
## Set the proportions of interest. Enter observed and totals. Because the number of
compared entities will be six or fewer, it is just as easy to manually enter the values. For
larger datasets, then reading a CSV file makes sense.
## this is the illustration for the High School comparison in Table 2

obs = c(314,53,60,42,95,84)
total = c(476,73,130,66,198,168)
p = obs/total
## Enter names of the high schools
names=c("Flour Bluff","George West", "Edna High", "Karnes City", "Cleveland High",
"Liberty High")
N = length(p)
absdiff = critical.range = c()
d = c(NULL,NULL)
e = c(NULL,NULL)
f=c()

## Compute critical values. The Chi Square critical value is calculated for the
appropriate degrees of freedom and α/2=0.025
for (i in 1:(N-1))
  { for (j in (i+1):N)
        {d <- rbind(d,c(names[i], names[j]))
         e <- rbind(e,c(p[i], p[j]))

         { absdiff = c(absdiff,(abs(p[i]-p[j])))
         critical.range = c(critical.range,
         sqrt(qchisq(.95,(N-1)))*sqrt(p[i]*(1-p[i])/total[i] + p[j]*(1-p[j])/total[j]))

   }
   }
   }

f<-cbind(ifelse (absdiff>critical.range,"Significant","NS"))
print(cbind(d,round(cbind(e,absdiff,critical.range),3),f),quote=FALSE)
```

Appendix C

The R Code for the Modified Marascuilo Procedure (Data Exploration Using Standard Z=1.96 For Critical Range Determination)

```
## Set the proportions of interest. For larger datasets using CSV files is easier than manual entry.
## Variables: Mean Countywide Score for Passing Science AIMS ("SciPass"), total schools within county ("Schools"), "County"
df <- read.csv("C:/users/…/filename.csv")
p <- c(as.numeric(df$SciPass))
county <- as.character(df$County)
##
total <- df$Schools
N = length(p)
absdiff = critical.range = c()
d = c(NULL,NULL)
e = c(NULL,NULL)
f=c()

## Compute critical values.
for (i in 1:(N-1))
  { for (j in (i+1):N)
        {d <- rbind(d,c(county[i], county[j]))
         e <- rbind(e,c(p[i], p[j]))
                { absdiff = c(absdiff,(abs(p[i]-p[j])))
                critical.range = c(critical.range,
                1.96*(sqrt(p[i]*(1-p[i])/total[i] + p[j]*(1-p[j])/total[j])))

  }
  }
  }

f<-cbind(ifelse (absdiff > critical.range,"Significant","NS"))
print(cbind(d,round(cbind(e,absdiff,critical.range),3),f),quote=FALSE)
```

References

Assessment Results. (n.d.). Retrieved April 05, 2017, from http://www.azed.gov/research-evaluation/aims-assessment-results/aims-science-2016.xlsx

Character Education. (n.d.). Retrieved February 28, 2017, from http://tea.texas.gov/Academics/Learning_Support_and_Programs/Character_Education/Character_Education/

The Comprehensive R Archive Network. (n.d.). Retrieved February 28, 2017, from https://CRAN.R-project.org/

Marascuilo, L. A. (1966). Large-sample Multiple Comparisons. *Psychological Bulletin,65*(5), 280-290. Retrieved March 20, 2017.

Marascuilo, L. A., & Serlin, R. C. (1998). *Statistical methods for the social and behavioral sciences*. New York: W. H. Freeman.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

The R Project for Statistical Computing. (n.d.). Retrieved February 28, 2017, from https://www.r-project.org/

RStudio. (n.d.). Retrieved February 28, 2017, from https://www.rstudio.com/products/rstudio/#Desktop