Open Access Library Journal
2017, Volume 4, e3927
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Working with Aggregate Data: An Excel Macro for Pairwise Comparison Using *Z* Test for Two Proportions

Victor L. Landry

College of Doctoral Studies, Grand Canyon University, Phoenix, AZ, USA
Email: victor.landry@my.gcu.edu

How to cite this paper: Landry, V.L. (2017) Working with Aggregate Data: An Excel Macro for Pairwise Comparison Using *Z* Test for Two Proportions. *Open Access Library Journal*, **4**: e3927.
https://doi.org/10.4236/oalib.1103927

Received: September 8, 2017
Accepted: October 16, 2017
Published: October 19, 2017

Abstract

A Visual Basic for Applications (VBA) Excel macro was created for doing a pairwise, two-sample *Z*-test of within-column proportions for k data rows in an Excel spreadsheet. By program iteration, the *Z*-score for $k(k-1)/2$ unique, non-repeating and non-duplicated within-column comparisons was generated and the null hypothesis is tested against a two-tailed *Z*-score critical value. This within-column process is useful for extracting potential meaning from large aggregate columnar data. The procedure was demonstrated using aggregate internet acquired summary data in the public domain. The VBA macro is provided and it is also available at the author's website.

Subject Areas

Education, Mathematical Statistics, Politics, Sociology, Statistics

Keywords

Pairwise Comparison, Two-Proportion *Z*-Test, Aggregated Data, Internet Data, Excel VBA Macro

1. Introduction

Aggregate data refers to numerical or non-numerical information that is: 1) collected from multiple sources and/or on multiple measures, variables, or individuals; and 2) compiled into data summaries or summary reports, typically for the purposes of public reporting or statistical analysis—*i.e.*, examining trends, making comparisons, or revealing information and insights that would not be observable when data elements are viewed in isolation [1]. Because the unit of analysis in aggregated data is no longer at the individual entity level, researchers

must exercise care in trying to conduct correlational or inferential statistics to avoid spurious results.

Aggregate data might still yield important information by moving to the next higher unit of analysis that provides a grouping unifier. Various versions of Chi Square, time series and proportional analyses may still be performed on aggregate datasets where a proper unifier exists.

Proportional aggregate analysis is the focus of this paper. A method and an Excel VBA macro is demonstrated that compares and contrasts a spreadsheet (above row to row) for unique and non-repeating pairwise row comparisons. This procedure incorporates the familiar Z-Test for Two Proportions to test paired data for statistical significance at $\alpha \leq 0.05$ [2].

2. Fundamental Principles

The VBA macro uses an "up one row", "down one row" iteration that populates the variables for p_i and p_j. The built-in Z-test for proportions has a two-tailed null hypothesis of no statistically significant difference between two proportions, H_0: $P_i = P_j$. The alternate hypothesis is Ha: $P_i \neq P_j$. There are three assumptions inherent in this procedure: 1) sampling independence; 2) sufficient size (≥5; the macro will reject if violated); and 3) randomness of selection. A pooled proportion is used to compute the standard error of the sampling distribution, using the individual proportions, p_i and p_j and the associated population for each, n_i and n_j. The test statistic is a Z-score which is the ratio of the absolute proportion difference divided by the standard error. Significance is determined as $Z \geq 1.96$, the two-tailed critical value for a normal distribution.

3. An Illustration

For illustration purposes, a mock research question was created that asked if there were any statistically significant between-county differences in the proportion of registered voters for the Green Party within the state of Arizonain January 2017 [3]. After minor cleansing, the data were inserted into a blank Excel macro-enabled (pairwise.xlsm) spreadsheet which incorporates the pairwise macro described in this report. The order of insertion must be followed exactly (Group Name, Sample Size and Total) starting in cell "A1" which is required by the macro (**Figure 1**).

The goal for this mock research question was to determine if there were any statistically significant proportional differences of Green Party registered voters between compared counties. For example, is the proportion of Green Party registered voters in Apache County significantly different from the proportions of Green Party registered voters in other counties? How many matched pairings of county-county data would be significantly different? This information could be pursued to investigate trends and patterns.

Because of the requirement of the Z-test for proportional differences, the minimum number of registered voters per county was 5. Only one county, Greenlee,

	A	B	C
1	County	Green Voters	Total
2	Apache	42	47648
3	Cochise	129	72673
4	Coconino	231	72709
5	Gila	26	28293
6	Graham	9	17060
7	Greenlee	0	4540
8	La Paz	10	8546
9	Maricopa	2817	2056458
10	Mohave	113	109616
11	Navajo	51	61751
12	Pima	1562	509310
13	Pinal	199	178474
14	Santa Cruz	36	25178
15	Yavapai	244	130335
16	Yuma	43	78020

Figure 1. The correct order of data insertion starting at Cell "A1".
Note: As of January 2017 per
https://www.azsos.gov/elections/voter-registration-historical-election-data/voter-registration-counts

failed to meet the minimum sample size and all of its combinations were eliminated.

4. Results

Output begins in cell "F1" and continues for $k(k-1)/2$ rows. For the fifteen rows illustrated, an output of 105 rows is generated (Table 1). The output grows exponentially and while the macro can accommodate very large datasets, there is a practical output limitation. For example, 50 rows of input would create 1225 matched pairs of unique data. The size of the input range is the researcher's choice.

This exercise was primarily for illustration but it did use real data which produced real results. Of the 105 county-county combinations, 59 (56%) showed statistically significant differences. Questions need to be asked of the data so that the differences in Green Party registered voters could perhaps be explained. For those in the social or political sciences, these differences might be important to pursue.

5. The Macro Methodology

The VBA macro uses an "up one row", "down one row" iteration that populates the upper row/lower row variables with their respective proportions, P_1 and P_2. With these values, the null hypothesis ($P_1 = P_2$) can be tested using the following standard proportion equations.

1) The pooled proportion:

$$p = \frac{p_i * n_i + p_j * n_j}{n_i + n_j}$$

Table 1. Results of pairwise comparison of between-county Z-test of proportions for green party registered voters.

Compared Groups	Group 1 p_1	N_1	P_1	Group 2	N_2	P_2	Z-Score	Result
Apache-Cochise	42	47,648	0.0009	129	72,673	0.0018	4.0241	Sig.
Apache-Coconino	42	47,648	0.0009	231	72,709	0.0032	8.1869	Sig.
Apache-Gila	42	47,648	0.0009	26	28,293	0.0009	0.167	NS
Apache-Graham	42	47,648	0.0009	9	17,060	0.0005	1.4135	NS
Apache-Greenlee	42	47,648	0.0009	0	4540	0	2.0013	N ≤ 5
Apache-La Paz	42	47,648	0.0009	10	8546	0.0012	0.8082	NS
Apache-Maricopa	42	47,648	0.0009	2817	2,056,458	0.0014	2.861	Sig.
Apache-Mohave	42	47,648	0.0009	113	109,616	0.001	0.8677	NS
Apache-Navajo	42	47,648	0.0009	51	61,751	0.0008	0.3127	NS
Apache-Pima	42	47,648	0.0009	1562	509,310	0.0031	8.5128	Sig.
Apache-Pinal	42	47,648	0.0009	199	178,474	0.0011	1.388	NS
Apache-Santa Cruz	42	47,648	0.0009	36	25,178	0.0014	2.1517	Sig.
Apache-Yavapai	42	47,648	0.0009	244	130,335	0.0019	4.6199	Sig.
Apache-Yuma	42	47,648	0.0009	43	78,020	0.0006	2.1853	Sig.
Cochise-Coconino	129	72,673	0.0018	231	72,709	0.0032	5.3778	Sig.
Cochise-Gila	129	72,673	0.0018	26	28,293	0.0009	3.1205	Sig.
Cochise-Graham	129	72,673	0.0018	9	17,060	0.0005	3.7421	Sig.
Cochise-Greenlee	129	72,673	0.0018	0	4540	0	2.8412	N ≤ 5
Cochise-La Paz	129	72,673	0.0018	10	8546	0.0012	1.2798	NS
Cochise-Maricopa	129	72,673	0.0018	2817	2,056,458	0.0014	2.8883	Sig.
Cochise-Mohave	129	72,673	0.0018	113	109,616	0.001	4.2726	Sig.
Cochise-Navajo	129	72,673	0.0018	51	61,751	0.0008	4.7425	Sig.
Cochise-Pima	129	72,673	0.0018	1562	509,310	0.0031	6.0526	Sig.
Cochise-Pinal	129	72,673	0.0018	199	178,474	0.0011	4.1534	Sig.
Cochise-Santa Cruz	129	72,673	0.0018	36	25,178	0.0014	1.1507	NS
Cochise-Yavapai	129	72,673	0.0018	244	130,335	0.0019	0.4894	NS
Cochise-Yuma	129	72,673	0.0018	43	78,020	0.0006	7.0312	Sig.
Coconino-Gila	231	72,709	0.0032	26	28,293	0.0009	6.3968	Sig.
Coconino-Graham	231	72,709	0.0032	9	17,060	0.0005	6.0315	Sig.
Coconino-Greenlee	231	72,709	0.0032	0	4540	0	3.8036	N ≤ 5
Coconino-La Paz	231	72,709	0.0032	10	8546	0.0012	3.2273	Sig.
Coconino-Maricopa	231	72,709	0.0032	2817	2,056,458	0.0014	12.6668	Sig.
Coconino-Mohave	231	72,709	0.0032	113	109,616	0.001	10.3402	Sig.

Continued

Coconino-Navajo	231	72,709	0.0032	51	61,751	0.0008	9.3913	Sig.
Coconino-Pima	231	72,709	0.0032	1562	509,310	0.0031	0.5014	NS
Coconino-Pinal	231	72,709	0.0032	199	178,474	0.0011	11.3375	Sig.
Coconino-Santa Cruz	231	72,709	0.0032	36	25,178	0.0014	4.5813	Sig.
Coconino-Yavapai	231	72,709	0.0032	244	130,335	0.0019	5.8355	Sig.
Coconino-Yuma	231	72,709	0.0032	43	78,020	0.0006	11.959	Sig.
Gila-Graham	26	28,293	0.0009	9	17,060	0.0005	1.4541	NS
Gila-Greenlee	26	28,293	0.0009	0	4540	0	2.0434	N ≤ 5
Gila-La Paz	26	28,293	0.0009	10	8546	0.0012	0.6513	NS
Gila-Maricopa	26	28,293	0.0009	2817	2,056,458	0.0014	2.0411	Sig.
Gila-Mohave	26	28,293	0.0009	113	109,616	0.001	0.5289	NS
Gila-Navajo	26	28,293	0.0009	51	61,751	0.0008	0.4435	NS
Gila-Pima	26	28,293	0.0009	1562	509,310	0.0031	6.4799	Sig.
Gila-Pinal	26	28,293	0.0009	199	178,474	0.0011	0.9293	NS
Gila-Santa Cruz	26	28,293	0.0009	36	25,178	0.0014	1.7327	NS
Gila-Yavapai	26	28,293	0.0009	244	130,335	0.0019	3.5255	Sig.
Gila-Yuma	26	28,293	0.0009	43	78,020	0.0006	2.0811	Sig.
Graham-Greenlee	9	17,060	0.0005	0	4540	0	1.5479	N ≤ 5
Graham-La Paz	9	17,060	0.0005	10	8546	0.0012	1.7807	NS
Graham-Maricopa	9	17,060	0.0005	2817	2,056,458	0.0014	2.9697	Sig.
Graham-Mohave	9	17,060	0.0005	113	109,616	0.001	1.9715	Sig.
Graham-Navajo	9	17,060	0.0005	51	61,751	0.0008	1.2506	NS
Graham-Pima	9	17,060	0.0005	1562	509,310	0.0031	5.9809	Sig.
Graham-Pinal	9	17,060	0.0005	199	178,474	0.0011	2.2488	Sig.
Graham-Santa Cruz	9	17,060	0.0005	36	25,178	0.0014	2.7891	Sig.
Graham-Yavapai	9	17,060	0.0005	244	130,335	0.0019	3.9894	Sig.
Graham-Yuma	9	17,060	0.0005	43	78,020	0.0006	0.1194	NS
Greenlee-La Paz	0	4540	0	10	8546	0.0012	2.3058	N ≤ 5
Greenlee-Maricopa	0	4540	0	2817	2,056,458	0.0014	2.4955	N ≤ 5
Greenlee-Mohave	0	4540	0	113	109,616	0.001	2.1644	N ≤ 5
Greenlee-Navajo	0	4540	0	51	61,751	0.0008	1.9371	N ≤ 5
Greenlee-Pima	0	4540	0	1562	509,310	0.0031	3.7371	N ≤ 5
Greenlee-Pinal	0	4540	0	199	178,474	0.0011	2.2511	N ≤ 5
Greenlee-Santa Cruz	0	4540	0	36	25,178	0.0014	2.5494	N ≤ 5
Greenlee-Yavapai	0	4540	0	244	130,335	0.0019	2.918	N ≤ 5

Continued

Greenlee-Yuma	0	4540	0	43	78,020	0.0006	1.5822	N ≤ 5
La Paz-Maricopa	10	8546	0.0012	2817	2,056,458	0.0014	0.4982	NS
La Paz-Mohave	10	8546	0.0012	113	109,616	0.001	0.3845	NS
La Paz-Navajo	10	8546	0.0012	51	61,751	0.0008	1.013	NS
La Paz-Pima	10	8546	0.0012	1562	509,310	0.0031	3.161	Sig.
La Paz-Pinal	10	8546	0.0012	199	178,474	0.0011	0.149	NS
La Paz-Santa Cruz	10	8546	0.0012	36	25,178	0.0014	0.562	NS
La Paz-Yavapai	10	8546	0.0012	244	130,335	0.0019	1.4713	NS
La Paz-Yuma	10	8546	0.0012	43	78,020	0.0006	2.1962	Sig.
Maricopa-Mohave	2817	2,056,458	0.0014	113	109,616	0.001	2.9751	Sig.
Maricopa-Navajo	2817	2,056,458	0.0014	51	61,751	0.0008	3.6219	Sig.
Maricopa-Pima	2817	2,056,458	0.0014	1562	509,310	0.0031	26.2683	Sig.
Maricopa-Pinal	2817	2,056,458	0.0014	199	178,474	0.0011	2.813	Sig.
Maricopa-Santa Cruz	2817	2,056,458	0.0014	36	25,178	0.0014	0.2557	NS
Maricopa-Yavapai	2817	2,056,458	0.0014	244	130,335	0.0019	4.7033	Sig.
Maricopa-Yuma	2817	2,056,458	0.0014	43	78,020	0.0006	6.1361	Sig.
Mohave-Navajo	113	109,616	0.001	51	61,751	0.0008	1.3175	NS
Mohave-Pima	113	109,616	0.001	1562	509,310	0.0031	11.7704	Sig.
Mohave-Pinal	113	109,616	0.001	199	178,474	0.0011	0.6666	NS
Mohave-Santa Cruz	113	109,616	0.001	36	25,178	0.0014	1.718	NS
Mohave-Yavapai	113	109,616	0.001	244	130,335	0.0019	5.3256	Sig.
Mohave-Yuma	113	109,616	0.001	43	78,020	0.0006	3.5535	Sig.
Navajo-Pima	51	61,751	0.0008	1562	509,310	0.0031	9.9095	Sig.
Navajo-Pinal	51	61,751	0.0008	199	178,474	0.0011	1.9206	NS
Navajo-Santa Cruz	51	61,751	0.0008	36	25,178	0.0014	2.5543	Sig.
Navajo-Yavapai	51	61,751	0.0008	244	130,335	0.0019	5.4688	Sig.
Navajo-Yuma	51	61,751	0.0008	43	78,020	0.0006	1.9677	Sig.
Pima-Pinal	1562	509,310	0.0031	199	178,474	0.0011	14.0414	Sig.
Pima-Santa Cruz	1562	509,310	0.0031	36	25,178	0.0014	4.6444	Sig.
Pima-Yavapai	1562	509,310	0.0031	244	130,335	0.0019	7.2539	Sig.
Pima-Yuma	1562	509,310	0.0031	43	78,020	0.0006	12.5348	Sig.
Pinal-Santa Cruz	199	178,474	0.0011	36	25,178	0.0014	1.3774	NS
Pinal-Yavapai	199	178,474	0.0011	244	130,335	0.0019	5.49	Sig.
Pinal-Yuma	199	178,474	0.0011	43	78,020	0.0006	4.2792	Sig.
Santa Cruz-Yavapai	36	25,178	0.0014	244	130,335	0.0019	1.5155	NS
Santa Cruz-Yuma	36	25,178	0.0014	43	78,020	0.0006	4.3833	Sig.
Yavapai-Yuma	244	130,335	0.0019	43	78,020	0.0006	7.8683	Sig.

Note: All comparisons using Greenlee county were rejected because of small sample size (less than or equal to 5).

where:

p = the pooled sample proportion,

p_i = first proportion,

p_j = second proportion,

n_i = population size associated with the first proportion,

n_j = population size associated with the second proportion.

2) The standard error of the weighted samples:

$$se_{pi-pj} = \sqrt{p(1-p)*\left[\left(\frac{1}{n_i}\right)+\left(\frac{1}{n_j}\right)\right]}$$

where:

$se_{pi\text{-}pj}$ = the standard error,

p = the weighted estimate of two populations,

n_i = sample size associated with the first proportion,

n_j = sample size associated with the second proportion.

3) The determination of the Z-score:

$$Z = \frac{\left|p_i - p_j\right|}{se}$$

where:

Z = the Z-score,

p_i = first proportion,

p_j = second proportion,

$se_{pi\text{-}pj}$ = the standard error.

The null hypothesis is rejected if the Z-score exceeds 1.96, the two-tailed critical value that is associated with a p-value ≤ 0.05.

6. Conclusions

An Excel macro procedure has been demonstrated as a screening tool to reveal patterns within aggregate data. It creates unique within-column pairwise comparisons and tests the data for proportional statistical significance. This method could be applied where aggregated data is available that includes, as a minimum, the named group, a proportion or count of a desired variable and a total for each row. The exponential growth of the output as the number of rows (k) increases will be a practical limiting factor.

The Excel macro can be saved as an Excel macro file (*.xlsm) and various internet references can be accessed for instructions for using an Excel macro files as an add-in.

This macro is also available for download at

http://www.viclandry.com/pairwise-comparison.html

The VBA Macro

```
Sub Pairwise()
Dim i As Integer
Dim j As Integer
```

```
Dim k As Integer
Dim lastrow As Long
Dim answer As Variant
Dim n1 As Variant
Dim n2 As Variant
Dim p As Variant
Dim p1 As Variant
Dim p2 As Variant
Dim z As Variant
Dim se As Variant
Dim r As Variant
MsgBox ("You must have HEADERS with category names in Column A; place data in Column B; place interval COUNTS in Column C")
lastrow = (Cells(Rows.Count, "A").End(xlUp).Row)-1
Range("f1").Value = "Compared Groups"
Range("f1").Offset(0, 1) = "Group 1"
Range("f1").Offset(0, 2).Value = "N1"
Range("f1").Offset(0, 3).Value = "P1"
Range("f1").Offset(0, 4).Value = "Group 2"
Range("f1").Offset(0, 5).Value = "N2"
Range("f1").Offset(0, 6).Value = "P2"
Range("f1").Offset(0, 7).Value = "Z-Score"
Range("f1").Offset(0, 8).Value = "Result"
For i = 1 To lastrow
For j = i + 1 To lastrow
k = k + 1
Range("f1").Offset(k, 0).Value = (Range("a1").Offset(i, 0).Value & " - " & Range("a1").Offset(j, 0).Value) 'first row header
p1 = Range("a1").Offset(i, 1).Value/Range("a1").Offset(i, 2).Value 'value for first proportion
p2 = Range("a1").Offset(j, 1).Value/Range("a1").Offset(j, 2).Value 'value for second proportion
r = (Abs(p1 - p2)) 'find absolute difference
n1 = Range("a1").Offset(i, 2).Value
n2 = Range("a1").Offset(j, 2).Value
p = ((p1 * n1) + (p2 * n2))/(n1 + n2)
se = Sqr((p * (1 - p)) * ((1/n1) + (1/n2)))
z = r/se
Range("f1").Offset(k, 1).Value = Range("a1").Offset(i, 1).Value 'first count
Range("f1").Offset(k, 2).Value = n1 'first total
Range("f1").Offset(k, 3).Value = Round(p1, 4) 'first proportion
Range("f1").Offset(k, 4).Value = Range("a1").Offset(j, 1).Value 'second count
Range("f1").Offset(k, 5).Value = n2 'second total
```

```
Range("f1").Offset(k, 6).Value = Round(p2, 4) 'second proportion
Range("f1").Offset(k, 7).Value = Round(z, 4) 'z score
If z > 1.96 Then
Range("f1").Offset(k, 8).Value = "Sig."
Else
Range("f1").Offset(k, 8).Value = "NS"
End If
If n1 * p1 < 6 Or n2 * p2 < 6 Then
Range("f1").Offset(k, 8).Value = "N<=5"
End If
Next j
Next i
Range("f1:m1").EntireColumn.AutoFit
End Sub
```

References

[1] Concepts, L. (2015) Aggregate Data Definition. http://edglossary.org/aggregate-data/

[2] StatisticsLectures.com (2017) Z-Test for Proportions, Two Samples. http://www.statisticslectures.com/topics/ztestproportions/

[3] Arizona Secretary of State (2017) Voter Registration Counts. https://www.azsos.gov/elections/voter-registration-historical-election-data/voter-registration-counts