

# A Note on Data and Methods for the Afghanistan Casualty Study

by Michael Porter\*

September 2011

## Data

The analysis is based on US casualties from Operation Enduring Freedom from its beginning in October 2001 to December 31, 2010. Casualties after this date are not included in the analysis. Data about the casualties came from the icasualties.org website (<http://icasualties.org/OEF/Fatalities.aspx>) which includes the casualty's name, rank, age, the cause of death, where the death occurred, and home state and city. I matched each casualties home city and state to places listed in the casualty data with the United States Geological Survey's (USGS) list of populated places ([http://geonames.usgs.gov/domestic/download\\_data.htm](http://geonames.usgs.gov/domestic/download_data.htm)). Using this method I identified counties for the 1,391 of the 1,446 casualties. All of the demographic, social, and economic data came from the 2005-2009 American Community Survey 5 year estimate. Details about data sources are available in Appendix C. Vote tallies for the 2008 election came from the New York Times vizlab (<http://vizlab.nytimes.com/datasets/2008-presidential-election-results-b-5/versions/1.txt>).

## Methods

In this project I test whether the characteristics of communities that were once home to US casualties in the Afghan war (hereafter casualties) differ from the

---

\*Doctoral candidate in the Department of Earth and Environmental Sciences at the Graduate Center of the City University of New York. His research interests include spatial data analysis, geographic information systems, urban environmental policy, and environmental justice.

characteristics of communities for the population age 18 to 65. Statistically, this is an issue of comparing two distributions for each variable of interest. In the first distribution there will be one value for each casualty. The first distribution, for instance, could include the median household income for each of the 1,391 casualties' home counties. In the second distribution there will be one value for each person in the United States. Using the above example, the second distribution would include the median household income for each of the home counties of for all of the roughly 300 million people in the United States. If these two distributions are statistically different we can confidently assert that the median household income for casualty counties is higher or lower than the country in general.

The most commonly used method for comparing distributions is the t-test. The t-test, however, is a parametric test. In general parametric methods make assumptions about the underlying data. One assumption of the test is that data are normally distributed (under a normal distribution the distribution's median equals the distribution's mean). In the analysis, however, the distribution of almost all variables of interest are not normal or skewed (the median and the mean are not equal). Because the data do not meet the underlying assumptions parametric methods, like a t-test, will produce spurious and potentially misleading results. As an alternative I use a non-parametric method to compare distributions – the Mann Whitney U (see below for details). Unlike t-tests and other parametric methods, non-parametric methods make no assumptions about the underlying data and are therefore more likely to produce accurate results.

One challenge in using the Mann Whitney U test is that for large data sets analysis can be computationally intensive. This application, however, presents a unique workaround to the problem. Because distributions are comprised of county level data and there are only 3,221 counties in the United States, there will necessarily be many repeat values. I can exploit this redundancy by modifying the Mann Whitney U to include a weight matrix where each the weight is equal to each county's population between the ages of 18 and 65.

As part of the analysis we classify each county into one of five different groups: core city, suburban, small city, town or village, or rural. Each county's classification is based on core based statistical area (CBSA) data compiled by the US Office of Management and Budget (OMB). In general, core counties are central counties in CBSAs with a population greater than 1,000,000, suburban counties are non-core counties in CBSAs with a population less than 1,000,000, small city counties are counties in CBSAs with an urbanized population less than 1,000,000 and greater than 50,000, towns and villages are micropolitan areas, and rural counties are counties that are not in a CBSA. Details about the classification

are in Appendix D.

To create the maps in the text, I mapped both the ratio of casualties to the total population by county, and then the ratio of casualties to total area by county. To help identify trends, I converted the vector data into a raster and then use a moving window smoother to remove noise in the data.

## Modified Mann Whitney U

The Mann Whitney U is a non-parametric statistic used to compare two distributions. Unlike the student t test which compares actual values of two distributions, the Mann Whitney U does the comparison based on ranks.(Mann and Whitney 1947, Nachar 2008). This makes the Mann Whitney U test less susceptible to outliers. The use of ranks also relaxes the assumption that the two samples have a normal distribution. The Mann Whitney U initially breaks down into three calculations (eq 1, 2, and 3):

$$U_x = nxny + \frac{nx(nx+1)}{2} - R_x \quad (1)$$

$$U_y = nxny + \frac{ny(ny+1)}{2} - R_y \quad (2)$$

$$U = \min(U_x, U_y) \quad (3)$$

where:

- $nx$  = the number of observations in group  $x$
- $ny$  = the number of observations in group  $y$
- $R_x$  = The sum of ranks of the observations in group  $x$
- $R_y$  = The sum of ranks of the observations in group  $y$

The Mann Whitney U, of course, assumes that all of the observations have equal weights. In this example, however, this assumption does not hold, as each county will have a weight equal to its population. To accommodate weights we would have to re-calculate  $R_x$  and  $R_y$ . The first step in doing this is to calculate the weighted rank for each observation.

$$r_i = \frac{1}{2} \left( \left( \sum_{j=1}^g h(i,j) * n_j \right) + n_i \right) \quad (4)$$

Where:

- $r_i$  weighted rank for observation  $i$
- $g$  = the total number of observations, in this case 6
- $h(i, j) = 1$  if value for observation  $j$  is greater than value for observation  $i$ , otherwise 0
- $n_j =$  weight for observation  $j$

Using the weighted rank and the group size we can now calculate  $R_x + R_y$  as follows:

$$R_x = \sum_{i=1}^{g_x} n_i * r_i \quad (5)$$

and

$$R_y = \sum_{i=1}^{g_y} n_i * r_i \quad (6)$$

Where:

- $g_x =$  the number of observations of type  $x$
- $g_y =$  the number of observations of type  $y$
- $n_i =$  the weight for observation  $i$
- $r_i =$  the weighted rank of observation  $i$

To calculate  $U_x$  and  $U_y$  we now have to recalculate  $nx$  and  $ny$  as follows:

$$nx = \sum_{i=1}^g xn_i \quad (7)$$

and

$$ny = \sum_{i=1}^g yn_i \quad (8)$$

Using the  $nx$ ,  $ny$ ,  $R_x$ , and  $R_y$  as calculated in equations 5 - 8 we can recalculate the  $U$  using equations 1 - 3.

### ***Significance Testing***

For large samples,  $U$  will follow a normal distribution, in that case the test statistic  $z$  (equation 7) is normally distributed. Significance values for  $z$  can be checked in tables of the normal distribution (appendix A).

$$z = |(U - \mu) / \sigma| \quad (9)$$

Where the mean and standard deviation are shown in equations 8 and 9. It should be noted that equation 9 is the estimate when there are a large number of ties.

$$\mu = \frac{(nxtp)}{2} - \frac{(U_x + U_y)}{2} \quad (10)$$

$$\sigma = \sqrt{\frac{nxtp}{N(N-1)} \left( \frac{N^3 - N}{12} - \sum_{j=1}^g \frac{t_j^3 - t_j}{12} \right)} \quad (11)$$

Where  $g$  is the number of ties,  $t_j$  is the number of tied values in the second group, and  $N$  is the total number of observations ( $nx + tp$ ).

### ***Example***

To illustrate the weighted Mann Whitney U test, I start with the following example from Nachar (2008:17):

An experimenter read that there is an antibiotic often tested, well documented, and known to help information storage in memory. This experimenter also knows through scientific reports and guidelines that the behavioral therapy has an established efficacy for the treatment of the social phobia (APA, 1998; INSERM, 2004; BPSCORE, 2001). In addition, he knows that the behavioral therapy requires the learning of new behaviours which implies information storage.

The number of symptoms of social phobia after two types of therapy was investigated. Two groups of individuals with social phobia were compared. The first group received the behavioral therapy; the second group received the behavioral therapy combined with the antibiotic. After each therapy, both groups showed a decreased in the number of symptoms of social phobia. The number of these symptoms was

measured and a test was run to decide whether the combined therapy had more effect on the symptoms than the behavioral therapy alone.

In other terms, the experimenter wishes to compare two random variables having continuous cumulative distribution functions. He wishes to test the hypothesis that his variables are stochastically equal (their distributions are similar) against the alternative that C is stochastically smaller than B. C corresponds to the numbers of symptoms under investigation in the combined therapy group, B corresponds to the numbers of symptoms in the behavioral therapy group.

Nachar then provides the following hypothetical results:

Numbers of symptoms	1	1	2	2	3	3	4	4	5	5	5	7	7	7
Behavioral therapy (b) / Combined therapy (c)	c	c	c	c	b	b	b	b	c	c	c	b	b	b
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Using these data we can calculate the rank sum for combined therapy ( $R_c$ ) =  $1 + 2 + 3 + 4 + 9 + 10 + 11 = 40$ , and the rank sum for behavioral therapy ( $R_b$ ) =  $5 + 6 + 7 + 8 + 12 + 13 + 14 = 65$ .

We can then plug these numbers into equations 1 – 3:

- $U_x = (7)(7) + ((7(7 + 1))/2) - 40 = 49 + (56/2) - 40 = 37$
- $U_y = (7)(7) + ((7(7 + 1))/2) - 40 = 49 + (56/2) - 65 = 12$
- $U = \min(37, 12) = 12$

We then consult a standard table to determine significance.

Now let's say instead of considering each sample independently (as Nachar does) we grouped them with each group having a weight equal to the group size. Now the data looks like this:

Numbers of symptoms	1	2	3	4	5	7
Behavioral therapy (b) / Combined therapy (c)	c	c	b	b	c	b
Weight	2	2	2	2	3	3

One advantage of doing this is that we have reduced the original vector from 14 elements to 6. Using equation 4 we can calculate the weighted rank for each group as

Numbers of symptoms	1	2	3	4	5	7
Behavioral therapy (b) / Combined therapy (c)	c	c	b	b	c	b
Weight	2	2	2	2	3	3
Weighted Rank	1.5	3.5	5.5	7.5	10	13

You will notice that the sum of the product of weight and weighted rank is equal to the sum of ranks in the first example and is also equal to  $R_x + R_y$  in Nachar's example.

$$(2 * 1.5) + (2 * 3.5) + (2 * 5.5) + (2 * 7.5) + (3 * 10) + (3 * 13) = 3 + 7 + 11 + 15 + 30 + 39 = 105$$

Using equations 5 and 6 we can calculate  $R_x$  and  $R_y$ :

- $R_x = (2 * 1.5) + (2 * 3.5) + (3 * 10) = 3 + 7 + 30 = 40$ , and
- $R_y = (2 * 5.5) + (2 * 7.5) + (3 * 13) = 11 + 15 + 39 = 65$

And using equation 7  $nx = 2 + 2 + 3 = 7$  and equation 8  $ny = 2 + 2 + 3 = 7$ . We can then plug these values back into equations 1 – 3 to calculate  $U_x$ ,  $U_y$ , and  $U$  and will yield the same result.

### ***Applying this method to the analysis of casualty data***

To apply the above method to the analysis of county data the first vector of values will include a value for each casualty's home county with a weight of 1 (table A1). The second vector will include a value for each county in the United State with a weight equal to that county's population (table A2).

Table A1: Percentage white population for casualty counties

Casualty	County	% White	Weight
1	Jackson County, Kentucky	98.40%	1
2	Grundy County, Iowa	98.30%	1
3	Elk County, Pennsylvania	98.30%	1
4	Monroe County, Ohio	98.10%	1
5	Clark County, Missouri	98.00%	1
6	Lewis County, Kentucky	98.00%	1
7	Knox County, Missouri	97.70%	1
8	Clay County, West Virginia	97.60%	1
9	Russell County, Virginia	97.60%	1
10	Kingsbury County, South Dakota	97.50%	1
	...		
1,387	Yauco Municipio, Puerto Rico	1.30%	1
1,388	San Juan Municipio, Puerto Rico	1.20%	1
1,389	Naranjito Municipio, Puerto Rico	0.90%	1
1,390	Cidra Municipio, Puerto Rico	0.80%	1
1,391	Cidra Municipio, Puerto Rico	0.80%	1

Table A2: Percentage white population for counties in general

County	Name	% White	Weight (Pop)
1	Hooker County, Nebraska	100.00%	661
2	Robertson County, Kentucky	100.00%	2,234
3	Tucker County, West Virginia	99.96%	6,861
4	Hand County, South Dakota	99.88%	3,268
5	Owsley County, Kentucky	99.78%	4,648
6	Garfield County, Montana	99.74%	1,135
7	Liberty County, Montana	99.71%	2,100
8	Nicholas County, West Virginia	99.53%	26,084
9	Wolfe County, Kentucky	99.52%	7,080
10	Switzerland County, Indiana	99.45%	9,627
	...		
3,217	Canvanas Municipio, Puerto Rico	0.20%	47,174
3,218	Loza Municipio, Puerto Rico	0.18%	33,678
3,219	Santa Isabel Municipio, Puerto Rico	0.13%	22,809
3,220	Florida Municipio, Puerto Rico	0.00%	15,629
3,221	Maricao Municipio, Puerto Rico	0.00%	6,321

## Sources

Mann, H. & D. Whitney (1947) On a Test Whether One of Two Random Variables Is Stochastically Larger Than the Other. *Annals of Mathematical Statistics*, 18, 50 - 60.

Nachar, N. (2008) The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. *Tutorials in Quantitative Mehtods for Psychology*, 4, 13 - 20.