ASSESSING SMALL SAMPLE WAR-GAMING DATASETS

W. J. HURLEY*, R. N. FARRELL**

* Department of Mathematics and Computer Science, Royal Military College of Canada, Kingston, Ontario, Canada
**Royal Military College of Canada

One of the fundamental problems faced by military planners is the assessment of changes to force structure. An example is whether to replace an existing capability with an enhanced system. This can be done directly with a comparison of measures such as accuracy, lethality, survivability, etc. However this approach does not allow an assessment of the force multiplier effects of the proposed change. To gauge these effects, planners often turn to war-gaming. For many war-gaming experiments, it is expensive, both in terms of time and dollars, to generate a large number of sample observations. This puts a premium on the statistical methodology used to examine these small datasets. In this paper we compare the power of three tests to assess population differences: the Wald-Wolfowitz test, the Mann-Whitney U test, and re-sampling. We employ a series of Monte Carlo simulation experiments. Not unexpectedly, we find that the Mann-Whitney test performs better than the Wald-Wolfowitz test. Resampling is judged to perform slightly better than the Mann-Whitney test.

Key words: *hypothesis testing, small sample, war gaming, Mann-Whitney, Wald-Wolfowitz, resampling.*

1. INTRODUCTION

One of the fundamental problems that military planners face is the assessment of changes to force structure. Whether this be the addition of a new weapon, or the substitution of a new weapon system for an old, planners must assess whether the proposed change will be beneficial. Typically this involves an assessment of whether the military benefit of the change is worth the incremental cost.

There are a number of ways to measure the military benefit. In the case of a new weapon system replacing an old, the most common way is to compare the two systems directly using criteria such as accuracy, lethality, survivability, etc. The problem with this approach is that it is difficult to gauge the force multiplier effects of the change. One way to get at this effect is to war-game both systems and then compare them on measures of effectiveness such as force exchange ratios, survivability ratios, etc. However it is usually the case that war-gaming iterations are expensive to generate both in terms of time and dollars. Hence it is important that the statistical tools used to assess populations differences in a particular measure of effectiveness be methodologically correct and as powerful as possible given the number of iterations that can reasonably be produced.

Hence this paper has two purposes. The first is to review hypothesis testing techniques for

detecting population differences in the context of war- gaming output. The second is to assess the relative power of some small sample tests to detect these differences. In particular, we examine the Wald - Wolfowitz test, the Mann-Whitney U test, and re-sampling. To do this we employ a series of Monte Carlo experiments to measure Type I and Type II errors for each test. We find that resampling performs favorably against the other two tests. However, in the case where sample size is low and the underlying variance is high, all tests give substantial Type II errors. Consequently, the choice of the number of iterations to run for each scenario is an important consideration.

2. THE PROBLEM

Suppose we are interested in the effectiveness of a *New* force structure relative to the Base force structure. To do this, we will run n iterations of a war game with the New force structure and n iterations with the *Base* force structure. We assume that there is a single measure of effectiveness, say force exchange ratio, which measures the military benefit of a given force structure, and that higher values of this measure are better than lower values. Suppose that war-gaming both force structures produces two random samples of the measure of effectiveness, one for the New force structure,

$$X = \{X_1, X_2, ..., X_n\} \quad (1)$$

and one for the Base force structure,

$$B = \{B_1, B_2, ..., B_n\}. \quad (2)$$

We assume that the X_i are iid from a distribution with mean $E(X_i) = \mu_X$ and that the Bi are iid from a distribution with $E(B_i) = \mu_B$.

In practice, it is usually the case that both force structures are run in a small number of well-defined scenarios, and multiple measures of effectiveness are considered in each scenario. This should not present a problem for the development herein. It simply means more work to test hypotheses about the performance of each measure of effectiveness in each scenario. Hence we focus on a single measure of effectiveness in a single scenario.

The other problem this paper ignores are the statistical issues associated with closely related datasets. For instance, if the New and Base force structures are run in slightly different scenarios, there is not likely to be much difference in the resulting samples, and a joint test (ANOVA) is likely to be more powerful than a series of scenario-based tests.

Let

$$\Delta_i = X_i - B_i, \quad i = 1, 2, ..., n \quad (3)$$

be the sample differences in the measure of effectiveness. We consider a simple hypothesis test about the underlying mean of these differences. By assumption we have that

$$\mu_\Delta = E(\Delta_i) = \mu_X - \mu_B \quad (4)$$

The null and alternative hypotheses are these:

$$H0: \mu_\Delta \leq 0$$
$$H1: \mu_\Delta > 0. \quad (5)$$

Most of the standard tests of this null require us to examine the difference in sample means:

$$\bar{\Delta} = \bar{X} - \bar{Y}. \quad (6)$$

To develop the properties of these tests we need to compute the sampling distribution of $\bar{\Delta}$. There are a number of possible distributions depending on the size of the samples and their underlying distributions.

Case I (Any Sample Size, Underlying Distributions are Normal): If X and B are drawn from a normal distribution with known standard deviations, σ_X and σ_B, then $\bar{\Delta}$ is normally distributed with mean $\mu_X - \mu_B$ and standard deviation

$$\sigma_\Delta = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_B^2}{n}} \quad (7)$$

Case II (Large Sample, Any Underlying Distribution): If X and B are drawn from arbitrary distributions with respective means μ_X and μ_B and unknown standard deviations and n is large enough (in this case, about15), then Δ is approximately normally distributed with mean $\mu_X - \mu_B$ and standard deviation

$$s_\Delta = \sqrt{\frac{s_X^2}{n} + \frac{s_B^2}{n}} \quad (8)$$

where s_X and s_B are the respective sample standard deviations.

Case III (Small Sample, Underlying Distributions are Normal): If X and B are drawn from normal distributions with respective means μ_X and μ_B and unknown standard deviations, and n is less than 30, then the statistic

$$t = \frac{\overline{X} - \overline{B} - (\mu_X - \mu_B)}{s_\Delta} \quad (9)$$

follows a t distribution with degrees of freedom

$$df = \left[\frac{s_X^2}{n} + \frac{s_B^2}{n}\right]^2 \left[\frac{s_X^4}{n^2(n-1)}\right] \quad (10)$$

$$= \frac{(n-1)(s_X^2 + s_B^2)^2}{s_X^4 + s_B^4} \quad (11)$$

In the case where df is not an integer, we round down to the nearest integer to be conservative.

Case IV (Small Sample, Underlying Distributions Unknown): If X and B are drawn from arbitrary distributions with unknown standard deviations, and n is less than 30, then we cannot say much about the sampling distribution of the difference in means. One approach would be to assume the underlying distributions of the measure of effectiveness were normal. However, this is not usually the case for war-gaming output. These distributions tend to be non-symmetric.

The hypothesis testing for Cases I through III is well developed. What we are interested in is statistical techniques for Case IV, as this is the most likely situation we will find ourselves in when we are studying war-gaming output.

3. TECHNIQUES FOR SMALL SAMPLES

There are a number of non-parametric tests available for small samples. We consider three: the Wald-Wolfowitz test, the Mann-Whitney U Test, also known as the Wilcoxon rank sum test, and a re-sampling procedure. The details of the first two can be found in any statistics text. See for example Aczel (1996) [1]. Over the past ten years, resampling has come to the fore as a legitimate contender to traditional parametric statistics. Good references for this technique are Efron and Tibshirani (1993) [2] and Simon (1997) [3].

3.1. The Wald-Wolfowitz Test

Suppose we have the following measure of effectiveness output for 6 iterations of a war-game with the New force structure and 6 with the Base force structure in the same scenario:

Obs.#	New, X	Base, Y
1	169.4	110.1
2	126.6	110.5
3	155.4	114.2
4	152.5	64.5
5	118.2	83.1
6	81.2	96.7

The question is whether there is sufficient evidence to conclude that

the New force structure is better.

The Wald-Wolfowitz test examines the following hypotheses:

H0 The two populations have the same distribution.

H1 The two populations have different distributions.

To apply it, we first pool the datasets and order the data from lowest to highest observation:

Obs. #	Ordered
1	64.5
2	81.2
3	83.1
4	96.7
5	110.1
6	110.5
7	114.2
8	118.2
9	126.6
10	152.5
11	155.4
12	169.4

Next, in the adjacent column, we mark from which sample the observation came:

Obs. #	Ordered	Sample
1	64.5	B
2	81.2	X
3	83.1	B
4	96.7	B
5	110.1	B
6	110.5	B
7	114.2	B
8	118.2	X
9	126.6	X
10	152.5	X
11	155.4	X
12	169.4	X

We count the number of "runs" in the sample column. A run is a sequence of like elements that is preceded and followed by a different element. For instance, in the above table, Observations 3 through 7 are all "B"; this sub- sequence is preceded by an "X" and followed by an "X". Hence observations 3-7 comprise a run. The sample statistic we are interested in is the number of runs in the Sample column. There are four of them: Observation 1, Observation 2, Observations 3-7, and Observations 8-12. Now if the New system were better we would expect a small number of runs with most of the "B" observations having low observation numbers and most of the "X" observations having high observation numbers. Tables of the cumulative distribution function for the number of runs, r, for various sample sizes can be found in Aczel (1996). If we look up the tabled value for a sample size vector (6, 6) and $r = 4$, we get 0.067. This means that, under the null that the two distributions are the same, there is only a 6.7% chance that we would observe 4 or fewer runs. Consequently we would be inclined to reject the null hypothesis and conclude that the New force structure is superior.

3.2. The Mann-Whitney U Test

Again, this test examines the same hypotheses as does the Wald-Wolfowitz test:

H0 The two populations have the same distribution

H1 The two populations have different distributions

Suppose we have the same samples as above:

Obs.#	New, *X*	Base, *Y*
1	169.4	110.1
2	126.6	110.5
3	155.4	114.2
4	152.5	64.5
5	118.2	83.1
6	81.2	96.7

In the same way as above we order the sample observations:

Obs. #	Ordered	Sample
1	64.5	B
2	81.2	X
3	83.1	B
4	96.7	B
5	110.1	B
6	110.5	B
7	114.2	B
8	118.2	X
9	126.6	X
10	152.5	X
11	155.4	X
12	169.4	X

Next we add the "ranks" of those sample observations coming from the "X" sample:

$R_X = 2 + 8 + 9 + 10 + 11 + 12 = 52.$ (12)

We then form the test statistic *U*:

$$U = n^2 + \frac{n(n+1)}{2} - R_X = 6 \cdot 6 + \frac{6 \cdot 7}{2} - 52 = 5 \quad (13)$$

The cumulative distribution for the *U* statistic are also tabled (see Aczel (1996)). The chance that we would observe a value of *U* of 5 or less is .0206. Hence we would reject the null hypothesis that both samples are from the same distribution. Our conclusion is that the New force structure has a higher measure of effectiveness.

3.3. Resampling

Here is how resampling works with the dataset used above. We first resample from the "New, X" sample by picking 6 points with replacement. This can be accomplished in EXCEL using 6 calls of the function

=SMALL(data_range,INT(n* RAND())+1)

where data_range contains the original 6 datapoints and $n = 6$. Next we do the same thing with the "Base, B" dataset. One resampling of each dataset is shown in the following table:

Obs.#	New, *X*	Base, *Y*
1	81.2	96.7
2	152.5	83.1
3	81.2	110.1
4	81.2	83.1
5	152.5	64.5
6	126.6	83.1

Note that there are repeats in both resamples. For instance in the "New, X" resample, 81.2 appears three times and 152.5 appears twice.

Next we compute a sample average for each resample. For the "New, X" resample, it is 112.5 and for the "Base, B" it is 91.5. We then record which sample mean is higher. Finally we repeat this experiment a large number of times (I usually do 10,000 iterations using the @RISK add-in to EXCEL) and count the number of times that the "New, *X*" average is greater than the "Base, B" average. In the 10,000 iterations I did, the "New, *X*" average exceeded the "Base, B" average 9,950 times. From this we conclude that the p-value of the null is

$$1 - \frac{9950}{10000} = .005 \qquad (14)$$

Hence we would reject the null hypothesis and conclude that the New force structure gives a higher measure of effectiveness.

3.4. A Monte Carlo Experiment

In order to measure the effectiveness of these three tests, we designed a Monte Carlo experiment with the following steps:

Step 1: Generate two random samples of size *n* of the measure of effectiveness for each force structure from normal distributions with the following parameters:

	New, X	New, Y
Mean	100+d	100
Standard Deviation	σ	σ

where we will vary the parameters *n, d* and *σ*. In the case where *d = 0*, there is no difference in the two systems. Where *d > 0*, the New system is superior.

Step 2: Using the samples generated in Step 1, measure *p*-values for each of the three tests described above.

We did Steps 1 and 2 a large number of times for fixed values of *d, n,* and *σ*. This gave us distributions of *p*-values for each of the three tests. We then compared these distributions to determine which is more effective. The case *d = 0* was used to measure Type I error. For each test we observed the percentage of time that it produces a *p*-value lower than .05 (and, in addition, .10). The cases *d = 10* (μ_X = 110, μ_B = 100) and *d = 20* (μ_X = 120, μ_B = 100) were used to measure Type II error.

4. RESULTS

We first examine Type I error results using $\mu_X = \mu_B = 100$ (the proposed system provides no additional benefit). We are interested in the proportion of time each of the tests provides a p-value less that 5% and 10%. Table 1 summarizes our results for two values of n (n = 5, 10) and 5 values of σ (σ = 5, 10, 15, 20, 25). First note that the resulting p-values are consistent with the various relative values for n and σ. For instance, for a given n, Type I errors increase as σ gets larger. We would expect this to happen as, in general, a larger standard deviation tends to mask the correct conclusion.

In addition, for a given σ, Type I errors fall as n increases. With larger sample sizes you would expect to have lower Type I errors.

In general, all three tests return empirical Type I errors that are low. The Wald-Wolfowitz and Mann-Whitney tests tend to perform about the same. Each gives slightly lower Type I errors than resampling.

Now to Type II Error. We examine this error for two assumptions, μX = 110, μB = 100, and μX = 120, μB = 100. Table 2 gives our results for μX = 110, μB = 100; Table 3 corresponds to μX = 120, μB = 100. First note that the Mann-Whitney and resampling perform much better than the Wald-Wolfowitz test. They each give uniformly lower Type II errors. Note as well that resampling outperforms the Mann-Whitney test, although this enhanced performance is quite small for some parameter values. Finally, we note that none of these tests is very good if data uncertainty, σ, is large relative to μX and μB. For instance, if μX = 110 and n = 5, the Type II errors are at least 75% for each test.

Table no. 1: Simulation Results for μ_X =100, μ_B = 100

TYPE I ERROR, $\mu_X = \mu_B = 100$						
n=5						
	Fraction of p-values ≤ 5%			*Fraction of p-values ≤ 10%*		
σ	Wald	Mann	Resampling	Wald	Mann	Resampling
5	.038	.048	.090	.038	.076	.138
10	.044	.052	.083	.044	.078	.134
15	.041	.048	.088	.041	.073	.137
20	.039	.048	.091	.039	.075	.139
25	.040	.045	.089	.040	.073	.143
n=10						
	Fraction of p-values ≤ 5%			*Fraction of p-values ≤ 10%*		
σ	Wald	Mann	Resampling	Wald	Mann	Resampling
5	.018	.044	.068	.050	.094	.120
10	.020	.044	.067	.049	.096	.116
15	.018	.040	.076	.054	.091	.125
20	.019	.048	.068	.050	.094	.119
25	.019	.048	.071	.049	.099	.124

Table no. 2: Simulation results for μ_X = 110, μ_B = 100

TYPE II ERROR, *μX = 110, μB = 100*						
n=5						
	Fraction of p-values > 5%			*Fraction of p-values > 10%*		
σ	Wald	Mann	Resampling	Wald	Mann	Resampling
5	.613	.135	.053	.613	.083	.026
10	.894	.601	.442	.894	.501	.326
15	.934	.768	.640	.934	.685	.526
20	.946	.837	.729	.946	.771	.631
25	.951	.861	.776	.951	.805	.687
n=10						
	Fraction of p-values > 5%			*Fraction of p-values > 10%*		
σ	Wald	Mann	Resampling	Wald	Mann	Resampling
5	.272	.007	.002	.265	.002	.001
10	.833	.351	.254	.798	.209	.157
15	.930	.621	.519	.893	.464	.394
20	.953	.747	.660	.918	.612	.541
25	.965	.805	.736	.932	.675	.621

Table no. 3: Simulation results for $\mu_X = 120$, $\mu_B = 100$

Type II Error, $\mu_X = 120$, $\mu_B = 100$						
n=5						
	Fraction of p-values > 5%			*Fraction of p-values > 10%*		
σ	Wald	Mann	Resampling	Wald	Mann	Resampling
5	.042	.000	.000	.042	.000	.000
10	.609	.136	.055	.609	.083	.027
15	.833	.419	.264	.833	.319	.171
20	.899	.602	.449	.899	.503	.333
25	.922	.708	.558	.922	.618	.442
n=10						
	Fraction of p-values > 5%			*Fraction of p-values > 10%*		
σ	Wald	Mann	Resampling	Wald	Mann	Resampling
5	.000	.000	.000	.000	.000	.000
10	.275	.007	.002	.269	.002	.001
15	.671	.147	.083	.644	.068	.042
20	.837	.355	.254	.801	.208	.162
25	.892	.514	.404	.856	.352	.286

5. CONCLUSIONS

In this paper we have reviewed some non-parametric methods for hypothesis tests of differences in population means for small sample war-gaming datasets. Our experiments, although limited in scope, indicate that resampling compares quite favorably with other tests. However, it should be noted that no test is capable of overcoming a shortage of data, data that may be quite variable. Planners should be aware that if sample variance is high relative to the mean measure of effectiveness, there is a good chance that a small-sample dataset will not be able to pick up that the New force structure is better than the Base when that is actually the case.

Future work will assess the accuracy of ANOVA tests in multiple, smallsample scenarios.

REFERENCES

[1] Aczel, A. D., Complete Business Statistics, Irwin, Chicago, 1996.

[2] Efron, Bradley, and Robert J. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall, New York, 1993.

[3] Simon, Julian L., Resampling: The New Statistics, Resampling Stats, Arlington VA, 1997.