A Methodology for Appropriate Testing When Data is Heterogeneous Using EXCEL

Nguyen Khanh[1], Jimin Lee[2], Susan Reiser[3], Donna Parsons[4], Sara Russell[5], and Robert Yearout[1]

[1] Management and Accountancy
[2]Departments of Mathematics,
[3]New Media
The University of North Carolina Asheville
Asheville, NC 28804

[4]Department of Business
Mars Hill University
Mars Hill, NC 28754

[5]Department of Statistics
North Carolina State University
Raleigh, NC 27606

Corresponding author's Email: yearout@unca.edu

Author Note: Nguyen Khanh graduated from the University of North Carolina Asheville with a degree in Accountancy in May 2015. Jimin Lee, Associate Professor of Statistics, has published many articles in statistics and bio-statistics; she has also published in international industrial engineering journals and spoken in conference proceedings. Donna Parsons, Assistant Professor of Business, has published numerous business articles as well as in industrial engineering journals and spoken in proceedings. Susan Reiser, Professor of New Media, has published a significant number of articles in international journals and proceedings. Sara Russell graduated from North Carolina State University with a master of Statistics degree with a concentration in biostatistics in May 2015, and Robert Yearout, Professor of Industrial Engineering Management, has published a significant number of articles in international journals and proceedings.

Abstract: *A Methodology for Appropriate Testing When Data is Heterogeneous* was originally published and copy written in the mid-1990s in Turbo Pascal and a 16-bit operating system. While working on an ergonomic dissertation (Yearout, 1987), the author determined that the perceptual lighting preference data was heterogeneous and not normal. Drs. Milliken and Johnson, the authors of *Analysis of Messy Data Volume I: Designed Experiments* (1989), advised that Satterthwaite's Approximation with Bonferroni's Adjustment to correct for pairwise error be used to analyze the heterogeneous data. This technique of applying linear combinations with adjusted degrees of freedom allowed the use of t-Table criteria to make group comparisons without using standard nonparametric techniques. Thus data with unequal variances and unequal sample sizes could be analyzed without losing valuable information. Variances to the 4th power were so large that they could not be reentered into basic calculators. The solution was to develop an original software package which was written in Turbo Pascal on a 7 ¼ inch disk 16-bit operating system. Current operating systems of 32 and 64 bits and more efficient programming languages have made the software obsolete and unusable. Using the old system could result either in many returns being incorrect or the system terminating. The purpose of this research was to develop a spreadsheet algorithm with multiple interactive EXCEL worksheets that will efficiently apply Satterthwaite's Approximation with Bonferroni's Adjustment to solve the messy data problem. To ensure that the pedagogy is accurate, the resulting package was successfully tested in the classroom with academically diverse students. A comparison between this technique and EXCEL's Add-Ins Analysis ToolPak for a *t-test Two-Sample Assuming Unequal Variances* was conducted using several different data sets. The results of this comparison were that the EXCEL Add-Ins returned incorrect significant differences. Engineers, ergonomists, psychologists, and social scientists will find the developed program very useful. A major benefit is that spreadsheets will continue to be current regardless of evolving operating systems' status.

Keywords: Heterogeneous Data, Unequal Sample Sizes, Satterthwaite's Approximation with Bonferroni's Adjustment, EXCEL

1. Introduction

This project began as an effort to obtain a more efficient, user friendly, replacement for the *Messy Data Assistant* (1999) that was published in the International Journal of Industrial Ergonomics. After reviewing numerous articles for publication in industrial engineering, ergonomics, and business journals, the authors discovered that testing for heterogeneous and unequal data sets was prevalent. This oversight in many cases resulted in reporting significant differences when there were none (R. Barger, R. Yearout, and G. Yates, 1995).

1.1. Background and Problem

For many experiments, the investigator examines and compares the effects of different treatments or the means of treatment populations. Heterogeneous variances, unequal sample sizes, and non-normal data are quite common. Perceptional or survey data gathered by industrial engineers, ergonomists, or social scientists are especially vulnerable. While working on an ergonomic dissertation (Yearout, 1987), the author determined that the perceptual data was determined to be heterogeneous and not normal. Drs. Milliken and Johnson, the authors of *Analysis of Messy Data Volume I: Designed Experiments* (1989), advised that Satterthwaite's Approximation with Bonferroni's Adjustment to correct for pairwise error be used to analyze the heterogeneous and not normal data. Unequal sample sizes also will contribute to making incorrect inferences. Since variances to the 4th power were so large that they could not be reentered into basic calculators, *The Messy Data Assistant* (Yearout, R. Barger, R. Yates, G. and Lisnerski D., 1999) was published and copyrighted. The software package's algorithm was written in Turbo Pascal on a 7¼ inch disk 16-bit operating system. This technique of applying linear combinations with adjusted degrees of freedom allowed the use of t-Table criteria to make group comparisons without using standard nonparametric techniques. Thus not normal data with unequal variances and unequal sample sizes could be analyzed without losing valuable information.

Current operating systems of 32 and 64 bits and more efficient programming languages have made the software obsolete and unusable. Using the old system could result either in many returns being incorrect or the system terminating when executed.

1.2 Why Use Satterthwaite's Approximation

Such examinations may include the following type of hypotheses (equation 1, 2, and 3) (Milliken and Johnson, 1984)

$$H_{01}: \sum c_i \mu_i = a \quad (1)$$

for some given set of coefficients c_1, c_2,..., c_t and given constant a and:

$$H_{02}: \mu_1 = \mu_2 = ... = \mu_t \quad (2)$$

$$H_{03}: \mu_i = \mu_{i'} \quad (3)$$

for at least one $i \neq i'$

Unfortunately, anecdotal conversations with practicing industrial engineers, ergonomists, social scientist, and statisticians suggest that these types of hypotheses do not conform to the assumption of normality; i.e., that error terms are independently and identically distributed. Also, such error terms in groups must have a mean equal to zero or variances for homogeneity. Both conditions, however, are often violated. Thus, a parametric test which depends on the crucial assumption that the investigator is sampling randomly from a distribution belonging to a particular family may be inappropriate (Sprent, 1989). Additionally, small and equal sample groups can create complicated observations. The t-test is considered sufficient to handle non-normal distribution. Its reliability, however, is questionable when unequal variances and an unequal sample condition exists. In general, the issues of this inequality are more critical than the distribution of the data. As a result, the t-test may not be appropriate. When these conditions are present, the investigator must change the techniques from traditional (parametric) to non-traditional (distribution free). Satterthwaitte's Approximation estimates the variance of a mean and variance components, or is used to construct an approximate F-test. It is necessary to utilize such approximation to form a linear function of mean squares, $\sigma^2 = \sum c_i^2 \sigma^2_{i'}$, where c_i are known constants (Satterthwaite, 1946). The distribution of error terms may or may not be strictly normal when the conditions can be assumed to approximate normality.

The procedure is illustrated as follows (equation 4):

$$\nu = \frac{\left(\sum C^2{}_i \sigma^2{}_{i'}/n_i\right)}{\left(\sum_i \left[\frac{C^4{}_i \sigma^4{}_{i'}}{n_i^2(n_i-1)}\right]\right)} \tag{4}$$

Summarizing, one rejects the hypothesis (Eq. (5)):

$$H_o: \sum C_i^2 \sigma_i^{\prime 2} = a, \tag{5}$$

If (Eq. 6)

$$|t_c| = \frac{|\sum c_i \mu_i - a|}{\sqrt{\frac{\Sigma C^2{}_i \sigma^2{}_i}{n_i}}} > t_{\alpha/2, \nu} \tag{6}$$

This technique is appropriate for unequal variance (heterogeneous data) and unequal sample sizes. The distribution of error terms may or may not be strictly normal when the conditions can be assumed to approximate normality. This method is allowed for a good approximation by estimating the degree of freedom ν for unequal variances.

The adjusted "degree of freedom" and "critical t" (t-test) are used to test the above hypotheses. The t-test retains the original information and is insensitive to unequal sample sizes as well. Yearout (1987) used a simple Turbo Pascal (1984) program to demonstrate the method. This program, however, required the user to have a Pascal compiler and be Pascal literate. Therefore, the program had limited utility and was not user-friendly.

1.3 Bonferroni's Adjustment for Pairwise Error

Another common analytical error is failure to consider reduced reliability of the stated confidence level. As a result, the user may make an error when concluding the significant differences between groups. Bonferroni proposed a method to determine the appropriate minimum significant level to obtain a desired confidence level (Neter. et al., 1990). The procedure to determine the confidence of any comparison is illustrated by equation 7.

Confidence level = 100(1-kα),

where the number of k intervals are calculated by:

$$k = \binom{I}{2} = \frac{I(I-1)}{2} \tag{7}$$

1.4 Research Purpose

Current operating systems of 32 and future 64 bits and programming languages have made the software obsolete and may return incorrect solutions. The purpose of this research is to develop a spreadsheet algorithm with multiple interactive worksheets that will solve the problem of messy data analysis. To insure that the pedagogy is accurate the resulting package was successfully tested in the classroom with academically diverse students.

2. Algorithm Development

The Algorithm was developed using the data set, by permission, described in Milliken and Johnson (1989). One spreadsheet file with three interactive worksheets (Data Entry, Calculations, and Results) was used. Each of these worksheets with specific instructions is contained in the following sections.

2.1 Data Entry Worksheet

Format the Excel worksheet as follows. All strings and sample data (noted in black) such as A1: Enter "Data Set 1" (table 1) appear in the appropriate cells on all three tabs. All calculations and results will be displayed individually and be colored in blue. Confidence level is 95%.

A1: Enter "Data Set 1".
C2: Enter =AVERAGE(B8:B15).
C3: Enter =STDEV(B8:B15).
C4: Enter =(C3^2).
C5: Enter =(A15).
C6: Enter confidence level
G6: Enter number of groups
A8:A15: Enter number of each sample within Data Set 1
B8:B15: Enter Data Set 1.

Table 1. Data Set 1 (Data Entry Worksheet) Columns A, B, and C

	A	B	C
1	***Data Set 1***		
2		mean (x-bar$_1$)	8.625
3		stdev (s_1)	3.113909
4		Variance (s^2_1)	9.696429
5		Sample (n_1)	8
6		Confidence Level	0.95
7	***Sample***	***Data Set 1***	
8	1	12	
9	2	4	
10	3	11	
11	4	7	
12	5	8	
13	6	10	
14	7	12	
15	8	5	

Repeat procedure for the same data set (Data Set 2). Copy table 1 to column E, F, and G. Then modify to perform required statistical calculations for Data Set 2 as shown in table 2.

Table 2. Data Set 2 (Data Entry Worksheet) Columns E, F, and G

	E	F	G
1	***Data Set 2***		
2		mean (x-bar$_2$)	11.66667
3		stdev (s_2)	1.36626
4		Variance (s^2_2)	1.866667
5		Sample (n_2)	6
6		Number of Groups	3
7	***Sample***	***Data Set 2***	
8	1	12	
9	2	10	
10	3	13	
11	4	13	
12	5	12	
13	6	10	

Repeat procedure for the same data set (Data Set 2). Copy Table 1 to column I, J, and K. Then modify to accept and perform required statistical calculations for Data Set 3 as shown in table 3.

Table 3. Data Set 3 (Data Entry Worksheet) Columns I, J, and K

	I	J	K
1	***Data Set 3***		
2		mean (x-bar$_3$)	13.750
3		stdev (s_3)	1.669046
4		Variance (s^2_3)	2.785714
5		Sample (n_3)	8
6			
7	***Sample***	***Data Set 3***	
8	1	13	
9	2	14	
10	3	14	
11	4	17	
12	5	11	
13	6	14	
14	7	13	
15	8	14	

2.2 Computation Worksheet

For Column A and cells B1, B20, B22, C17, C22, E1, E20, E22, and F22 enter the labels as shown in table 4.

B2: Enter =('Data Entry'!C2).
B3: Enter ==('Data Entry'!C3).
B4: Enter =('Data Entry'!C4).
B5: Enter =('Data Entry'!C5).
B6: Enter =(B4)^2.
B7: Enter =(B2-E2)
B8: Enter =((B4/B5)+(E4/E5))^0.5
B9: Enter =(B6+E6)
E9: Enter =(E6)
B10: Enter =(B5^2).
B11: Enter =(B5-1).
B12: Enter =(B10*B11).
B13: Enter =(B6/B12).
B14: Enter =(B13+E13).
B15: Enter =(B8^2)
B16: Enter =(B7/B8).
B17: Enter =(B15^2)/(B14).
B18: Enter =('Data Entry'!C6).
B23: Enter =('Data Entry'!B8). Drag down to B30.
C23: Enter =('Data Entry'!A8). Drag down to C30.
D16: Enter =-ABS(B16)
D17: Enter =T.DIST(D16,B17,1)*('Data Entry'!G6).
E2: Enter =('Data Entry'!G2).
E3: Enter ==('Data Entry'!G3).
E4: Enter =('Data Entry'!G4).
E5: Enter =('Data Entry'!G5).
E6: Enter =(E4)^2.
E10:Enter =(E5^2).
E11: Enter =(E5-1).
E12: Enter =(E10*E11).
E13: Enter(E6/E12).
E23: Enter =('Data Entry'!F8). Drag down to E28.
F23: Enter =('Data Entry'!E8). Drag down to F28.

Table 4. Data Set 1 and Data Set 2 (Calculations) Worksheet Columns A through F

	A	B	C	D	E	F
1		**Data Set 1**			**Data Set 2**	
2	Mean (x-bar)	8.625			11.66667	
3	Standard Deviation (s)	3.113909			1.36626	
4	Variance (s^2)	9.696429			1.866667	
5	n =	8			6	
6	Varriance $(s^2)^2$	94.02073			3.484444	
7	$l_1 = \mu_1 - \mu_2 =$	-3.04167				
8	s.e. (combination) =	1.234166				
9	Σs_i^4	97.50517			3.484444	
10	n^2	64			36	
11	n-1	7			5	
12	n^2 *(n-1)	448			180	
13		0.209868			0.019358	
14	correction factor	0.229226				
15	s^2 of combination	1.523165				
16	***Critical t =***	-2.46455		-2.46455		
17	***Adjusted d.f. (v) =***	10.12116	p =	0.050123		
18	Confidence Level	0.95				
19						
20	**Data Entry**	**Data Set 1**			**Data Set 2**	
21						
22		**Value**	**Sample**		**Value**	**Sample**
23		12	1		12	1
24		4	2		10	2
25		11	3		13	3
26		7	4		13	4
27		8	5		12	5
28		10	6		10	6
29		12	7			
30		5	8			

Repeat procedure for Data Set 1/Data Set 3(using columns I through N), and Data Set 2/Data Set 3(using columns Q through V). Copy table 4 to column I through N. Then modify to accept and perform required statistical calculations for Data Set 1 and Data Set 3 as shown in table 5.

Table 5. Data Set 1 and Data Set 3 (Calculations) Worksheet Columns I through N

	I	J	K	L	M	N
1		**Data Set 1**			**Data Set 3**	
2	Mean (x-bar)	8.625			13.75	
3	Standard Deviation (s)	3.113909			1.66905	
4	Variance (s^2)	9.696429			2.78517	
5	n =	8			8	
6	Variance $(s^2)^2$	94.02073			7.7602	
7	$l_2 = \mu 1 - \mu 3 =$	-5.125				
8	s.e. (combination) =	1.24911				
9	Σs_i^4	101.781			7.7602	
10	n^2	64			64	
11	n-1	7			7	
12	n^2 *(n-1)	448			448	
13		0.209868			0.01732	
14	correction factor	0.22719				
15	s^2 of combination	1.56027				
16	***Critical t =***	-4.10293		-4.10293		
17	***Adjusted d.f. (v) =***	10.71544	p =	0.003202		
18	Confidence Interval	0.95				
19						
20	**Data Entry**	**Data Set 1**			**Data Set 3**	
21						
22		**Value**	**Sample**		**Value**	**Sample**
23		12	1		13	1
24		4	2		14	2
25		11	3		14	3
26		7	4		17	4
27		8	5		11	5
28		10	6		14	6
29		12	7		13	7
30		5	8		14	8

Repeat procedure for the same Data Set 2 and Data Set 3. Copy Table 4 to column Q through V. Then modify to accept and perform required statistical calculations for Data Set 2 and Data Set 3 as shown in table 6.

Table 6. Data Set 2 and Data Set 3 (Calculations) Worksheet Columns Q through V

	Q	R	S	T	U	V
1		**Data Set 2**			**Data Set 3**	
2	Mean (x-bar)	11.666667			13.750	
3	Standard Deviation (s)	1.3662601			1.66905	
4	Variance (s^2)	0.811989			2.785714	
5	N	6			8	
6	Variance $(s^2)^2$	3.4844444			7.760204	
7	$l_3 = \mu_2 - \mu_3 =$	-2.083333				
8	s.e. (combination) =	0.811989				
9	Σs^4	11.24465			7.760204	
10	n^2	36			64	
11	n-1	5			7	
12	n^2 *(n-1)	180			448	
13		0.019358			0.031669	
14	correction factor	0.03668				
15	s^2 of combination	0.659325				
16	***Critical t* =**	-2.565718		-2.565718		
17	***Adjusted d.f. (v)* =**	11.85145	p =	0.0393748		
18						
19						
20	**Data Entry**	**Data Set 2**			**Data Set 3**	
21						
22		**Value**	**Sample**		**Value**	**Sample**
23		12	1		13	1
24		10	2		14	2
25		13	3		14	3
26		13	4		17	4
27		12	5		11	5
28		10	6		14	6
29					13	7
30					14	8

2.3 Results Worksheet

The result summary connects all results from Data Entry and Calculations worksheets. This worksheet displays all outcomes on the Result tab (table 7). The detailed results include critical t, adjusted degrees of freedom (d.f.), Bonferroni, and significant level. Same calculations apply for data set 1, 2 and 3. The users need to navigate back to the Data Entry and Calculations tab to construct the Results worksheet. Spreadsheet cells are as follows:

For Column A and cells B1, B20, B21, D2, D5 through D8, G2, G5 through D8, enter the labels as shown in tables 7 and 8. To obtain color highlight in the results, refer to the main Tab conditioning format procedure.

B5: Enter =(Calculations!B2). Drag down to B8. Repeat for E5:E8 and H5:H8.
B10: Enter =(Calculations!B7).
B11: Enter =(Calculations!B8)
B12: Enter =(Calculations!B16)
B13: Enter =(Calculations!B17)
B14: Enter =('Data Entry'!C6)
B15: Enter =(1-B14).
B17: Enter =(Calculations!D17).
B18: Enter =IF(B17>B15, B20, B21).

Table 7. Results of Analysis Worksheet Columns A and B

	A	B
1		
2	***Data Set 1***	
3		
4		
5	Mean (x-bar) =	8.625
6	Standard Deviation (s) =	3.113908889
7	Variance (s^2) =	9.696428571
8	n =	8
9		
10	**$l_1 = \mu_1 - \mu_2$ =**	-3.041666667
11	s.e. (combination) =	1.234165581
12	***Critical t* =**	-2.464553146
13	***Adjusted d.f. (v)* =**	10.12116206
14	Confidence level =	0.95
15	p normal=	0.05
16		
17	***Bonferroni p* =**	0.050122512
18	***Significant Differences* =**	***Not Significant***
19		
20		***Not Significant***
21		***Significant Difference***
22		
23	**$l_2 = \mu_1 - \mu_3$ =**	-5.125
24	s.e. (combination) =	1.249106824
25	***Critical t* =**	-4.102931713

Note: To obtain color highlight in the results, refer to the main Tab conditioning format procedure.

Table 7A. Results of Analysis Worksheet Columns A and B (Continued)

	A	B
26	Adjusted degrees of freedom (v) =	10.71543776
27	Confidence Level =	0.95
28	p normal =	0.05
29		
30	***Bonferroni p =***	0.003201815
31	***Significant Differences =***	***Significant Difference***
32		
33	**$l_3 = \mu_2 - \mu_3 =$**	-2.083333333
34	s.e. (combination) =	0.811988545
35	***Critical t =***	-2.565717641
36	***Adjusted degrees of freedom (v) =***	11.85144656
37	Confidence Level =	0.95
38	p normal =	0.05
39		
40	***Bonferroni p =***	0.039374843
41	***Significant Differences =***	***Significant Difference***

Table 8. Results of Analysis Worksheet Columns D Through H

	D	E	F	G	H
1					
2	***Data Set 2***			***Data Set 3***	
3					
4					
5	Mean (x-bar)	11.6666667		Mean (x-bar)	13.75
6	Standard Deviation (s)	1.3662601		Standard Deviation (s)	1.669045921
7	Variance (s^2)	1.86666667		Variance (s^2)	2.785714286
8	n	6		n	8

3. Results and Comparison

By using this method, the researcher can analyze data with unequal variances and sample sizes without losing valuable information. The results also include critical t, adjusted d.f, Bonferroni, and significance level. Figure 1 illustrates that there is no significant difference between Data Set 1 and Data Set 2, and that there is a significant difference between Data Set 1 and Data Set 3 and between Data Set 2 and Data Set 3.

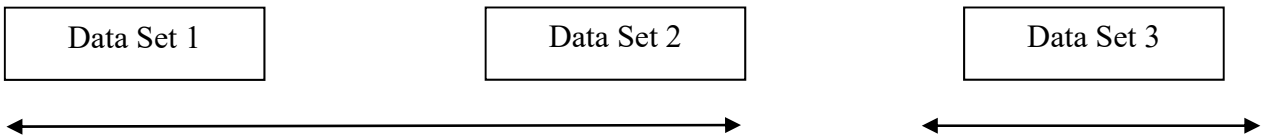

Figure 1. Results Diagram for Satterthwaite's Approximation with Bonferroni's Adjustment

4. Comparison to EXCELs Module for Heterogeneous Data

A comparison was made with the same data set used in *Analysis of Messy Data* (Yearout, Barger, Yates, and Lisnerski, 1999). This paper's Data Analysis module for 't-Test Two Samples Assuming Unequal Variances' (heterogeneous) program in EXCEL presents completely different results when compared to the above method. Figure 2 indicates that there is a significant difference between Data Set 1 and Data Set 2, Data Set 1 and Data Set 3, and Data Set 2 and Data Set 3.

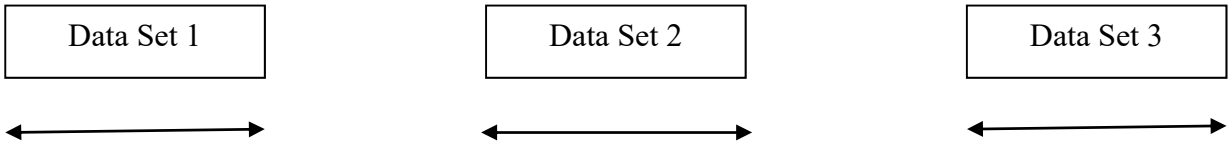

Figure 2. Results Diagram for EXCEL's t-Test Two Samples Assuming Unequal Variances

5. Discussion

EXCEL's Function provides an absolutely different result because it does not contain Satterthwaite's Approximation, Bonferoni's Adjustment, and appropriate Student t-Table. Failure to consider the pairwise error rate by not using Bonferroni's Adjustment can result in making statistical inferences that contain Type I errors. The algorithm illustrated in this paper is not only user-friendly and operates from any operating system, but it also contains the appropriate adjustments. This comparison clearly illustrates its value to practicing engineers, ergonomists, and social scientists who would find the developed program very useful. A major benefit is that spreadsheet algorithms will continue to be current regardless of evolving operating systems' status.

6. Conclusion

The resulting package was successfully tested in classrooms in different universities with academically diverse students to insure that the pedagogy is accurate. The results of the comparison were that the EXCEL Add-Ins returned incorrect significant differences. The value of this research is that spreadsheet algorithms will continue to be current regardless of the evolving operating systems' status. In addition, EXCEL is available to the engineer and researcher worldwide.

7. References

Barger, R., Yearout, R., and Yates, G. (1995). *Statistical Package for Analyzing Messy Ergonomic Data, Advances in Industrial Ergonomics and Safety VII*, Proceedings of the Annual International Foundation for Industrial Ergonomics and Safety Research Conference. London: Taylor and Francis.

Microsoft (2013*). Microsoft Excel*. Redmond, Washington: Microsoft, 2013. Computer Software.

Milliken, G. and Johnson, D. (1984). *Analysis of Messy Data Volume I: Designed Experiments.* Belmont: Lifetime Learning Publications.

Netter, J., Wasserman, W., and Kutner, M., (1990). *Applied Linear Statistical Models, 3ed*. Boston: Irwin.

Satterthwaite, F.E. (1946). *Biometrics Bulletin, 2,* pp. 11-114.
Sprent, P. (1989). *Applied Non-Parametric Statistical Methods*. New York: Chapman & Hall.
Yearout, R. (1987). *Task Lighting for Visual Display Unit Work Stations*. Manhattan: Kansas State University.
Yearout, R., Barger, R., Yates, G., and Lisnerski, D. (1999). A Methodology for Appropriate Testing When Data are Heterogeneous. *International Journal of Industrial Ergonomics*, *24*, pp. 129-134.