



Assessing Normality: Applications in Multi-Group Designs

^{1*}Abdul R. Othman, ²H. J. Keselman and ³Rand Wilcox

¹*School of Distance Education,
Universiti Sains Malaysia,
11800 Pulau Pinang, Malaysia*

²*P501, Duff Roblin Building, Department of Psychology,
University of Manitoba, 190 Dysart Road,
Winnipeg, MB R3T 2N2, Canada*

³*Department of Psychology,
University of Southern California,
Los Angeles, CA 90089-1061,
United States of America*

E-mail: oarahman@usm.my

*Corresponding author

ABSTRACT

Warr and Erich (2013) compared a frequently recommended procedure in textbooks: the interquartile range divided by the sample standard deviation, against the Shapiro-Wilk's test in assessing normality of data. They found the Shapiro-Wilk's test to be far superior to the deficient interquartile range statistic. We look further into the issue of assessing non-normality by investigating the Anderson-Darling goodness-of-fit statistic for its sensitivity to detect non-normal data in a multi-group problem where Type I error and power issues can be explored from perspectives not considered by Warr and Erich. In particular, we examined the sensitivity of this test for 23 non-normal distributions consisting of g- and h-distributions, contaminated mixed-normal distributions and multinomial distributions. In addition, we used a sequentially-rejective Bonferroni procedure to limit the overall rate of Type I errors across the multi-groups assessed for normality and defined the power of the procedure according to whether there was at least one rejection from among the three group tests, whether all three non-normal groups of data were detected and the average of the per-group power values. Our results indicate that the Anderson-Darling test was generally effective in detecting varied types of non-normal data.

Keywords: Non-normal data, Anderson-Darling, goodness-of-fit statistics, Power, Familywise control over the multiple significance tests for normality.

1. INTRODUCTION

Warr and Erich (2013) demonstrate how the interquartile range divided by the sample standard deviation is not a good method for assessing whether data is normal in form. Their finding is not surprising in that it is well known that the sample standard deviation can be inflated by outlying values from skewed non-normal distributions and accordingly the value of the interquartile range divided by such an inflated value would result in small values and thus accordingly underestimate the degree of non-normality (See Wilcox (2012)). Warr and Erich's simulations confirm this lack in power to detect non-normality.

Testing the normality assumption has been a matter of concern for quite some time since it is well known that classical test statistics such as Student's two-independent sample t-test and the analysis of variance (ANOVA) F-test can produce too many false rejections (i.e., inflated rates of Type I error) and/or lack sensitivity to detect effects (i.e., low rates of power) when data do not conform to the normality derivational assumption. This anomaly can be exacerbated when other derivational assumptions are not also satisfied [e.g., homogeneity of variances (See Zimmerman (2004), (2010))]. Accordingly, it is quite common for authors of textbooks and research articles to recommend that normality and/or variance homogeneity be assessed prior to performing the main test for say treatment group mean equality.

Recently, Schoder *et al.* (2006) examined the Kolmogorov-Smirnov (K-S) goodness-of-fit statistic (See D'Agostino and Stephens, (1986)) for detecting various types of non-normal distributions in a one-sample pretest-posttest design and found that it lacked sufficient power to detect non-normal distributions with sample sizes less than 100. Keselman *et al.* (2013) followed-up on the work of Schoder *et al.* but they included two additional goodness-of-fit statistics, namely the Cramer-von Mises (CvM) and Anderson-Darling (A-D) statistics (See Muller and Fetterman (2002)) and found the tests to have sufficient power to detect many types of non-normal distributions with sample sizes less than 100. Keselman *et al.* (2013) recommended the A-D test with a significance level of .15 or .20. Later Keselman *et al.* (2014) extended their investigation to the case of detecting non-normality in the multi-group problem. In particular, they examined a one-way design having three treatment groups in which the main statistical test would be the ANOVA *F*-test.

Again Keselman *et al.* (2014) found that the A-D goodness-of-fit test was most powerful of the fit statistics and once again recommended that applied researchers adopt a more liberal criterion for significance (i.e., $\alpha = .15$ or $\alpha = .20$). However, in the multi-group problem they suggested that researchers set these more liberal significance levels over the family of fit-statistics with Hochberg's (1988) sequentially-rejective Bonferroni procedure familywise critical values. Thus, as they claimed, researchers can detect many different forms of non-normality with the A-D goodness-of-fit statistic and control the overall rate of Type I error with a procedure that limits the multiplicity of testing problem.

2. METHOD

We will present some simulation results to extend the findings reported by Keselman *et al.* (2014) regarding the A-D goodness-of-fit statistic. In our study we investigated non-normal distributions that differed in degree of skew and kurtosis in a multi-group setting. We created three groups with the same distributions. One set of non-normal distributions was created by converting standardized unit normal variates to g and h -variates (See Headrick, Kowalchuk and Sheng (2008); Hoaglin (1985), (1988); Kowalchuk and Headrick (2010)). Thirteen g - and h - distributions were created (See Table 1). The reader can gauge whether a distribution is a symmetric short-tailed distribution, a symmetric long-tailed distribution, or asymmetric by comparing the skewness and kurtosis indices reported in Table 1 against 1. For example, the $g = 0, h = k$ distributions with $k \geq 1$ are all symmetric long-tailed distributions because all of their kurtosis values are greater than 2. As well, we examined six contaminated mixed-normal distributions displayed in Table 2 (See Zimmerman (2004)) and four Likert-type data from multinomial distributions displayed in Table 3 (See Schoder *et al.* (2006)).

We chose contaminated mixed-normal and Likert type data (i.e., multinomial) because these distributions were examined in previous investigations and because they typify psychological data collected by applied researchers. As for the contaminated mixed normal distributions, they are of course symmetric. The length of their tails depend upon the proportion of $N(0, \sigma^2)$, $\sigma^2 > 1$ added to the $N(0, 1)$. Together, a total of twenty three non-normal distributions were used in the investigation. All data were generated using the SAS (2013) system on a Sun Fire X4600 M2 x64 server: 8 x AMD Opteron Model 8220 processor (2.8GHz-dual-core).

A program used to generate the three groups with $g=4$ and $h=0$ data per group is presented as Program 1 in the Appendix. We set the sample size at 60 observations per group and adopted familywise significance levels of .15 and .20 adopting Hochberg's (1988) sequentially-rejective Bonferroni procedure to assess the A-D goodness-of-fit test. We chose this sample size because Keselman *et al.* (2014) in their study had samples of size 80 per group and we wanted to see if equally good results could be obtained with a smaller sample size per group. Lastly, we defined power as (1) the probability of detecting at least one non-normal distribution, (2) the probability of detecting all three non-normal groups of data, and (3) the average probability of detecting non-normal data averaged across the three groups.

Hochberg's step-up Bonferroni procedure is another example of a stepwise Bonferroni method of Type I error control and hence can also be better than the usual Dunn-Bonferroni method. In this procedure, the p -values corresponding to the m statistics for testing the hypotheses $H_{(1)}, \dots, H_{(m)}$ are ordered from smallest to largest $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$. Then, for any $i = m, m - 1, \dots, 1$, if $p_{(i)} \leq \alpha/(m - i + 1)$, the Hochberg procedure rejects all $H_{(i')}$ ($i' \leq i$). According to this procedure one begins by assessing the largest p -value, $p_{(m)}$. If $p_{(m)} \leq \alpha$ all hypotheses are rejected. If $p_{(i)} > \alpha$, then $H_{(m)}$ is accepted and one proceeds to compare $p_{(m-1)} \leq \alpha/2$. If $p_{(m-1)} \leq \alpha/2$, then all $H_{(i)}$ ($i = m - 1, \dots, 1$) are rejected; if not, then $H_{(m-1)}$ is accepted and one proceeds to compare $p_{(m-2)}$ with $\alpha/3$, and so on. The program enabling power to be defined as the probability of at least one group detected to be non-normal using the Hochberg procedure is available in the Appendix as Program 2.

3. RESULTS

Table 1 contains power (and Type I error) rates for the A-D test for numerous non-normal distributions which vary by degree of skewness and kurtosis. The proportion of empirical power values $\geq .80$ was .64 when power was defined as the probability of detecting at least one non-normal set of data in the three groups. The proportion was .29 and .43 when power was defined as the probability of detecting all three non-normal data sets or the average probability of rejecting over the three groups, respectively. The rates did not differ due to level of significance. Also notable from Table 1 was that the A-D test was more sensitive to detect non-normality when data was both skewed and kurtotic. Keselman *et al.* (2014) found larger proportions to detect non-normal data based on group sizes of 80 per group.

Assessing Normality: Applications in Multi-Group Designs

TABLE 1: Power rates for the Anderson-Darling test adopting familywise Type I error control.

Distribution	Skewness	Kurtosis	$\alpha = .15$	$\alpha = .20$
At Least One Rejection				
Normal*	0.00	0.00	0.1436	0.1796
$g=0, h=.075$	0.00	1.49	0.4692	0.5196
$g=0, h=.1$	0.00	2.51	0.6084	0.6630
$g=0, h=.125$	0.00	4.16	0.7402	0.7806
$g=0, h=.15$	0.00	7.17	0.8414	0.8666
$g=0, h=.2$	0.00	33.22	0.9468	0.9566
$g=.2, h=0$	0.61	0.68	0.6450	0.6962
$g=.4, h=0$	1.32	3.26	0.9918	0.9954
$g=.6, h=0$	2.26	10.27	1.0000	1.0000
$g=1, h=0$	6.19	110.94	1.0000	1.0000
$g=.2, h=.1$	1.08	5.50	0.8674	0.8922
$g=.4, h=.1$	2.45	20.30	0.9938	0.9952
$g=.6, h=.1$	4.69	89.80	1.0000	1.0000
$g=.8, h=.1$	9.27	603.61	1.0000	1.0000
Rejecting all Three				
Normal*	0.00	0.00	0.1436	0.1796
$g=0, h=.075$	0.00	1.49	0.0404	0.0670
$g=0, h=.1$	0.00	2.51	0.0882	0.1194
$g=0, h=.125$	0.00	4.16	0.1510	0.2002
$g=0, h=.15$	0.00	7.17	0.2392	0.2894
$g=0, h=.2$	0.00	33.22	0.4270	0.4814
$g=.2, h=0$	0.61	0.68	0.1034	0.1464
$g=.4, h=0$	1.32	3.26	0.7338	0.7820
$g=.6, h=0$	2.26	10.27	0.9832	0.9880
$g=1, h=0$	6.19	110.94	1.0000	1.0000
$g=.2, h=.1$	1.08	5.50	0.2598	0.3112
$g=.4, h=.1$	2.45	20.30	0.7360	0.7728
$g=.6, h=.1$	4.69	89.80	0.9658	0.9748
$g=.8, h=.1$	9.27	603.61	0.9978	0.9982
Average Probability of Rejecting				
Normal*	0.00	0.00	0.0544	0.0741
$g=0, h=.075$	0.00	1.49	0.2156	0.2619
$g=0, h=.1$	0.00	2.51	0.3157	0.3677
$g=0, h=.125$	0.00	4.16	0.4235	0.4797
$g=0, h=.15$	0.00	7.17	0.5335	0.5837
$g=0, h=.2$	0.00	33.22	0.7095	0.7462
$g=.2, h=0$	0.61	0.68	0.3435	0.4025
$g=.4, h=0$	1.32	3.26	0.8859	0.9100
$g=.6, h=0$	2.26	10.27	0.9943	0.9959
$g=1, h=0$	6.19	110.94	1.0000	1.0000
$g=.2, h=.1$	1.08	5.50	0.5619	0.6137
$g=.4, h=.1$	2.45	20.30	0.8916	0.9095
$g=.6, h=.1$	4.69	89.80	0.9885	0.9916
$g=.8, h=.1$	9.27	603.61	0.9993	0.9994

Note: *Normal data ($g = 0, h = 0$); thus Type I error rates.

Table 2 presents the empirical rejection rates for the non-normal contaminated mixed-normal distributions examined in the study. The proportion of rejections for the three definitions of power (at least one, all three, and average power rates) were .83, .17, and .50, respectively. Proportion of rejections did not differ across levels of significance.

TABLE 2: Power of the Anderson-Darling goodness-of-fit test with data that were obtained from contaminated mixed-normal distributions.

At Least One Rejection				
Distribution	Distance	Number	$\alpha = .15$	$\alpha = .20$
(.983) $N(0,1)$ + (.017) $N(0,25)$	5	1	0.7228	0.7478
(.967) $N(0,1)$ + (.033) $N(0,25)$	5	2	0.9224	0.9348
(.933) $N(0,1)$ + (.067) $N(0,25)$	5	4	0.9930	0.9944
(.983) $N(0,1)$ + (.017) $N(0,100)$	10	1	0.9480	0.9552
(.967) $N(0,1)$ + (.033) $N(0,100)$	10	2	0.9972	0.9974
(.933) $N(0,1)$ + (.067) $N(0,100)$	10	4	1.0000	1.0000
Rejecting All Three				
Distribution	Distance	Number	$\alpha = .15$	$\alpha = .20$
(.983) $N(0,1)$ + (.017) $N(0,25)$	5	1	0.1086	0.1368
(.967) $N(0,1)$ + (.033) $N(0,25)$	5	2	0.2980	0.3408
(.933) $N(0,1)$ + (.067) $N(0,25)$	5	4	0.6714	0.7066
(.983) $N(0,1)$ + (.017) $N(0,100)$	10	1	0.3448	0.3862
(.967) $N(0,1)$ + (.033) $N(0,100)$	10	2	0.7260	0.7480
(.933) $N(0,1)$ + (.067) $N(0,100)$	10	4	0.9648	0.9686
Average Probability of Rejecting				
Distribution	Distance	Number	$\alpha = .15$	$\alpha = .20$
(.983) $N(0,1)$ + (.017) $N(0,25)$	5	1	0.3881	0.4230
(.967) $N(0,1)$ + (.033) $N(0,25)$	5	2	0.6229	0.6561
(.933) $N(0,1)$ + (.067) $N(0,25)$	5	4	0.8660	0.8827
(.983) $N(0,1)$ + (.017) $N(0,100)$	10	1	0.6691	0.6973
(.967) $N(0,1)$ + (.033) $N(0,100)$	10	2	0.8925	0.9023
(.933) $N(0,1)$ + (.067) $N(0,100)$	10	4	0.9879	0.9893

Note: Distance is measured in standard deviations. Number represents the number of outliers present.

We do not table our rejection rates for the five cases of Likert-type non-normal data that were investigated (See Table 3 for a description of these non-normal distributions) since all proportions equaled 1.00.

TABLE 3: Multinomial distributions based upon Schoder, *et al.*'s (2006) probabilities simulated as Likert scales

Description	$(p_1, p_2, p_3, p_4, p_5)$
Even	(.2, .2, .2, .2, .2)
Symmetric	(.1, .2, .4, .2, .1)
Moderate skew	(.5, .3, .15, .04, .01)
Heavy skew	(.7, .2, .06, .03, .01)

4. DISCUSSION

Based on the results presented in this paper and those reported by Keselman *et al.* (2013, 2014) it appears that the Anderson-Darling goodness-of-fit statistic is moderately effective in locating non-normal distributions. When power is defined as the probability of detecting at least one case of non-normality among the groups being examined, the A-D rejection rate is reasonably high (i.e., $\geq .80$) in most cases examined in this study. Researchers only need to detect one normal group of data to conclude that this requirement is not satisfied. However, when a statistical procedure allows researchers to model different types of data shapes per group, e.g., SAS's PROC GLIMMIX, then detecting all non-normal data sets would be important and the rejection rates for detecting all non-normal groups is not particularly good, though the average rejection rates were reasonable. Keselman *et al.*'s (2014) findings would lead one to believe that the 60 observations per group that we examined should be increased to 80. However, another very beneficial aspect of these tests is that they can be used to test the fit of other distributions. The SAS system allows users to determine if the data is best fit by data modeled with responses distributed as exponential, gamma, inverse Gaussian, lognormal and tcentral (i.e., the non-central t distribution). Thus users can use the Akaike (1974) statistic to determine which distribution best fits the data. Major statistical packages enable the applied research to compute these statistics (e.g., See Muller and Fetterman (2002)).

As a postscript we believe it is important to enumerate choices researchers could or should not follow after concluding that their data do not conform to normality. Many researchers routinely adopt non-parametric methods when data are presumed to be non-normal to compare treatment groups. This is a strategy that is not supported by empirical research. That is, it is quite well known that non-parametric tests are also substantially affected when data are non-normal; furthermore, the deleterious effects of non-normality are exacerbated when groups have unequal variances (Zimmerman (2004, 2010)). Another technique is to use a test statistic, and sample estimators, which are not negatively biased when data are non-normal (and when variances are unequal). A great deal of evidence has been published indicating the a Welch-type statistic that does not pool over heterogeneous variances and uses robust estimators such as trimmed means and Winsorized variances provides good protection against inflated rates of Type I error and/or decreases in power (See Keselman *et al.* (2008a, b)).

Indeed, according to many this approach is most reasonable (See Hampel *et al.* (1986); Huber (1981); Keselman *et al.* (2008a, b)). That is, as Marazzi and Ruffieux (1999) note “the (usual) mean is a difficult parameter to estimate well: the sample mean, which is the natural estimate, is very nonrobust.” (p. 79). And another very reasonable approach to adopt is to use a statistical procedure that allows in its estimation process for data to assume other shapes from the normal. SAS’s (2012) GLIMMIX procedure is a generalized linear model approach to data analysis that does just this (See Breslow and Clayton (1993); McCullah and Nelder (1989); Nelder and Wedderburn (1972)). One really nice aspect to this method of analysis is that researchers can even specify different distributional shapes across treatment groups. However, to date, empirical findings are not positive regarding using preliminary analyses prior to adopting a generalized linear model statistic (See Keselman *et al.*, (under review)). Consequently, faced with non-normal data these authors would choose to use the Welch-James statistic to test for treatment group mean equality.

REFERENCES

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*. **AC-19**: 716–723.
- Breslow, N. R., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*. **88**: 9-25.
- D’Agostino, R. B., and Stephens, M. I. A. (1986). (Editors). *Goodness-of-Fit Techniques*. New York: Marcel Dekker.
- Hampel, F., R., Ronchetti, E., M., Rousseuw, P., J. and Stahel, W., A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.
- Headrick, T., C., Kowalchuk, R., K. and Sheng, Y. (2008). Parametric probability densities and distribution functions for Tukey g-and-h transformations and their use for fitting data. *Applied Mathematical Sciences*. **2**(9): 449-462.
- Hoaglin, D., C. (1983). *G-and-h distributions*, in *Encyclopedia of Statistical Sciences*, Vol. 3, eds. S. Kotz, and N. L. Johnson (Wiley, New York), p. 298-301.

- Hoaglin, D., C. (1985). *Summarizing shape numerically; The g-and h-distributions*, in Exploring Data, Tables, Trends, and Shapes, eds. D. C. Hoaglin, F. Mosteller, and J. W. Tukey (Wiley, New York), p. 461-511.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. **75**: 800–803.
- Huber, P., J. (1981). *Robust Statistics*. New York: Wiley.
- Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., and Deering, K. (2008a). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*. **13**: 110-129.
- Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., and Deering, K. (2008b). Supplemental materials to Keselman *et al.* (2008). A SAS program to implement a general approximate degrees of freedom solution for inference and estimation. <http://dx.doi.org/10.1037/1082-989X.13.2.110.supp>.
- Keselman, H. J., Othman, A. R., and Wilcox, R. R. (2013). Preliminary testing for normality: Is it a good practice? *Journal of Modern Applied Statistical Methods*. **12**(2): 2-19.
- Keselman, H. J., Othman, A. R., & Wilcox, R. R. (2014). Testing for normality in the multi-group problem: Is this a good practice? *Clinical Dermatology*. **2**(1): 29-43.
- Keselman, H., J., Othman, A., R. and Wilcox, R., R. (under review). Generalized linear model analyses for treatment group equality when data are non-normal. *Psychological Methods*.
- Kowalchuk, R., K. & Headrick, T., C. (2010). Simulating multivariate g-and-h distributions. *British Journal of Mathematical and Statistical Psychology*. **63**: 63-74.
- Marazzi, A., and Ruffieux, C. (1999). The truncated mean of an asymmetric distribution. *Computational Statistics & Data Analysis* . **32**: 79-100.

- McCullagh, P. and Nelder, J., A. (1989). *Generalized Linear Models*. 2nd ed. London, England: Chapman Hall.
- Muller, K., E. and Fetterman, B., A. (2002). *Regression and ANOVA: An Integrated Approach Using SAS Software*. Cary, NC: SAS Institute, Inc.
- Nelder, J., A., and Wedderburn, W., M. (1972). Generalised linear models. *Journal of the Royal Statistical Society, Series A*. **135**: 370-384.
- SAS Institute. (2012). *Statistics: ANOVA and Regression*. Cary, NC: SAS Institute, Inc.
- SAS Institute. (2013). *SAS 9.3 for Unix and Linux*. Cary, NC: SAS Institute, Inc.
- Schoder, V., Himmelmann, A., and Wilhelm, K., P. (2006). Preliminary testing for normality: Some statistical aspects of a common concept. *Clinical Dermatology*. **31**: 757-761.
- Shapiro, S., S., and Wilk, M., B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*. **52**: 591-611.
- Warr, R. L. and Erich, R. A. (2013). Should the interquartile range divided by the standard deviation be used to assess normality? *American Statistician*. **67**(4): 242-244.
- Wilcox, R., R. (2012). *Introduction to Robust Estimation and Hypothesis Testing*. 3rd ed. San Diego, CA: Academic Press.
- Zimmerman, D., W. (2004). A note on the influence of outliers on parametric and nonparametric tests. *Journal of General Psychology*. **121**: 391-401.
- Zimmerman, D., W. (2010). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education*. **67**: 55-68.

Appendix

PROGRAM 1: Generating 5000 $g=.4, h=0$ distributions in three groups.

```
options nocenter nonumber nodate ps=60;

*Creating 5000 datasets of (g=.4,h=0) random variates of size 60;
data random;
  call streaminit(439839383);
  g = .4;
  mu=(exp((g**2)/2)-1)/g;
  sigmasq = exp(g**2)*(exp((g**2)/2)-1)/g**2;
  do dataset=1 to 5000;
    do group = 1 to 3;
      do j = 1 to 60;
        z=rand('normal');
        y=(exp(g*z)-1)/g;
        x=(y-mu)/sqrt(sigmasq); *Standardizing the g distribution;
        output;
      end;
    end;
  end;
  drop g mu sigmasq z y;
run;

*Running the goodness of fit tests on the 5000 data sets;
ods select GoodnessOFFit;
proc univariate data=random;
  var x;
  class dataset group;
  histogram / normal;
run;
```

PROGRAM 2: Extracting power of non-normality with at least one group being not normal using $g=4$, $h=0$ distribution

```
options center;

data one;
  infile 'E:\aro pj\ PC\aro_from_c\research\canada2013\power\3-
  groups\ss60\skewed\data\gof_norm_g.4h.1_5k_s60_3grp.lst';
  input test $ 1-18 pvalue $57-62;
  if test not in ('Kolmogorov-Smirnov', 'Cramer-von Mises',
  'Anderson-Darling') then delete;
  if pvalue =: '<' then pvalue = substr(pvalue,2) - 0.001;
  else if pvalue =: '>' then pvalue = substr(pvalue,2) + 0.001;
run;

data two;
  set one;
  if test ne 'Anderson-Darling' then delete;
run;

data three;
  set two;
  if mod(_n_, 3) = 1 then group = 1;
  else if mod(_n_, 3) = 2 then group = 2;
  else group = 3;
run;

data four;
  set three;
  if group ne 1 then delete;
  pvalue1 = pvalue;
  drop pvalue group;
run;

data five;
  set three;
  if group ne 2 then delete;
  pvalue2 = pvalue;
  drop pvalue group;
run;

data six;
  set three;
  if group ne 3 then delete;
  pvalue3 = pvalue;
  drop pvalue group;
run;

data seven;
  merge four five six;
```

PROGRAM 2 (continued) : Extracting power of non-normality with at least one group being not normal using $g=4$, $h=0$ distribution

```

p3 = max(pvalue1, pvalue2, pvalue3);
p1 = min(pvalue1, pvalue2, pvalue3);
if p1=p3 then p2=p3;
if pvalue1 ne p1 and pvalue1 ne p3 then p2 = pvalue1;
if pvalue2 ne p1 and pvalue2 ne p3 then p2 = pvalue2;
if pvalue3 ne p1 and pvalue3 ne p3 then p2 = pvalue3;
if pvalue1=p1 and pvalue2=p1 then p2=p1;
if pvalue1=p1 and pvalue3=p1 then p2=p1;
if pvalue2=p1 and pvalue3=p1 then p2=p1;
if pvalue1=p3 and pvalue2=p3 then p2=p3;
if pvalue1=p3 and pvalue3=p3 then p2=p3;
if pvalue2=p3 and pvalue3=p3 then p2=p3;
run;

data eight;
set seven;
if p3 <= 0.05 then cpval05 = 1;
else if p2 <= 0.025 then cpval05 = 1;
else if p1 <= 0.05/3 then cpval05 = 1;
else cpval05 = 0;
if p3 <= 0.10 then cpval10 = 1;
else if p2 <= 0.05 then cpval10 = 1;
else if p1 <= 0.10/3 then cpval10 = 1;
else cpval10 = 0;
if p3 <= 0.15 then cpval15 = 1;
else if p2 <= 0.075 then cpval15 = 1;
else if p1 <= 0.05 then cpval15 = 1;
else cpval15 = 0;
if p3 <= 0.20 then cpval20 = 1;
else if p2 <= 0.10 then cpval20 = 1;
else if p1 <= 0.20/3 then cpval20 = 1;
else cpval20 = 0;
run;

proc freq;
title 'Hocheberg Monte Carlo 1 minus pvalue of the Anderson-
Darling Test';
title2 'H0: Groups 1 2 and 3 are normally distributed';
title3 'H1: At least one group is not normally distributed';
tables cpval05 cpval10 cpval15 cpval20;
run;

```