International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 1613-1618 www.ijaem.net ISSN: 2395-5252

Robustness and Asymptotic Theory of Tests of Location under Violation of Normality Assumption Based on Infectious Diseases Data (a Study of Borno and Yobe States)

Ali Aji, Ibrahim Sani, Baba Alhaji Umar
Department of StatisticMai Idris Alooma Polytechnic Geidam, Yobe State, Nigeria.

Submitted: 15-05-2021 Revised: 26-05-2021 Accepted: 28-05-2021

ABSTRACT: Borno and Yobe States are among the states that are have affected by various infectious diseases based on the analysis of the data obtained in this study, one can safely conclude that the Z – test is the smallest, most powerful and robust p – values among all the test especially for infectious diseases data that occur due to some change in atmospheres or environmental factors.

Keywords:robustness, asymptotic, infectious, diseases, data

I. INTRODUCTION

World health organisation report (2003) Infectious diseases remain key agents of the debilitating poverty afflicting so much of the world today. Each year these diseases kill almost 9 million people, many of them children under five, and they also cause enormous burdens through life-long disability. Stepping up research into their causes and how to effectively treat them and prevent them from spreading could have an enormous impact on efforts to lift people out of poverty and to build a better world for future generations.

Most emerging infectious disease pathogens in humans cross from their natural zoonoticreservoir to human populations where early mutated, reasserted or recombined forms begin tospread from person-to-person [Antia et al. (2003)]. Examples include humanimmunodeficiency virus, monkey pox, severe acute respiratory syndrome and pandemic influenza. Currently, a highly pathogenic avian influenza strain (H5N1) has been spreadingfrom poultry to humans, mostly in Southeast Asia, with possible limited human-to-humanspread through close contact in Indonesia [Butler (2006)]. A concern is that this virus couldcause a large scale pandemic as it becomes more adapted to human-to-human transmission.Real-time surveillance provides limited information on small clusters of human cases in termsof symptom onset times and physical location. It is critical to answer two questions in real time: 1. Is the infectious agent spreading from person to person? and 2. If it is, how transmissible is it? The first question is novel and, to our knowledge, has not been addressed in the statistical literature. The second question is an estimation problem, and various statistical methods using household data are applicable, such as the models based on observed final infection status [Longini and Koopman (1982), Becker and Hasofer (1997), and Halloran (2006). The statistical questions hinge on inference about the transmissibility of the infectious agent. The basic reproductive number, R0, is the fundamental measure of the transmissibility of an emerging infectious agent. Given that the emerging infectious agent is transmissible, estimates of R0 will generally be small and are not very informative.

Although statisticians have discussed asymptotic tests for a limited set of scenarios [Feng and McCulloch (1992)], more often such an asymptotic null distribution is not available for a specific case. Furthermore, the validity of asymptotic tests depends on relatively large sample sizes, which may compromise the power of such tests to detect person-to-person transmission if applied to a small sample size, such as those generated by avian influenza. These challenges motivate our investigation in exact rather than asymptotic testing methods. those based on a discrete-time sequence of symptom onset [Rampey et al. (1992), Yang, Longini and Halloran (2006)].

Robust statistics assesses the changes in estimates due to small changes in basic assumptions. Olive (2005) defined robust statistics as a method that is designed to perform well when the shape of the true underlying model deviates slightly from the assumed parametric model, such as assumptions of normality. Robust statistical methods therefore, have been developed for many

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 1613-1618 www.ijaem.net ISSN: 2395-5252

common problems, such as estimating location, scale, and regression parameters (Stigler, 2010; Bosse et al, 2016). One motivation is to produce statistical methods that are not unduly affected by the violation of the normality assumption; another motivation is to provide methods with good performance when there are small departures from parametric.

The beginning of robust statistics dates back to the eighteenth century, when the first rule for rejection of outliers was developed (Filzmoser and Rousseeuw, 2005). statistics, asymptotic theory is a framework for assessing properties of estimators and statistical tests. Where, it is assumed that as the sample size n increases, the properties of estimators and tests can be evaluated in the limit as $n \rightarrow \infty$. In practice, limit evaluation is treated as being approximately valid for large finite sample sizes. Most statistical problems begin with a dataset of size n. The asymptotic theory proceeds by assuming that it is possible (in principle) to keep collecting additional data, so that the sample size grows infinitely, i.e. $n \rightarrow \infty$. Under this assumption, many results can be obtained that are unavailable for samples of finite size, for example the lawoflargenumbers which states that for a sequence of independentand identically distributed (IID) random variables $X_1, X_2, \ldots$, if one value is drawn from each random variable and the average of the first n values is computed as X_n, then the X_n converge in probability to the population mean $E[X_i]$ as $n \rightarrow \infty$ *(Balakrishnan et al, 2001).

A test could be robust to data that violate normality assumption or data that contain outlier.Outliers are observations that stand too different from others in a set of observations. When present in a data set, they affect both descriptive and inferential statistics (Kayode et al, 2016). This study therefore, studies the asymptotic and robustness of one sample test statistics to outliers and non-normality so as to know the appropriate one to test hypothesis about the population parameter when outliers are present in a particular distribution family. The study examines the robustness and asymptotic property of four tests (t, z, sign and Wilcoxon tests), because they are the commonly used test for one sample location to examine which of them is more robust to non-normal data and data containing outliers. It is well known that classical tests for comparing location like means and medians are very sensitive to departures from normality, therefore it is considered in this study, some hypothesis tests in situations where the data come from a probability distribution whose underlying distribution may be normal or non-normal (e.g. uniform, exponential, Gamma), with and without outliers and different sample sizes are considered for each scenario of data set from each diseases.

Infectious diseases emerging throughout history have included some of the most feared plagues of the past. New infections continue to emerge today, while many of the old plagues are with us still. These are global problems (William Foege, former CDC director, terms them “global infectious disease threats”). As demonstrated by influenza epidemics, under suitable circumstances, a new infection first appearing anywhere in the world could traverse entire continents within days or weeks. We can define as “emerging” infections that have newly appeared in the population, or have existed but are rapidly increasing in incidence or geographic range (Morse S.S,&Schluederberg A,1990). Recent examples of emerging diseases in various parts of the world include HIV/AIDS; classic cholera in South America and Africa; cholera due to Vibrio choleraeO139; Rift Valley fever; hantavirus pulmonary syndrome; Lyme disease; and hemolytic uremic syndrome, a foodborne infection caused by certain strains of Escherichia coli (in the United States, serotype O157:H7).

Robustness statistics assesses the changes in estimates due to small changes in the basic assumptions and to create new estimates that are insensitive to small changes in some of the assumptions.In statistics it is conventional to assume that observations are normally distributed. The entire statistical framework is based on this assumption and if this assumption is violated the inference breaks down. For this reason, it is essential to check or test whether this assumption hold before any statistical analysis of data. Many researchers do not recognize the importance of test of assumption before it is used. For instance, t-test has some assumptions (normality, homoscedasticity and continuity of the data set). If any of these assumptions are violated, may lead to its insufficient and a more robust (good) test can be used to obtain valid result. In all branches of knowledge, it is necessary to apply statistical methods in a sensible way. The most commonly used statistical methods are correlation, regression and experimental design. But all of them are based on one basic assumption, that the observation follows normal (Gaussian) distribution (Das and Imon, 2016). So, it is assumed that the population from where the sample is drawn is normally distributed. For this reason, the inferential methods require checking the normality assumption. The purpose of this work is to provide an asymptotic

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 1613-1618 www.ijaem.net ISSN: 2395-5252

theory of robustness to non-normal which shows when the various selection procedures are robust as the sample sizes increase.

II. MATERIAL METHOD

this study investigating the robustness of one sample test statistics to non-normal (skewed distributions) and outliers using both power and type I error as criteria. This will enable us to know the appropriate test for testing hypothesis about the population mean when infectious diseases data are non-normal and or contain outliers. The tests considered in this study are t-test and z-test under parametric test, while Wilcoxon signed rank and sign test under the nonparametric test through procedures from normal, uniform, exponential and gamma distributions.

III. TABLE AND DISCUSSION

The analysis of 54 weeks of four selected infectious diseases data COVID19, Measles, Meningitis and Cholera of Borno and Yobe state for the four tests z, t, sign and Wilcoxon sign rank test under deference sample sizes small, medium and large which are 15, 25 and >30 with extreme values (outlies) and their p – values are recorded below

Table 1. COVID19 P – values of the four tests for various sample sizes

Test Sample size	Z	t	Sign	Wilcoxon
15	$2.2e^{-16}$	0.014	$1.2e^{-04}$	0.001
25	$2.2e^{-16}$	$2.2e^{-05}$	$1.192e^{-07}$	$1.933e^{-05}$
>30	$2.2e^{-16}$	$1.628e^{-08}$	$1.776e^{-15}$	$7.759e^{-10}$

The table1 shows that the p – values of the four tests for all the sample size small, medium and large Z < sign < t < Wilcoxon.

Table2. Measles P – values of the four tests for various sample sizes

Test Sample size	Z	t	Sign	Wilcoxon
15	$2.2e^{-16}$	$7.593e^{-04}$	$6.104e^{-05}$	$7.247e^{-04}$
25	$2.2e^{-16}$	$1.133e^{-04}$	$5.96e^{-08}$	$1.3e^{-05}$
>30	$2.2e^{-16}$	$4.137e^{-08}$	$2.22e^{-16}$	$2.444e^{-10}$

The table2 shows that the p – values of the four tests for all the sample size small, medium and large which is 15, 25, and >30 the test is Z < sign < t < Wilcoxon.

Table3. Meningitis P – values of the four tests for various sample sizes

Test Sample size	Z	t	Sign	Wilcoxon
15	$2.2e^{-16}$	0.2131	$3.906e^{-03}$	$7.914e^{-03}$
25	$2.2e^{-16}$	0.1417	$3.052e^{-05}$	$3.931e^{-04}$
>30	$2.2e^{-16}$	0.07709	$2.98e^{-08}$	$5.973e^{-06}$

The above table3 also shows that the p – values of the four tests for all the sample size small, medium and large which is 15, 25, and >30 the test is Z < sign < t < Wilcoxon.

Table4. Cholera P – values of the four tests for various sample sizes

Test Sample size	Z	t	Sign	Wilcoxon
15	0.8614	0.0701	$4.883e^{-03}$	$2.516e^{-03}$
25	0.1247	0.03109	$3.815e^{-06}$	$1.419e^{-04}$
>30	0.0329	0.0474	$1.164e^{-10}$	$3.802e^{-07}$

Table4 shows that the p – values of the four tests for the sample size of 15 is Z > sign > t > Wilcoxon, for the sample size of 25 is Z > t > Wilcoxon > sign and for sample size of >30 is t > Z > Wilcoxon > Sign.

P – values of the four test with various sample size after removing extreme values (outlies) from the data set for the four selected diseases

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 1613-1618 www.ijaem.net ISSN: 2395-5252

Table 5 COVID19 p – values of the four tests with various sample sizes after removing outlies

Test Sample size	COVID19 Z	t	WilcoxonSign	
15	$2.2e^{-16}$	$2.902e^{-04}$	$1.221e^{-04}$	$1.088e^{-03}$
25	$2.2e^{-16}$	$6.014e^{-06}$	$1.92e^{-07}$	$1.933e^{-05}$
>30	$2.2e^{-16}$	$1.059e^{-09}$	$2.276e^{-13}$	$1.155e^{-08}$

The table 5 above shows that the p – values for the sample sizes of 15 is Z < Sign < Wilcoxon < t, for sample sizes 25 and > 30 is Z < Sign < t < Wilcoxon.

Table 6 Measles p – values of the four tests with various sample sizes after removing outlies

Test Sample size	Measles Z t	Sign	Wilcoxon	
15	$2.2e^{-16}$	3.893^{-05}	$6.104e^{-05}$	$7.211e^{-04}$
25	$2.2e^{-16}$	$4.139e^{-06}$	$5.96e^{-08}$	$1.296e^{-05}$
>30	$2.2e^{-16}$	$3.896e^{-11}$	$1.144e^{-13}$	$7.832e^{-09}$

Table 6 above shows that the p – values for the sample sizes of 15 is Z < t < Wilcoxon < Sign for sample sizes 25 and > 30 is Z < Sign < t < Wilcoxon.

Table 7 Meningitis p – values of the four tests with various sample sizes after removing outlies

Test Sample size	Meningitis Z tSign	Wilcoxon		
15	$7.891e^{-04}$	$2.54e^{-03}$	$7.812e^{-03}$	0.01154
25	$6.334e^{-05}$	$1.095e^{-04}$	$2.441e^{-04}$	$1.212e^{-03}$
>30	$2.43e^{-05}$	$2.621e^{-06}$	$1.907e^{-06}$	$5.497e^{-05}$

The table 7 above shows that the p – values for the sample sizes of 15 is Z < t < Sign < Wilcoxon for sample sizes 25 and > 30 is Sign < t < Z < Wilcoxon.

Table 8 Cholera p – values of the four tests with various sample sizes after removing outlies

Test Sample size	Cholera Z	t Sign	Wilcoxon	
15	0.8614	0.0701	$4.883e^{-04}$	$2.516e^{-03}$
25	0.7374	0.1185	$1.52e^{-05}$	$3.198e^{-04}$
>30	0.8005	0.1177	$3.725e^{-09}$	$2.695e^{-05}$

Table 8 above shows that the p – values all sample sizes of 15, 25, > 30 is Z > t > Wilcoxon > Sign.

IV. SUMMARY OF THE MAJOR FINDINGS:

The study reveal that;

1. The t – test has the highest p – values which is not powerful (not robust) which is very weak for all the sample size especially for the data with extreme values (outlies) for the diseases of COVID19, Measles, Meningitis and weak for Cholera.
2. Wilcoxon Sign rank test has the highest p – values after t – test which is weak and not powerful (robust) for all the sample size especially for the data set with extreme values (outlies)
3. Sign – test has the small p – values compare with and Wilcoxon Sign rank test with almost all the data set with and without extreme values for difference sample sizes
4. Z – test has the smallest p – values compare with all of the above tests t, Sign and Wilcoxon sign rank test with almost all the sample sizes for various sample sizes of the three diseases of COVID19, Measles and Meningitis. It is the most powerful and robust except for cholera diseases data set.

V. CONCLUSION:

From the analysis of the data obtained in this study, one can safely conclude that the Z – test is the smallest, most powerful and robust p – values among all the test especially for infectious diseases data that occur due to some change in atmospheres or environmental factors

Recommendation: On basis of the result of this study the researcher made this recommendation that any data set that has the characteristics of

infectious diseases especially with extreme values (outlies) under difference sample sizes a robust and powerful test for it is Z – test.

Acknowledgment: This research was Supported and Funded by the Tertiary Education Trust Fund (TETFund) Nigeria

REFERENCES

[1]. Morse S.S and Schluederberg A (1990) Emerging viruses: the evolution of viruses and viral diseases. J Infect Dis;162: page1-7.

[2]. Antia R, Regoes R.R, Koella J.C, Bergstrom C. (2003) The role of evolution in the emergence of infectious diseases. Nature; 426:658–661. [PubMed: 14668863]

[3]. Becker N.G, Hasofer A.M. (1997) Estimation in epidemics with incomplete observations. J Roy Statist SocSerB; 59:415–429.MR1440589

[4]. Butler D. (2006) Family tragedy spotlights flu mutations. Nature; 442:114–115. [PubMed: 16837983]

[5]. Cauchemez S, Boëlle P, Thomas G, ValleronA. (2006) Estimating in real time the efficacy of measures to control emerging communicable diseases. American J Epidemiology; 164:591–597.

[6]. Feng Z.D and McCulloch C.E. (1992) Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. Statist Probab Lett; 13:325–332. MR1160755

[7]. O'Neill P, and Roberts G.O. (1999) Bayesian inference for partially observed stochastic epidemics. J Roy Statist SocSer A; 162:121–129.

[8]. Rampey A.H, Longini I.M, Haber M.J, Monto A. (1992) S. A discrete-time model for the statistical analysis ofinfectious disease incidence data. Biometrics; 48:117–128. [PubMed: 1316178]

[9]. Wallinga J, Teunis P. (2004) Different epidemic curves for severe acute respiratory syndrome reveal similarimpacts of control measures. American J Epidemiology; 160:509–516.

[10]. Akeyede, I. Usman M. and Chiawa M. A (2014). On Consistency and Limitation of Paired t-test, and Wilcoxon Signed Rank test. IOSR Jornal of Mathematics, IOSR Journal International Organisation of Scientific Research. Vol. 10, Issue 1, Version 4 Page 1-6.

[11]. Balakrishnan, N Ibragimov, I.A V B and Nevzorov, (2001). Asymptotic Methods in Probability and Statistics with Application,Birkhauser.

[12]. Bechtel, M.M. and Schneider, G., (2010). Eliciting substance from hot air, Financial market responses to EU summit decisions on European defense. International Organization, 64(2): 199-223.

[13]. Barnett, V. and Lewis, T. (1994).Outliers in Statistical Data (3rd ed.), Chichester:John Wiley.

[14]. Bosse M., Agamennoni, G. and Gilitschensiki, I. (2016). Robust Estimation and Application in robotics. Foundation and Trends in Robotics, vol. 4, pp. 225-269,

[15]. Bradley, J.V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

[16]. Bradley, J.V. (1980). Nonrobustness in one-sample Z and t tests: A large-scale sampling study. Bulletin of the Psychonomic Society, 15(1), 29-32.

[17]. Bruno D. Zumbo and Martha J. Jennings (2002). The Robustness of Validity and Efficiency of the Related Samples t-test in the Presence of Outliers. Psicologi (2002), 23,415-450

[18]. David J. Olive (2005), Southern Illinois University, Department of Mathematics. Mailcode 4408 Carbondale, IL 62901-4408. Applied Robust Statistics

[19]. Feltovich, N. (2003), Nonparametric Tests of Differences in Medians: Comparison of the Wilcoxon-Mann-Whitney and Robust Rank-Order Tests, Experimental Economics, vol. 6, pp. 273-297.

[20]. Filzmozer P. and Rousseeuw P.J. (2005), Vienna University of Technology, Austria, UniversitaireInstellingAntwerpen, Belgium. Probability and Statistics. Robust Statistics

[21]. Freidlin, B., Gastwirth, J.L. (2000). Should the Median Test Be Retired from General Use? American Statistician, vol. 54, pp. 161-164.

[22]. G. Cicchitelli (1989), On The Robustness of The One Sample t Test. Gordon and Breach Science publishers, Inc. Printed in Great Britain

[23]. Gosset, W. S. (2016). The probable error of a mean. Biometrika-wikipedia (2016), the free encyclopedia, <http://www.encyclopedia.com.

[24]. Hampel F.R. (1974). The influence curve and its role in robust estimation. Journal of the American Statistical Association

69, 383-393. [Introduces the influence curve, an important tool in robust statistics.]

[25]. Haw kins, D. (1980). Identification of outliers, Chapman and Hall, Reading, London.

[26]. Huber, P.J. (1972). The 1972 Wald lecture robust statistics: a review. The Annals Mathematical Statistics 43, 1041-1067.

[27]. Kayode, A., Taiwo, J.A, Gbenga, S.S. (2016). A Study on Sensitivity and Robustness of One Sample Test Statistics to Outliers. Global Journal of Science Frontier Research (F) Volume XVI Issue VI Version 1.

[28]. Keefer, P. and Stasavage, D., (2002). "Checks and balances, private information,and the credibility of monetary commitments." International Organization, 56(4): 751-774.

[29]. Keya Rani Das, A. H. M. RahmatullahImon (2016). A Brief Review of Tests for Normality. American Journal of Theoretical and Applied Statistics.Vol. 5, No. 1, pp. 5-12. doi: 10.11648/j.ajtas.20160501.12

[30]. Mendenhall, W., Wackerly, D. D. and Scheaffer, R. L. (1990). Mathematical Statistics with Applications, PWS-KENT Publishing Company.

[31]. Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures.Psychological Bulletin, 105, 156-166.

[32]. Mohammad, R I. (2018) Sample Size and Its Role in Central Limit Theorem (CLT). Computational and Applied Mathematics Journal.Vol. 4, No. 1, 2018, pp. 1-7.

[33]. Nordås, R. and Davenport, C., (2013). "Fight the youth: Youth bulges and state repression." American Journal of Political Science, 57(4): 926-940.

[34]. Osborne, J. W. and Amy, O. (2004). The power of outliers (and why Researchersshould always check for them). Practical Assessment, Research and Evaluation. 9(6). North Carolina State University.

[35]. Sawilowsky, S.S., & Blair, R.C. (1992). A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. Psychological Bulletin, 3,352-360.

[36]. Stigler, S.M. (2010). The Changing History of Robustness. The American Statistician, 64(4): 227

[37]. Tanizaki, H. (1997), Power Comparison of Non-Parametric Tests: Small-Sample Properties from Monte Carlo Experiments, Journal of Applied Statistics, vol. 24, pp. 603-632.

[38]. Thomas Plümperand EricNeumayer (2016). Department of Socioeconomics, Vienna University of Economics and Business, Department of Geography & Environment, London School of Economics and Political Science (LSE). Robustness Tests and Statistical Inference.

[39]. Zimmerman, D.W. (1998), Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions, Journal of Experimental Education, vol.67, pp. 55-68.

[40]. Zimmerman, D.W. (2000), Statistical Significance Levels of Nonparametric Tests Biased byHeterogeneous Variances of Treatment Groups. Journal of General Psychology, vol.127, pp. 354-364.

[41]. Zimmerman, D.W., &Zumbo, B.D. (1993). Relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), A Handbook forData Analysis in the Behavioral Sciences. Volume 1: Methodological Issues (pp. 481-517). Hillsdale, NJ: Lawrence Erlbaum Associates.