International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2024): 6.623

Effect of Outliers on One Sample t-test

Mintu Kr. Das

Department of Statistics, D.H.S.K. Commerce College, Dibrugarh, Assam, India
Email: *mintu.dhskcc[at]gmail.com*

Abstract: ***The one sample t-test is a parametric test used to determine whether a sample mean differs significantly from the hypothesized value of the population mean. This one sample location test helps us to generalize our findings to the concerned population. The process of generalizing the results from any research study mostly depends on some summary measures of the test procedure. One such measure is the p-value whose smaller values leads to rejection of null hypothesis. The validity of the decision either to reject or accept the null hypothesis is subject to the fulfillment of the underlying assumptions. Sometimes the presence of outlying observations in a data set violates the assumptions leading to inappropriate p-values. In this paper an attempt has been made to illustrate the effect of outliers on one sample t-test for different sample sizes and varying magnitude of the outliers.***

Keywords: outlier, t-test, *p*-value, location test, hypothesis

1. Introduction

The inferential statistics deals with the generalization the findings based on sample data to the concerned population under study. The process of generalization of the findings from any research study mostly depends on some summary measures resulted from the test of hypothesis. One of such measures is *p-value*, which is mostly reported and used to established the statement about a population. An informal definition of *p*-value is that 'it is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value' [5]. In hypothesis testing, we form our decision rule based on comparison of calculated value with the tabulated value of the test statistic. An alternative decision rule is to use the *p*-value, which is easily computed in statistical software. For example, if we prespecify the level of significance as 5%, then *p*-value ≤ 0.05 leads to the rejection of the null hypothesis and vice versa. So, final inference about the study results is drawn only on the basis of *p*-value. We know that the data to be analysed is a random sample from the study population. The sample is used to compute the test statistic, which makes the test statistic a random variable. Now to each test statistic, there is a *p*-value. The *p*-value is a way to standardise the test statistic so that it can be interpreted more easily. Thus *p*-value is too a random variable and a small shift in the *p*-value can alter the whole inference. Outliers often represent a potential reason of violations of assumptions [6]. Sometimes the presence of outlier in a data set leading to erroneous *p*-values. An outlier is any value that is numerically distant from most of the other data points in a set of data. There are different views of defining an outlier. Gumbel defined the outliers as "that value which seem either too large or too small as compared to the rest of the observations" [4]. Grubbs defined outlier as "an observation which appears to deviate markedly from other members of the sample in which it occurs" [3]. Under a true null hypothesis, the *p*-value based on a continuous test statistic has a uniform distribution over the interval [0,1], regardless of the sample size. Whereas under alternative hypothesis, its distribution is affected by both the sample size and true value or range of the true values of the tested parameter [7]. Thus, it will be interesting to study the behaviour of *p*-values obtained from a one sample t-test in the presence of outlier.

Considering a random sample $X_1, X_2, \dots, X_n$ from a normal distribution defined as

$$(x;\ \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] ; -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

with the parameters $\mu\ (\text{mean}\ ) and\ \sigma(s.d.)$, Dixon proposed two types of mixtures: (i) the mean-shift model, where the distribution is a mixture of $N(\mu, \sigma^2)$ and $N(\mu+\lambda, \sigma^2)$ models, and (ii) the variance-inflation model is a mixture of $N(\mu, \sigma^2)$ and $N(\mu, \lambda^2\sigma^2)$ models with $\lambda>1$ [1],[2]. Under the mean-shift model, out of n observations some unknown set of k observations are suspected to have come from a different normal distribution $f(x|\mu+\lambda, \sigma)$, $0<\lambda$ whereas the remaining $(n-k)$ observations are from the distribution $f(x|\mu, \sigma)$.

The one sample t-test is a parametric test used to determine whether a sample mean differs significantly from the hypothesized value of the population mean. In performing this test, we state our hypotheses as:

Null hypothesis (H_0): There is no significant difference between the sample mean and population mean, against the Alternative hypothesis (H_1): There is significant difference. If we pre-specify the level of significance, α=0.05, then our *p*-value < 0.05 will allow us to reject H_0 otherwise not to reject H_0. In this paper we shall examine the behavior of *p*-values under the true H_0.

2. Simulation Study

To perform the Monte Carlo simulations, we are using the open-source software R. We first generate a large number of random samples from a standard normal distribution $N(\mu=0, \sigma^2=1)$. The test was performed under the condition that the null hypothesis is true. For each generated sample the test statistic and p-value are calculated and the *p*-values are compared for the significance level 0.05. If the observed p-value is less than or equal to 0.05 then it leads to the rejection of the null hypothesis. If it is rejected, it is counted and the process get repeated. Thus, we recorded the proportion of rejection of the null hypothesis. We have repeated 10,000 times for different sample sizes $(n = 10, 15, 25)$ and varying magnitude of outliers $(\lambda = 1, 2, 5, 7, 9)$.

3. Results and Discussion

First, we observe the behaviour of p-values for various sample sizes with and without outliers. Figure-1 indicates that the histogram of 10,000 simulated p-values for t-test confirms the Uniform distribution for both n=10 and n=25 sample sizes. This is for the case when μ=0, i.e., the H_0 is true. Whereas in Figure-2, we observed that the distribution of p-values deviates from the Uniformity when $\mu \neq 0$, i.e., the H_0 is false. The departure from Uniform distribution increases with the increase of the value of the hypothesized value of the population mean.

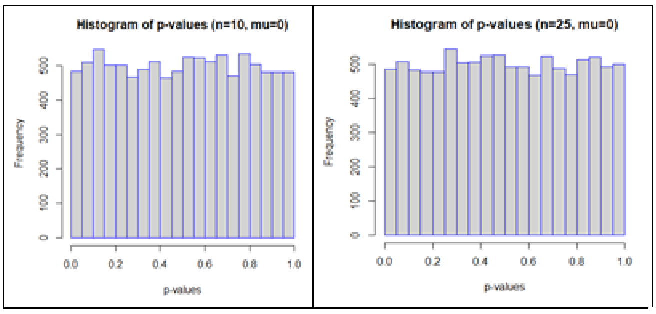

Figure 1: Histogram of p-values when H_0 is true

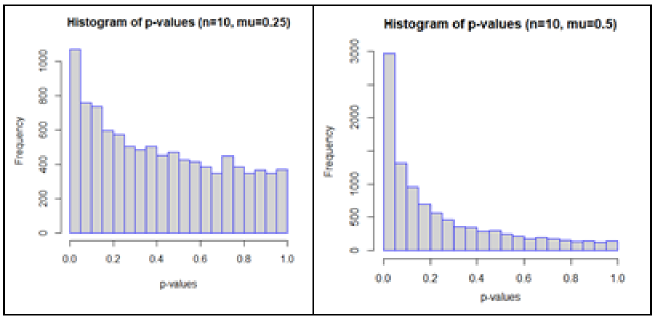

Figure 2: Histogram of p-values when H_0 is false

Then we moved to observed the behaviour of p-values when the sample is contaminated with outlying observation. Here for simplicity, i.e., to avoid the impact of masking or swamping effect of outliers we restrict the value of k to one (01) only. The empirical proportion of rejection of the true H_0 for n=10, 15, 25 has been given in Table-1. Along with the graphs of proportion of rejection for various values of λ shown in Fig-3 we observe the following points:

- For a sample without any outlier (i.e., λ=0) proportion of rejection when the null hypothesis is true is exactly 0.05 which support the fact that at even when H_0 is true for any given experiment, one of twenty p-values could be less than or equal to 0.05.
- When an outlier was added with varying magnitude then the proportion of rejection of the H_0 fluctuates around the value of 0.05 for λ=1,2.
- But as the value of λ increases the proportion of rejection of the H_0 gets decreases for t-test.

Table 1: Proportion of rejection of the true H_0

λ	n =10	n = 15	n = 25
0	0.0504	0.0510	0.0513
1	0.0489	0.0502	0.0510
2	0.0463	0.0492	0.0473
5	0.0158	0.0270	0.0376
7	0.0028	0.0122	0.0223
9	0.0010	0.0031	0.0099

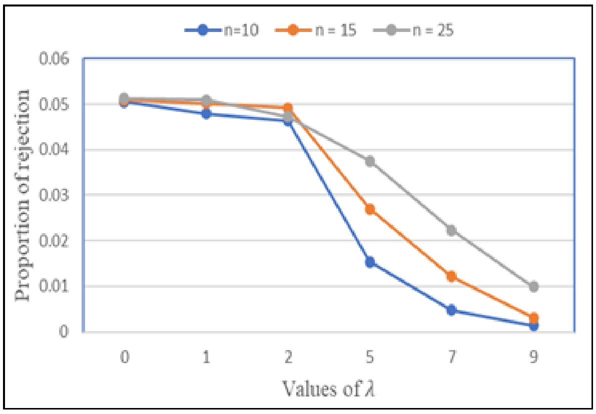

Figure 3: Proportion of Rejection of null hypothesis with outliners

Thus, the presence of outlying observations in a data set can lead to acceptance of the null hypothesis when it is false. In the following section, a numerical example with simulated data has been used to explain the impact of a single outlier on t-test.

4. Numerical Illustration

A sample consisting of 10 observations was generated from N (μ=5, σ=1) and we want to test whether this sample was drawn from a population with mean, μ=5. The simulated sample observations are:

(5.50, 5.44, 4.13, 5.03, 5.19, 6.24, 5.06, 5.56, 5.36, 5.28)

Here we state hypotheses as H_0: μ=5, against H_1: $\mu \neq 5$ and choose the level of significance α=0.05. Applying one sample t-test we get the following result:

Sample Mean	sd	t	p-value
5.279	0.529	1.667	0.129

Consider that due to an error any observation, say the 3rd observation was misread as 8.13 instead of 4.13. Now if we apply the same test to the contaminated sample, we got a different result:

Sample Mean	sd	t	p-value
5.679	0.927	2.317	0.045

Thus, we observed that for the outlier free sample our H_0 was not rejected which supported the fact that the sample was drawn from a population with μ=5. But in the presence of a single outlier the p-value 0.045 < 0.05 leads to rejection of the H_0. Also, even a single outlier can inflate the value of sample mean and standard deviation.

5. Conclusion

When our samples are contaminated with outliers the parametric tests may lead to misleading results. The results also shows that inference should not be solely based on the *p*-values, because outliers tend to distort the uniformity of the *p*-value distribution. This leads to a situation, where we are not able to reject the false null hypothesis. So, it is very essential to check for outlier in any data set before analysis. It should be noted that the outlier is not always an error. Sometimes outliers may be legitimate cases sampled from the correct population. Then such observations can therefore serve as important input for further investigation.

References

[1] Dixon, W.J., (Dec., 1950): 'Analysis of Extreme Values', The Annals of Mathematical Statistics, Vol. 21, No. 4, pp. 488-506.
[2] Dixon, W.J., (Mar., 1950): 'Ratios Involving Extreme Values', The Annals of Mathematical Statistics, Vol. 22, No. 1, pp. 68-78.
[3] Grubbs, F. E., (1969): 'Procedures for detecting outlying observations in samples', Technometrics, 11, pp. 1-21.
[4] Gumbel, E.: Discussion on rejection of outliers by Anscombe, F.J., Technometrics, 2, 165–166, 1960. 4853
[5] Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108.
[6] Philip H. Ramsey, Kyrstle Barrera, Pri Hachimine-Semprebom & Chang-Chia Liu (2011): Pairwise comparisons of means under realistic nonnormality, unequal variances, outliers and equal sample sizes, Journal of Statistical Computation and Simulation, 81:2, 125-135
[7] H. M. James Hung, O'Neill, R. T., Bauer, P., & Köhne, K. (1997). The Behavior of the P-Value When the Alternative Hypothesis is True. *Biometrics*, *53*(1), 11–22. https://doi.org/10.2307/2533093

Author Profile

Dr. Mintu Kr. Das received M.Sc. and Ph.D. degrees in Statistics from Dibrugarh University, Assam, India in 2007 and 2009, respectively. During 2010-2012, he worked on a contractual basis at Regional Medical Research Centre, N.E. Region, (ICMR) Dibrugarh, Assam. Before joining as an **Assistant Professor** in the Department of Statistics, DHSK Commerce College, he was a UGC-BSR Research Fellow in the Department of Statistics, Dibrugarh University during 2013-2016.