International Journal of Statistics and Probability; Vol. 10, No. 2; March 2021
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Large Sample Problems

Jai Won Choi[1], Balgobin Nandram[2]

[1] Statistical Consultant, Meho Inc., 9054 Mary Knoll Dr., Rockville, MD 20850. E-mail: kycho1937@yahoo.com

[2] Professor, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280. E-mail: balnan@wpi.edu

Correspondence: Jai Won Choi, 9504 Mary Knoll Drive. Rockville, MD 20850. E-mail: kycho1937@yahoo.com

Received: December 6, 2020 Accepted: January 26, 2021 Online Published: February 21, 2021,

doi:10.5539/ijsp.v10n2p81 URL: https://doi.org/10.5539/ijsp.v10n2p81

Abstract

Variance is very important in test statistics as it measures the degree of reliability of estimates. It depends not only on the sample size but also on other factors such as population size, type of data and its distribution, and method of sampling or experiments. But here, we assume that these other fasctors are fixed, and that the test statistic depends only on the sample size.

When the sample size is larger, the variance will be smaller. Smaller variance makes test statistics larger or gives more significant results in testing a hypothesis. Whatever the hypothesis is, it does not matter. Thus, the test result is often misleading because much of it reflects the sample size. Therefore, we discuss the large sample problem in performing traditional tests and show how to fix this problem.

Keywords: Bayesian methods, large population, random group method, sample size, testing hypothesis, traditional methods

1. Introduction

It is very expensive to list all the people of a large population. Therefore, we take a sample to minimize survey cost. We use the sample to investigate certain characteristics of the US population by implementing traditional methods. For many surveys and experiments, the sample sizes are large. The large samples can always reject the null hypothesis. This result reflects not only the real significance but also the sample sizes. Therefore, we cannot perform the traditional tests with the samples exceeding certain size.

Although a large sample provides more information, it also causes problems in performing statistical tests. We noticed this problem when we worked at the National Center for Health Statistics (NCHS). NCHS collects large probability samples from the U.S. population and we analyzed the data from several National Health Surveys. We have tried to use existing methods such as normal test or student t-test for the analysis of the data. During this trial time, we have encountered the large sample problems in calculating the variance and degrees of freedom.

Variance is very important factor of the test statistic in the traditional testing and depends directly on the sample size. i.e., the variance becomes too small when the sample size is large. As a result of it, the test statistic becomes too big giving significant test results.

A concrete example is as follows. Let $X_1, X_2, \ldots, X_n$ are identically independently distributed as N(μ, σ^2), where σ^2 is known and inference is required about $\boldsymbol{\mu}$. We want to test the null hypothesis H_0: $\mu \geq \mu_0$ versus alternative H_1: $\mu < \mu_0$ in a random sample of size of n. We denote $\bar{\boldsymbol{x}}_{\boldsymbol{0}} \leqslant \boldsymbol{\mu_0}$, where $\bar{\boldsymbol{x}}_{\boldsymbol{0}}$ is the observed value of the sample mean, $\bar{X}$. Then the p-value of the test is

$$\mathrm{P}(\bar{\boldsymbol{X}} \leq \bar{\boldsymbol{x}}_{\boldsymbol{0}} \mid \boldsymbol{H_0})$$

$$= \mathrm{P}\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq \frac{\bar{x}_0 - \mu_0}{\sigma/\sqrt{n}} \middle| \boldsymbol{H_0}\right)$$

$$= \boldsymbol{\Phi}\{\sqrt{\boldsymbol{n}}\,(\bar{\boldsymbol{x}}_{\boldsymbol{0}} - \boldsymbol{\mu_0})\}$$

Here $\boldsymbol{\Phi}(.)$ is the cdf of standard normal random variables. Therefore, if n is very large and $\bar{\boldsymbol{x}}_{\boldsymbol{0}} \leq \boldsymbol{\mu_0}$, p-value $\approx \boldsymbol{0}$.

In recent years, we have discussed a lot of experience in hypothesis testing. These tests are both non-Bayesian (as we

have done here and elsewhere) and Bayesian (as we have done elsewhere). Non-Bayesian test of independence for weighted and correlated data is discussed by Choi and McHugh (1999). Bayesian tests for independence in contingency tables are discussed by Nandram, Kim and Zhou (2019), Bhatta, Nandram and Sedransk (2018), Yu, Bhadra and Nandram (2017), Nandram, Bhatta and Bhadra (2013), Nandram, Bhatta and Sedransk (2013), Nandram and Choi (2007); the paper by Nandram and Choi (2007) is our pioneering work on hypothesis testing. The Bayesian test, which is based on the Bayes factor, uses both hypotheses directly where the non-Bayesian test, based on the p-value, mostly uses just the null hypothesis. This can partially help explain the problem with non-Bayesian tests.

We discuss the random group methods (RGM) to solve such large sample problems. RGM is a simple approach in using the traditional tests. RGM is to divide a large sample randomly into smaller samples to avoid the large sample problem and enable us to use existing methods. We perform the traditional test for each of the smaller samples. If the 90 % of these tests are significant at a given significant level (e.g., 5% significance level), then we call it 90% significant also for the entire original sample. In order to do so, we need to pass through the two processes. First, we calculate the size of the small samples. Second, we show the reason why the significant rate of small samples also represents at least the same rate for the original sample. These are shown in Section 2.

We have searched the literature for the large sample problem. We can find short mention in the text book by DeGroot and Schervish (2002). They gave three suggestions (p529-530) for large samples. First, one can make the significance level much smaller than traditional p = 0.01 or 0.05 to fit the large samples. This idea is somewhat similar as ours, but only reversing the process. Second, replace the single value of mean in the null hypothesis by an interval. Third, regarding it as one of the estimation problems rather than the problem of testing a hypothesis. There is no benchmark for the first suggestion.

Another reference is the text book by Petruccelli, Nandram, and Chen (1999). On page 300, they stated "virtually any null hypothesis will be rejected if there are enough data, i.e., with enough data, the test statistic will produce a small p-value" regardless of real significance.

In both references, they gave only a brief mention of this problem. They recognized the large sample problems in testing a hypothesis, but did not give any solution.

We raised this question whenever we attended meetings or seminars. People have already recognized this problem for a long time, but avoided to discuss the question. The reason is very clear: There is no solution. But we felt now is the time to discuss this problem, and motivate them to do more research on this important topic.

We have done many works at NCHS that use large samples from NCHS surveys. We have written many papers. Some of them are Nandram and Choi (2002a, 2002b, 2006, 2010). In these papers, we tried to analyze the various topics on small area estimation, non-ignorable nonresponse problems and many others topics. In these, we used a Bayesian approach to avoid large sample problems. But now we feel strongly to present the large sample problem as many researchers abuse hypothesis testing.

We will divide this paper into seven sections. Section 2 show the two processes. Process One describes how to divide a large sample to smaller size. Process Two shows how the percent of the significant tests of small samples implies the same percent of the original sample before dividing. Section 3 shows how to divide the entire universe of samples into three sizes. Section 4 introduces the three statisticians who used small samples. Section 5 illustrates the large sample problems with the Table 1. Section 6 gives a few examples of large samples. Finally, Section 7 gives a concluding remark.

2. Random Groups

Let $\boldsymbol{x_n} = x_1, x_2, \ldots, x_n$ be the random variables of size n from f(**x**| α , β), α , β are the parameters of inrerest. When n is a large number, we want to divide the sample into h smaller samples to be able to use a traditional method for testing hypothesis.

Process One: the size of small samples

Start from the initial number m, 1<m<n, the smallest number enables us to perform a traditional test. Let t_m be the test statistic

$$T(f_m | x_m, m) = t_m, \quad (1)$$

where f_m f($\boldsymbol{x_m}$ | α′ , β′), $\boldsymbol{x_m} = x_1, x_2, \ldots, x_m$ and α′ , β′ are the estimates from f(x_m | α , β). The test statistic t_m provides the probability p_m

Note that the test statistic t_m is also a function of sample size m of the given random variable $\boldsymbol{x_m}$.

Gradually increasing sample size from m to (m + k) until the inequality (3) below is achieved. Here, the test statistic of

the new sample size (m + k) is

$$T(f_{(m+k)} \mid x_{(m+k)}, (m + k)) = t_{(m+k)}, \quad (2)$$

where $f_{(m+k)}$ = f($\boldsymbol{x}_{(\boldsymbol{m+k})}$ | α′ , β′) , where $\boldsymbol{x}_{(\boldsymbol{m+k})} = x_1, x_2, \ldots x_{(m+k)}$, and α′ , β' are the estimates from f($x_{(m+k)}$| α , β). The test statistic $t_{(m+k)}$ provides the probability $p_{(m+k)}$.

At the sample size of (m + k - 1) , the inequality of $|p_m - p_{(m+k-1)}| < p$ is not changed. At the sample size (m + k), the inequality is reversed as

$$|p_m - p_{(m+k)}| > p, \quad (3)$$

where the change of inequality in (3) arises when the sample seize m increased exactly to (m + k). Here we choose one p, a significant probability, (0 < p < 1). Choose p conservatively so that smaller sample size is not too large for the traditional testing. Note that the inequality of the two test statistics from (1) and (2), we have $t_m < t_{(m+k)}$, which gives the inequality of two significant probabilities, $p_m > p_{(m+k)}$, where (1< (m+k) < n).

The sample size (m + k) is the change point from p_m to $p_{(m+k)}$ at the probability p to achieve the inequality (3). In this case, the size of small samples is (m + k).

If we still have $|p_m - p_{(m+k)}| < p$, the inequality is not reversed, at (m + k), further increase from (m + k) to (m + k + r), (1< r < (n-m- k) < n. As in (1), forming the test statistic,

$$T(f_{(m+k+r)} \mid \boldsymbol{x}_{(\boldsymbol{m+k+r})}, m+k+r) = t_{(m+k+r)}. \quad (4)$$

The test statistic $t_{(m+k+r)}$ provides the significant probability $p_{(m+k+r)}$: r is an increasing positive integer until the inequality (5) is achieved.

$$|p_m - p_{(m+k+r)}| > p. \quad (5)$$

In this case, (m + k + r) is the size of the small samples.

Once (3) or (4) reversed the inequality, then (m + k) or (m + k + r) is the size of the small samples. Once we found the number (m + k) or (m + k + r), we divide the large sample of size n into the h random groups of the size (m + k) or (m + k + r), Here, (1 < r < (m+k+r) < n).

If the size of the small sample, (m + k), is taken, we divide $\boldsymbol{x}_{\boldsymbol{n}} = x_1, x_2, \ldots, x_n$, into the h smaller samples of $x_{i(m+k)} = x_1, x_2 \ldots, x_{i(m+k)}$, i = 1, 2, ...,h. The small samples are the same size: i.e. $\boldsymbol{x}_{(\boldsymbol{m+k})} = \boldsymbol{x}_{\boldsymbol{i(m+k)}}$, for all i, and the original large sample of size n = h (m + k). However, the contents of the h small samples are different. Each sample $\boldsymbol{x}_{\boldsymbol{i(m+k)}}$ provides the different test statistic $t_{i(m+k)}$, which in turn gives different significant probability $p_{i(m+k)}$, i = 1, 2, …, h.

Below shows Process Two, the percent of the significant tests among the small samples also implies that at least the same significant percent for the original sample of size n.

Process Two: Significance of the sample of size n

Take a probability $\boldsymbol{\pi}$, 0 < $\boldsymbol{\pi}$ < 1, for the significant level of test (say 0.01 or 0.05). Note that here probability $\boldsymbol{\pi}$ is different from the p used in the process one. To count the number of small samples satisfying the condition ($p_{i(m+k)} < \boldsymbol{\pi}$), using indicator function I. We have

$$R_{(m+k)} = \sum_{i=1}^{h} \frac{I(p_{i(m+k)} \leq \pi)}{h}, \quad (6)$$

where $I(p_{i(m+k)} \leq \pi) = 1$ if $p_{i(m+k)} < \boldsymbol{\pi}$ and 0 otherwise, i = 1, …, h, and

$R_{(m+k)}$ x 100 is the % of the counts satisfying $p_{i(m+k)} < \boldsymbol{\pi}$ among the h small samples at the significant probability $\boldsymbol{\pi}$.

If 90% of $p_{i(m+k)}$ are $p_{i(m+k)} < \boldsymbol{\pi}$ among the h small samples, we define it 90% significant for all h groups at significant probability $\boldsymbol{\pi}$. Also, we claim that the p_n of the original sample of size n is also 90% significant at the same probability $\boldsymbol{\pi}$ for the entire sample of x_n.

Lemma. Suppose a large sample of size n is divided into h samples of sizes (m + k). If 90% of the h small samples are significant or $p_{i(m+k)} < \pi$ at the $\pi = 0.05$, then the test statistic p_n for the original sample of size n, is also at least 90% significant or $p_n < \boldsymbol{\pi}$ at the probability $\boldsymbol{\pi}$.=0.05. **Appendix A** shows the sketch of the proof.

3. Dividing all Samples Into Three Sizes

We may roughly divide the universe of all the samples into three sizes:

1. Small size (n<5);
2. Middle size 5<n<100;
3. Large size (100<n).

The boundary of these divisions is somewhat arbitrary. They can be readjusted by the equation (3).

We assume the test statistic depends only on the sample size of the random variables x. We discuss the size of the small, middle, and large samples separately. We have problems only for the sample of too small (n<5) and too large (100<n).

The middle size 5<n<100,

We can apply the existing methods (e.g., t-test or normal test) if the change of the inequality (3) occurs in this interval. Any sample falling in this interval can be tested with the traditional method, and one test is enough.

Large samples (100 < n).

This is the only place we can apply the method developed in Section 2. Take the sample size n = 400. We randomly divide a large sample of 400 into smaller sizes of (m + k) = 40 forming the h = 10 random groups. Then we calculate its mean and variance of each and apply traditional method to each for testing a hypothesis. If 90% out of 10 tests, 9 tests, are significant at π = 0.05, we define the overall test with the sample of n=400 is also 90% significant at the same π = 0.05.

If there are too many groups to handle for a large sample (say n= 10 millions), then we have 250,000 random groups of 40. We may use only a simple random sample of the 100 groups out of 250,000 groups. Performing the same test for all 100 groups, only 95 were significant at π = 0.05. We can conclude that it is true that the 250,000 groups also 95% significant at the same π = 0.05. Furthermore, this is also true for the large sample of n = 10 million. Note that the estimates from a simple random sample provide the unbiased estimates.

Small samples (n<5)

Test result may be unreliable if the sample size is too small. For the small sample, we do not have enough information to make proper inference based on the assumed distribution. For the small sample, e.g., n<5, the size of n is too small to form a distribution. One may use one of the distribution free methods or nonparametric methods such as Fisher's exact test, the sign test or U test.

If we have n = 2, (space between n and =) for example, the similarity between two persons or two companies. Choi and Nandram (2000) discussed this problem comparing two persons. We need more research in this area for statistical inference.

4. Three Statisticians Used Small Samples: Fisher, Bayes, and Gosset

In the early years in the 1900s, the only tool was the pen, and the early researchers calculated the mean and variance manually. One example is Fisher's tea tasting. Another example is Gosset's student t-test to choose better combinations in tasting beer. Another method for small sample is Bayesian method. Assuming the distribution of a small sample, we can generate as many samples as needed from this distribution for the statistical inference. If this assumption is correct, this is a good approach.

R. A. Fisher (1890-1962, Figure 1)

Fisher studied Mathematics at Cambridge University from 1909. Among many accomplishments, he began to publish the papers related to the maximum likelihood estimation during 1912 – 1922 years laying the foundation of current statistics. He developed the design of experiment at Rothamsted Experiment Station. He used the terms "variance" and "analysis of variance" for the first time. Fisher used small sets of data in his studies not exceeding more than he could calculate with his pen or pencil. For example, a few plots were used for his agricultural experiments at Rothamsted.

Lady's tasting tea is an example for small sample. It is a randomized experiment reported in the Chapter Two of Fisher's book, the Design of Experiments (1935). He used 8 cups, putting tea first in 4 cups and cream first in 4 cups randomly. He asked Ms. Muriel Bristol to identify which, tea or cream, was the first in his randomized blind test. The test used was Fisher's exact test, a nonparametric test on the 2 x 2 table. It is useful test for a small sample. Here he used only 8 cups and calculated 70 combinations of 4 cups out of 8 cups. This could be the maximum numbers he could calculate at the time of no calculator or computer.

Thomas Bayes (1702–1761, Figure 2), Bayesian method

Bayes was born in London, England, a nonconformist Presbyterian minister and mathematician, graduated from University of Edinburgh. Nonconformists could not go to Oxford or Cambridge at the time.Two years after his death, Richard Price edited and corrected "An Essay Towards Solving a Problem in the Doctrine of Chances", prior to

publication in 1763. He was the first to use the probability inductively calculating the probability given the prior probability. It has been used widely recently with the improvement of computing technologies.

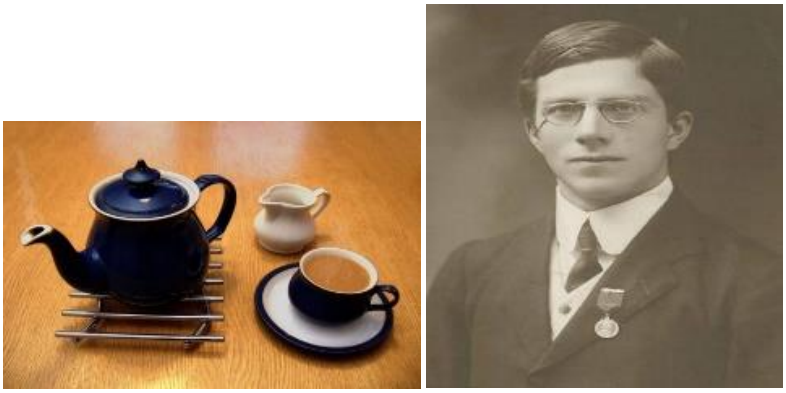

Figure 1. Fisher's Young Portrait

Figure 2. Bayes' Portrait

William Sealy Gosset (1876-1937) introduced the Student's t-distribution for small samples. Gosset reported "Student's test of statistical significance" in Biometrika,1908, under the pen name Student. He graduated from Oxford influenced by Karl Pearson and W.A. Spooner. It is useful for establishing confidence limits for the mean estimated from smaller samples. He was a chemist and brewer at Guinness Brewery. Pioneered in small sample experimental design, he was testing means to make a better brew.

In recent years, data size exploded and there is no way to analyze the data without a computer. For example, Affordable Health Care needs to control the cost of the plan. To do so, they need to know the number of doctor visits. National Health Interview Survey (1990) used a sample of about 12,000 people from the U.S. population of about 300 million and it reported that people visited doctors' office an average of 4 times a year excluding hospital inpatients, 48,000 visits in all. To complete cost calculation, need not only number of visits but also the related information such as whom and why they visited, and how much they paid for their visits and medications. The final data could be increased to millions. In this case, even if a small fraction of it, its data size will be prohibitively large. No current methods would work with such a huge sample size. Section 2 is useful tool in this case. Choi (2011) discussed this problem for testing hypothesis under the large sample situation and Nandram and Choi (2002b) studied these doctor visits.

5. Discussion of Table 1 Showing Large Sample Problems

The following is a hypothetical example showing a danger for large samples. Suppose a drug company submitted application to FDA to get approval of a medication. When its test result shows no significance, FDA rejects it. Then the drug company increased the sample size to obtain the significant test result. When they reapplied with the new significant result, FDA approved it. Here, the only thing changed is the sample size, keeping all others same. It has nothing to do with the efficacy of the drug.

Here we show that test score Z depends on the variance of the effective proportion p of clinical trial and a null hypothesis (say, p=0). This variance of p in turn depends on the sample size n. Table 1 shows the standard deviations S

and the normalized test scores Z for the sample sizes n = 5, 10, 20, 40,100, 400, 500, and 1000, and the proportions p = 0.1, 0.2, 0.4, 0.5, 0.6, 0.6, 0.8, and 0.9. If we take p = 0.1 on the first column and sample sizes n = 5, 10, 20, 40,100, 400, 500, and 1000, we have

S = 0.134, 0.095, 0.067, 0.047, 0.030, 0.015, 0.013, 0.009, and

Z = 0.745,1.054, 1.491. 2.108, 3.333, 6.666, 7.454, 10.540.

Here we see that, when n is increasing from 5 to 1000, the standard deviation S is decreasing from 0.134 to 0.009, while the normal test scores Z is increasing from 0.745 to 10.540.

Table 1 below also shows that, no matter what the size of proportion p is, if n

is greater than 40, all the test result Z is significant (i.e. Z > 1.65) at π = 0.05. The numbers of Zs of red color are less than 1.65. Here nothing is changed except the sample size for the given p.

Resampling hides real problems

National Health Interview Survey (NHIS) (Jack and Ries,1981) had the national sample of about 12,000 people and National Health and Nutrition Examination Survey (NHANES III 1988-1994) was a national probability sample of 39,695 persons aged 2 months and older (about 33,000 older than 18 years of age). We encountered large sample problems. For large samples, some at NCHS used the resampling methods for variance calculation: Replication or Balanced Half Sample (BHS) (McCarthy, 1966), Jackknife, or Bootstrap (Efron, 1982). For example, NHIS used BHS and generalized variance function (GVF) to calculate the variance of the NHIS data. (Choi, 1989). We had to manipulate the data to use BHS. The resulting variances were used to draw the GVF curve, GVF is g(x | a, b)= a + b/x, where x is the sample size n of the variable, and the parameters a and b are obtained by the least square estimation. The variances obtained from BHS and GVF cannot be used for testing a hypothesis. For example, if we test the mean, BHS and GVF do not give the variance of this mean, but the variance indirectly calculated by BHS with the manipulated data and approximation of GVF curve. This resulting variance is not related to the mean that we want to test.

Bootstrap is to take samples from an original sample. It does not give any better information than the original sample itself. If the original sample is biased, a sample from the original sample is also biased. Hence, we should be careful when we use a sample for Bootstrap.

Table 1. Standard deviation S and normal score Z for given sample size n and proportion p. Sample variance var(p) = p(1-p)/n, standard deviation S = $\sqrt{\text{var}(p)}$, Normal approximation Z = p/S for the estimate p under the null hypothesis p = 0. The Z = 1.65 at the probability π = 0.05. Red color numbers of Zs are smaller than 1.65

N		p=0.1	p=0.2	p=0.4	p=0.5	p=0.6	p=0.8	p=0.9
n=5	S	0.13416	0.17889	0.21909	0.22361	0.21909	0.17889	0.13416
	Z	0.74536	1.11803	1.82574	2.23607	2.73861	4.47214	6.70820
n=10	S	0.09487	0.12659	0.15491	0.15811	0.15491	0.12659	0.09487
	Z	1.05409	1.58114	2.58199	3.16228	3.87298	6.32456	9.48686
n=20	S	0.06708	0.08944	0.10954	3.65148	0.10954	0.08944	0.06708
	Z	1.49071	2.23607	3.65148	4.47214	5.47723	8.94427	13.4164
n=40	S	0.04743	0.06325	0.07746	0.07906	0.07746	0.06325	0.04743
	Z	2.10819	3.16228	5.16398	6.32456	9.74597	12.6491	18.9737
n=100	S	0.03	0.04	0.04899	0.05	0.04899	0.04	0.03
	Z	3.33333	5.0000	8.16497	10.0000	12.2475	20	30
n=400	S	0.015	0.02	0.02449	0.025	0.02449	0.02	0.015
	Z	6.66667	10	16.3299	20	24.4949	40	60
n=500	S	0.01342	0.01789	0.02191	0.02236	0.02191	0.01789	0.01342
	Z	7.45356	11.1803	18.2574	22.3607	27.3851	44.7214	67.0820
n=1000	S	0.00949	0.01265	0.01549	0.01581	0.01549	0.01265	0.00949
	Z	10.5409	15.8115	25.8199	31.6228	38.7289	63.2456	94.8683

6. Large Sample Examples

Example 1

A doctoral student presented her research results. The sample sizes of her studies were over 1,500 and her test results were all very significant. Our suggestion was that she could form 15 random groups of (m + k) = 100 to see how many of the 15 groups are significant. If the 95% of them (about 14 out of 15) are significant at π = 0.05, she may conclude that it is also actually 95% significant with the original sample of n = 1,500 at π = 0.05. Here the group size is 100 from inequality (3) in Section 2.

Example 2.

We can randomly divide NHANES III sample of 33,000 people for testing hypothesis of the mean of body mass index (BMI). We may choose sample size according to the method in Section 2. If the small sample sizes (m + k) =100 by the inequality (3) in Section 2, we can form the 330 random groups from n = 33,000 people.

We apply one of the current tests to each of 330 groups. If we have the 90% of them significant at π = 0.01, we can claim at least 90% of the original sample of size n = 33,000 is also significant at the significant level π = 0.01.

Nowadays, we have unlimited computing power in developing deep learning or artificial intelligence (AI) technique. For example, Google search gives the information on Thomas John Watson Sr. (February 17, 1874 – June 19, 1956). He was an American businessman, served as the chairman and CEO of International Business Machines (IBM). He oversaw the company's growth into an international force for 42 year from 1914 to 1956, developing Watson technology. IBM Watson Technology is reinventing The Way We Work, Discover More facts, Data Intelligence, Cognitive Technology, Cognitive Innovation, and Watson Ecosystem. Even if the data size is over billions, the calculation is not a problem with currently available computing power.

7. Conclusion

We can use the traditional methods for the samples of the middle range i.e. (5< n < 100). For the samples of large n (i.e., n > 100), the test results are unnecessarily too significant. So following the process presented in Section 2, we need to form random groups in traditional testing of a hypothesis at a given significance level π.(delete m).. For the small samples (n<5), one may use a nonparametric statistic that does not depend on the variance (or n size). A type of new nonparametric method, we hope, can be developed to replace the traditional parametric testing in order to accommodate the small samples better.

For the small sample i.e. (n < 5), Choi and Nandram (2000) discussed the measure of similarity between two persons using distances between them for each characteristic or trait. When there are at least five traits such as race, age, sex, height, and color of hair, the similarity of them is measured by the square root of the sum of the weighted distances of the five characteristics.

It is not simple to handle large data sets. In the era of COVID-19, the counts by states are very large. For example, on Saturday October 23, the COVID Tracking Project announced for the US states. For Massachusetts, among the 7883 tests, 2123 infected, and 142 deaths in the population of 6,547,629 people. The proportion of death among the tested is very small p = 142/7883 = 0.01. We have a test of null hypothesis $H_0: p \leq 0.05$ vs $H_1: p > 0.05$, the 0.05 or 5.0 percent is a threshold used to open a state.

Variance is very small as n= 7883. Hence the p-value is near 0 and reject the null hypothesis and accept the alternative hypothesis. Here the whatever the null hypothesis is, it does not matter. We always reject null hypothesis and accept the alternative hypothesis. The rejection of null hypothesis does not mean the acceptance of alternative automatically. For this type of large sample the RGM would not work well because we do not have the individual person's data.

Recently the data size has been increasing rapidly. Using the increased computing power, we can handle large data through AI technique. However, in the age of AI, we still use the old method for testing hypothesis that depends on the sample size. For large samples, we can still use the traditional tests via the RGM presented in this paper.

Appendix A. Sketch of the Proof of Lemma

Take n = 200, divide 200 into 10 groups (i.e., h=10) of size 20, i.e.,(m + k = 20). The test statistic $T_{20}(f_{i20} \mid x_{i20}, 20) = t_{i20}$, and t_{i20} *provides* p_{i20}. We count $p_{i20} < \boldsymbol{\pi}$ from these 10 random groups at $\boldsymbol{\pi}$= 0.05. Now, for the original sample of 200. Here h=1 and $T_{200}(f_{200} \mid x_{200}, 200) = t_{200}$, *and* t_{200} provides p_{200}.

It is clear that $t_{i20} < t_{200}$,and $p_{200} < p_{i20} < \boldsymbol{\pi}$, i = 1, … ,10. Then $R_{200} \geq R_{20}$. For the equality hold for h=1 for n= 200, and (m + k) = 20.

In general, it is true, observing $t_{(m+k)} \leq t_n$ and $p_n \leq p_{(m+k)} < \boldsymbol{\pi}$, for any numbers h, n, and (m + k), here (1≤ h < n), and (1< (m+k) < n). Then the equation (6) shows that $R_n \geq R_{(m+k)}$..

References

Bhatta, D. R., Nandram, B., & Sedransk, J. (2018). Bayesian Testing for Independence of Two Categorical Variables Under Two-Stage Cluster Sampling with Covariates. *Journal of Applied Statistics, 45*(13), 2365-2393.

Choi, & McHugh. (1989). An Adjustment Factor for Goodness of Fit and Independence Test for Test for Correlated and Weighted Observation. *Biometrics, 45,* 979-996. https://doi.org/10.2307/2531697

Choi, J. (1989). *Variance of Health Surveys at National Center for Health Statistics.* ASA 1988 Proceedings of the Section on Survey Research Methods, 734-739.

Choi, J. (2011). *A Thought on the Current Statistics, News Letter of Korean Statistical Association.* April, 2011, 23-26.

Choi, J., & Casady, R. (1984). Variance in the National Health Interview Survey Data. ASA 1983 Proceedings of the Section on Survey Research Methods, 343-146.

Choi, J., & Nandram, B. (2000). A Measure of Concordance When There Are Many Traits, ASA 1999 Proceedings of the Section on Survey Research Methods, 837-842.

DeGroot, M. H., & Schervish, M. J. (2002). *Probability and Statistics* (3rd ed). Addison Wesley.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans. Series 36, SIAM, 1405 Architects Building, 117 S.* 17th Street, Philadelphia. https://doi.org/10.1137/1.9781611970319

Fisher, R. A. (1937). *The Design of Experiments* (2 ed.). Oliver and Boyd, Edinburgh.

Jack, S. S., & Ries, P. W. (1981). *Current Estimates From the National Health Interview Survey: United States, 1979.* Series 10 No. 136. (PHS) 81-1564, PC A07 MF A02. Washington, D.C.

McCarthy, P. J. (1966). Replication. An Approach to the Analysis of Data From Complex Surveys. Development and Evaluation of a Replication Technique for Estimating Variance, Series 2, National Center for Health Statistics. DHEW Publication No. (HSM) 73-1269. Washington, D.C.

McCarthy, P. J. (1982). Estimated Variance for the Combined Ratio Estimate Stratified, Two~stage Samples Without Replacement. B. V. Sukhatrne Memorial Lecture. Iowa State University, Apr. 16, 1982.

Nandram, B., & Choi, J. W. (2002a). A Hierarchical Bayesian Nonresponse Models for Binary Data from Small Areas with Uncertainty About Ignorability. *Journal of American Statistical Association, 97*(457), 381-388. https://doi.org/10.1198/016214502760046934

Nandram, B., & Choi, J. W. (2002b). A Bayesian Analysis of a Proportion under Nonignorable Nonresponse. *Statistics in Medicine, 21,* 1189-1212. https://doi.org/10.1002/sim.1100

Nandram, B., & Choi, J. W. (2006). Hierarchical Bayesian Nonignorable Nonresponse Regression Models for Small Areas: An Application to the NHNES III Data. *Survey Methodology, 11*(1), 73-84.

Nandram, B., & Choi, J. W. (2010). A Bayesian Analysis of Body Mass Index Data from Small Domains under Nonignorable Nonresponse and Selection. *Journal of American Statistical Association, 105*(489), 120-133. https://doi.org/10.1198/jasa.2009.ap08443

Nandram, B., Bhatta, D. R., & Bhadra, D. (2013). A Likelihood Ratio Test of Quasi-Independence for Sparse Two-Way Contingency Tables. *Journal of Statistical Computation and Simulation, 85*(2), 284-304.

Nandram, B., Bhatta, D. R., & Bhadra, D. (2013). A Likelihood Ratio Test of Quasi-Independence for Sparse Two-Way Contingency Tables. *Journal of Statistical Computation and Simulation, 85*(2), 284-304.

Nandram, B., Bhatta, D. R., Sedransk, J., & Bhadra, D. (2013). A Bayesian Test of Independence in a Two-Way Contingency Table Using Surrogate Sampling. *Journal of Statistical Planning and Inference, 143,* 1392-1408.

Nandram, B., & Choi, J. W. (2007), Alternative Tests of Independence in Two-Way Categorical Tables. *Journal of Data Science, 5*(2), 217-237.

Nandram, B., Kim, D., & Zhou, J. (2019). A pooled Bayes test of independence for sparse contingency tables from small areas. *Journal of Statistical Computation and Simulation, 89*(5), 889-926.

Petruccelli, J. D., Nandram, B., & Chen, M-H. (1999). *Applied Statistics for Engineersand Scientists.* Prentice Hall, New Jersey.

Yu, Y., Bhadra, D., & Nandram, B. (2017), Tests of Independence for a 2 by 2 Contingency Table with Random Margins. *International Journal of Statistics and Probability, 6*(2), 106-121.