PROBABILITY THEORY AND INTRODUCTORY STATISTICS

ALY6010, FALL 2019

MODULE 4 PROJECT ASSIGNMENT

One-sample Confidence Intervals & Hypothesis Testing

SUBMITTED BY: PARINITA MAHENDRIKAR, SHIVANI ADSAR, ANUPREETA MISHRA
NUID: 001069853, 001399374, 001050752
CRN: 71447
SUBMITTED TO: DR. DEE CHILUIZA REYES
DATE: 10/13/2019

Introduction

The assignment aims at performing hypothesis and estimates on the population parameter. Using the data, we have performed the hypothesis testing to estimate and test a hypothesis regarding some parameter like population mean, population proportion and population standard deviation. The usage of parameters like Test Statistic, P-value and Significance level by using Null Hypothesis and Alternative Hypothesis has led to the decision of rejecting or not rejecting the null hypothesis. The null hypothesis, symbolized by H0, is a statistical hypothesis that states that there is no difference between a parameter and a specific value, or that there is no difference between two parameters. Hypotheses are always statements about the population (1). The alternative hypothesis, symbolized by H1, is a statistical hypothesis that states the existence of a difference between a parameter and a specific value, or states that there is a difference between two parameters (2). They help us solve problems such as if the average time a student spends studying each week equal to 20 hours, will this earn the student a GPA of 3.0 or higher.

Analysis

Part 1: Hypothesis Testing

Q1. Confidence intervals for sample size of 160 for Location Quotient parameter

On the basis of the Location Quotient, samples with size 160 were obtained using random sampling from Data analysis tool in Excel. Table 1 shows the Mean, Variance and Standard Deviation of the population. Table 2 indicates mean, standard deviation, sample size and sampling error for the samples. It can be observed the mean remains constant for the sample size of 160, however small variation is observed in the standard deviation.

Table A	
Population Mean:	1.04
Population Variance:	0.84
Population Standard deviation:	0.92

Table 1: Population statistics for Location Quotient

Table B	
Sample Mean	**1.04**
Sample Standard deviation	**0.87**
Sample size	**160**
Sampling (standard) Error	**0.07**

Table 2: Sample statistics for Location Quotient

Table3. denotes the margin error calculated for 160 samples of Location Quotient. Considering the sample size is >= 30, the z value will be calculated for given Confidence Levels (excel formula: $\boldsymbol{z} = \boldsymbol{NORM.S.INV}((\mathbf{1}+\boldsymbol{c})/\mathbf{2})$). The margin of error was calculated using the formula $\boldsymbol{E} = \boldsymbol{z}^*(s/\sqrt{\mathbf{n}})$, where s is the standard deviation. A margin error of 0.12 was calculated for 92% interval with an increasing trend for greater confidence intervals. Table 4. Denotes the minimum sample size needed for a desired margin of 0.03 units for confidence level of 92%, 96% and 98%.

Confidence Level CI	z value	Margin of Error	CI Lower Limit	CI Upper Limit	CI Width
92%	1.75	0.12	**0.92**	**1.16**	0.24
96%	2.05	0.14	**0.89**	**1.18**	0.28
98%	2.33	0.16	**0.88**	**1.20**	0.32

Table 3: Margin of Error for 160 Location Quotient

Confidence Level CI	Desired Margin of Error	Minimum Sample Size Needed
92%	0.03	**2601**
96%	0.03	**3579**
98%	0.03	**4593**

Table 4: Sample size for 0.03 margin of error

Q2 Confidence intervals for sample size of 23 for Hourly wage parameter

The random samples of size 23 have been calculated based on the Mean hourly wage. The sample mean of the data is 30.12 and standard deviation of 17.59. The degree of freedom is the number of values in the final calculation of a statistic that are free to vary. In the population samples, the degree of freedom is calculated to be 22.

Table A	
Sample Mean	30.12
Sample Standard deviation	17.59
Sample Size	23
Degrees of Freedom (DF)	22
Sampling Error	4

Table 5: Population Statistics for Mean Wage Sample

The Sampling Error occurs when the results of the sample do not represent the results that would be obtained from the population. The sampling error is 4 for the population data. The Confidence Interval is computed from the statistics that might contain the true value of an unknown population parameter.

Table B					
Confidence Level CI	z or t value corresponding to the CI level	Margin of Error	CI Lower Limit	CI Upper Limit	CI Width
92%	1.835416576	6.730475014	28.28	31.95	3.67
96%	2.18289265	8.004670238	27.93	32.30	4.37
98%	2.508324553	9.198029457	27.61	32.62	5.02

Table 6: T Value ,Margin Error and CI Upper and Lower Limit for Population

Since the sample size is 23, which is < 30, we are using the t-value formula of: ***t=T.INV(P(T<t), df)*** to calculate corresponding CI levels. The CI Width is calculated by subtracting the CI Upper Limit and the CI Lower Limit. On the basis of the Confidence Interval, the CI Lower Limit, Upper Limit, CI Width and Margin of error has been calculated. The Margin of Error for 92% CI is 6.73,which becomes 2.18 for 96% and increases to 2.51 at 98%.Hence we can see that the Margin of Error keeps increasing.

Q3. Location Quotient for Sample size 150000000

On the basis of the Location Quotient of the population, the random samples of size 1500 have been generated. On the basis of samples, 1275 denote the Total Samples lesser than 1.5. Based on the CI intervals of 90%, 95% and 99%,

Table A	
Sample Size	1500.0000
Total Samples < 1.5	1275.0000
Sample Proportion of Success (< 1.5)	0.8500
Sample Proportion of Failure (>= 1.5)	0.1500
Sampling Error	0.0256

Table 7: Location Quotient for the given Population

we have calculated the corresponding z-values of 1.6, 1.9 and 2.5 respectively. We use z, as the number of samples is 15000000 (excel formula: $\boldsymbol{z = NORM.S.INV((1+c)/2)}$ which is >=30.

Table B					
Confidence Level CI	z value corresponding to the CI level	Margin of Error	CI Lower Limit	CI Upper Limit	CI Width
90%	1.6449	0.042	1.01	1.0915	0.0842
95%	1.9600	0.050	1.00	1.0995	0.1003
99%	2.5758	0.066	0.98	1.1153	0.1318

Table 8: Z Value ,Margin Error and CI Upper and Lower Limit for Population

Based on the desired margin of error given as 0.04, minimum sample size has been calculated. We can see from the table that, the population is 7286, Total population lesser than 1.5 is 6200. Moreover, the Population Proportion of Success (<1.5) is 0.85 and the Population Proportion of Failure (>=1.5) is 0.14.We see that the Margin of Error has been greatly reduced, from 2.5 at 99% when sample size is 23 to 0.066 at 99% when it is 15000000,even though the population is the same for both the cases.

Q4. Hourly Median Wage for sample side 150

According to the population of Hourly Median Wage, random samples of size 150 have been created. The Sample Variance, Sample Standard Deviation and Degree of Freedom have been calculated as per the sample data.

Sample Variance	183.80
Sample Standard deviation	13.56
Sample Size	150
Degrees of Freedom (DF)	149

Table 8: Hourly Median Wage for the given Population

The confidence levels of 92%,96% and 98% are used to calculate the χ2Left, χ2Right, CI Lower for Variance, CI Upper for Variance and CI for Standard Deviation. The calculations of χ2Left are based on the formula CHISQ.INV(1+c/2,DF) and for χ2Right is CHISQ.INV(1-c/2, DF) where DF equals n-1.

Confidence Level CI	χ^2_{Left}	χ^2_{Right}	CI Lower for Variance	CI Upper for Variance	CI Lower for Standard deviation	CI Upper for Standard deviation
92%	120.198	180.553	151.68	227.85	12.32	15.09
96%	115.726	186.560	146.80	236.65	12.12	15.38
98%	111.802	192.073	142.58	244.96	11.94	15.65

Table 9: Chi-Square for Population

.

Q5. Location Quotient with sample size 200

According to the Location Quotient of the population, random samples of size 200 have been generated. The parameters like, the mean using AVERAGE formula, Standard Deviation using the STDEV formula and Sampling error using the Sample SD/SQRT(Sample Size) formula. As this is a two-tailed problem, we have used the hypothesis testing for population mean by using Ho: μ= μo for null hypothesis and Ha: μ ≠ μo for alternative hypothesis. The P-value has been calculated using $\mathbf{2} * (\mathbf{1} - \mathbf{norm.s.dist}(\boldsymbol{z}, \mathbf{1}))$ formula. Since the P-value ie. 0.8 is more than the Significance Level ie. 0.05, we will not reject the null hypothesis. Also, the critical value of -1.96 for both the lower and upper limit is calculated using norm.s.inv(*α/2)* formula.

Q6. Hourly Wage with sample size 29

Based on the mean hourly wage of population, we have created random samples using size 29. The parameters like, the mean using AVERAGE formula, Standard Deviation using the STDEV formula and Sampling error using the Sample SD/SQRT(Sample Size) formula. As this is a Right Tailed Problem, we have calculated the null hypothesis using Ho: μ<= μo and alternative hypothesis as : μ>μo. Also, the test statistic has been calculated based on the sample mean, standard deviation, size and the hypothesized mean. In order to calculate the P-value ie. 0.87, 1 – norm.s.dist(*z* , 1*) formula has been used.* The critical value of 3.40 as the lower limit is calculated using norm.s.inv(1 – α) formula**.** Since the P-value ie. 0.87 is more than the Significance Level ie. 0.01, we will not reject the null hypothesis.

Q7. Location Quotient with sample size 800

Using the Location Quotient as the population, the random samples size 800 have been generated.
The parameters like the Total samples<1.5, sample proportion of Success(<1.5), sample proportion of failure and the sampling error have been calculated. As this is a Left Tailed problem, we have calculated the Null Hypothesis using Ho: μ>= μo and Alternative Hypothesis using Ha: μ< μo.

The P-value ie. 1 for this problem is calculated using norm.s.dist(*z* , 1*) formula.* The critical value with upper limit of -1.6 is calculated using norm.s.inv(α) formula**.** Since the P-value ie. 1 is more than the Significance Level ie. 0.05, we will not reject the null hypothesis.

Q8. Hourly Median Wage for sample size 130

Based on the Hourly Median Wage as the population, we have generated random samples with size of 130. The parameters like Variance, Standard Deviation and Degree of Freedom have been calculated. This is a Two Tailed Problem considering the hypothesis testing for a population Standard Deviation. The Null Hypothesis Ho: σ= σo and Alternative Hypothesis Ha: σ≠ σo.In order to calculate the P-value, the 2* (1 – chisq.dist(*χ2* , df , 1*))* formula is used to calculate the minimum and maximum values, out of which the minimum value is considered as the P-value. The critical value 1 of 103.76 is calculated using chisq.inv(*α/2 , df)* formula and the critical value 2 of 156.50 is calculated using chisq.inv(1 – *α/2 , df).* Since, the P-value is lesser than the Significance Level, we will reject the Null Hypothesis.

CONCLUSION:

1. We see that the Margin of Error keeps decreasing with the increase in the number of samples.
2. By conducting major hypothesis tests we see that these tests can give a deeper insight into the population parameters.

   Hypothesis testing plays a crucial role in making business or strategic decisions.

It also helps in writing research papers.One such example would be the following statement we see: The mathematical formulation of a hypothesis testing problem in quantum statistics under consideration reduces to the following. Let {ψj} be a given basis in a d-dimensional unitary space. Find an orthonormal basis {ej} which approximates the basis {ψj} in the sense that the value of one is minimal. An asymptotic solution to this problem is given for «almost orthogonal» vectors ψj. An asymptotically optimal basis is ψˆj=Γ−1/2ψj, where Γ is the Gram operator of the system {ψj}(3).

It helps us to decide if something was the truth, or if certain treatments have positive effects, or if groups differ from each other or if one variable predicts another. By using hypothesis testing we want to proof if our data is statistically significant and unlikely to have occurred by chance alone. In other words, a hypothesis test is a test of significance.

Reference

1. Bulman, A. G. (n.d.). Probability and Counting Rules. In ELEMENTARY STATISTICS: A STEP BY STEP APPROACH, TENTH EDITION (10th ed., p. A-440). New York
2. Bulman, A. G. (n.d.). Probability and Counting Rules. In ELEMENTARY STATISTICS: A STEP BY STEP APPROACH, TENTH EDITION (10th ed., p. A-440). New York
3. A. S. Holevo, “On asymptotically optimal hypotheses testing in quantum statistics”, Teor. Veroyatnost. i Primenen., 23:2 (1978), 429–432; Theory Probab. Appl., 23:2 (1979), 411–415