PROBABILITY THEORY AND INTRODUCTORY STATISTICS

ALY6010, FALL 2019

MODULE 5 PROJECT ASSIGNMENT

TWO-SAMPLE CONFIDENCE INTERVALS & HYPOTHESIS TESTING

SUBMITTED BY: RISHHIKUMAR JANAKIRAMAN, SHIVANI ADSAR, AAYUSH SHETTY
NUID: 001068500, 001399374 , 001491160
SUBMITTED TO: DR. DEE CHILUIZA REYES
DATE: 10/21/2019

Probability Theory and Introductory Statistics

Introduction

The assignment aims at performing hypothesis with two population parameters and estimates them with statistical methodology. Using the given data, we have performed the hypothesis testing on two samples to estimate and test a hypothesis regarding some parameters like population mean, population proportion and population standard deviation. The parameters like Test Statistic, P-value and Significance level by using Null Hypothesis and Alternative Hypothesis has led to the decision of either rejecting the null hypothesis or not rejecting it.
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 (1). The alternative hypothesis, symbolized by H1 or Ha, 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 average time a person spends working each week equal to 10 hours, will he earn an increment of more than $5000 can be determined.

The confidence interval for two independent samples will be computed using either the Z or t distribution for the selected confidence level and the standard error of the point estimate. The use of Z or t depends on whether the sample sizes are large (n1 > 30 and n2 > 30) or small. The standard error of the point estimate will incorporate the variability in the outcome of interest in each of the comparison groups. If we assume equal variances between groups, we can pool the information on variability (sample variances) to generate an estimate of the population variability (3). Therefore, the standard error (SE) of the difference in sample means is the pooled estimate of the common standard deviation.

Analysis

Part 1: Hypothesis Testing

Q1. Confidence intervals for two different sample sizes of 500 for Domestic Migration Rate (DMR) and sample sizes of 650 for International Migration rate (IMR)

On the basis of the samples – Domestic Migration rate and the International migration rate, samples with sizes 500 and 650 were obtained using random sampling from Data analysis tool Pak in Excel. Table 1 shows the Mean, Variance and Standard Deviation of the population. Table 2 indicates mean, standard deviation, sample size and variance for the samples. It can be observed the mean changes for both the population, however small variation is observed in the standard deviation.

Table A		
	DMR	IMR
Population Mean:	0.92	1.24
Population Variance:	119.92	3.60
Population Standard deviation	10.95	1.90

Table A: Population statistics for both Domestic Migration Rate and International Migration Rate

Table B		
	DMR	IMR
Sample Size	500.00	650.00
Sample Mean (DMR: μ1, IMR: μ2)	1.59	1.14
Sample Variance	131.82	2.56
Sample Standard deviation	11.48	1.60
Sample Means Difference	0.4496	
Sampling Error	0.5173	

Table B: Sample statistics for both Domestic Migration Rate and International Migration Rate

Table C: Denotes the margin error calculated for the samples of Domestic Migration rate and International Migration rate. Considering the sample sizes are >= 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 $E = z_c \sigma_{\bar{x}1-\bar{x}2}$, where Zc is the Critical Value. An average margin error of 0.16 was calculated for 90% interval with an increasing trend for greater confidence intervals. We have also calculated the decision-making parameters like the Test statistic, P values which helps us to decide rejecting null Hypothesis.

Table C					
Confidence Level CI	z_c	Margin of Error	CI Lower Limit for the Means Difference	CI Upper Limit for the Means Difference	CI Width
90%	1.644853627	0.851	-0.401	1.300	1.701666986
95%	1.959963985	1.014	-0.564	1.463	2.027661278
98%	2.326347874	1.203	-0.754	1.653	2.407

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

Table D	
Test Statistic:	0.8692
P-value	0.807625251
Significance Level:	0.1

Table D: Shows the Hypothesis Testing

- Here, the Null Hypothesis, can be represented as, $\mu_1 - \mu_2 \geq \mu_0$ and the alternative hypothesis can be noted as, $H_A: \mu 1 - \mu 2 < \mu 0$. So, the problem is a Left-Tailed problem.
- On the basis of the Decision, Method 1: Reject H_0 if sample z $\leq$ Critical z*, and,

  Method 2: Reject H_0 if P value $\leq$, the null hypothesis is not rejected.
- The problem is a Left-Tailed problem using the Z-test for two samples. In this case, we are not rejecting the Null Hypothesis because the P value is greater than the Significant level which we have assumed as 0.1.
- The z test is a statistical test used to determine whether two population means are different when the variances are known and the sample size is large.
- The Confidence intervals can now be constructed for the inequality conditions and it is observed that the graph shows the skewness aligned to the left side. The left skewed distribution also known as the negatively skewed distribution because its long tail is on the negative direction.

Q2 Confidence intervals for 1000 random sample sizes for the Domestic Migration rate and the International Migration rate

The random samples of size 1000 have been calculated based on the Proportion of Success from the overall population. Table A shows – the Proportion of Success which is calculated from the Population numbers for both DMR and IMR which are greater than 9 and 2 respectively. Table B shows – the Proportion of Success which is calculated from the Sample numbers for both DMR and IMR which are greater than 9 and 2 respectively

Table A		
	DMR (*)	IMR (**)
Population Size	2431	2431
Population number of Success	471	442
Population proportion of Success	0.1937	0.1818
Population proportion of Failure	0.8063	0.8182

Table A: Location Quotient for the given Population

Table B		
	DMR (*)	IMR (**)
Sample Size	1000	1000
Sample number of Success *	200	181
Sample Proportion of Success	0.2000	0.1810
Sample Proportion of Failure	0.8000	0.8190

Table B: Location Quotient for the given Samples

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 0.0176 for the population data. The Confidence Interval is computed from the statistics that might contain the true value of an unknown population parameter.

Table 1	
Difference of Population Proportions	**0.0119**
Difference of Sample Proportions	0.0190
Sampling (standard) Error	0.0176

Table 1: Shows the difference of population proportions for both the population and the samples for the Domestic Migration rate and the International Migration rate also the Standard sampling error is calculated to be 0.0176.

We have used the formula, $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$ to calculate the population proportion for the samples.

This hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the population proportions. The difference of two proportions follows an approximate normal distribution. The larger the margin of error, the less confidence one should have that reported results are close to the "true" figures; that is, the figures for the whole population.

Q3. Hypothesized Sampling for the entire population of DMR and IMR against very small sample sizes

On the basis of the Domestic and International Migration rates population, the sample of 2431 are considered and On the basis of samples, 30 and 40 samples each are selected from the entire population of DMR and IMR.

Table 2		
	DMR	IMR
Sample Size	30.00	40.00
Sample Variance	68.69	6.90
Sample Standard deviation	8.29	2.63
Degrees of Freedom (DF)	29	39

Table 2: Location Quotient for the given Population

From the above Table 2, we have calculated the sample size, variance, standard deviation and the Degree of freedom for both the samples.

Table B. Hypothesis Testing.			
	Parameter:	Inequality Type	Hypothesized Quotient
Null Hypothesis H_o in terms of Variance:	σ_1^2/σ_2^2	=	1
Alternative Hypothesis H_a in terms of Variance	σ_1^2/σ_2^2	≠	1

Table B: Hypothesis Testing condition for the Samples in terms of variance

Table B shows the condition that we have assumed to consider the Null and alternate hypothesis based on the parameters $\sigma 1^2/\sigma 2^2$. From the calculations its seen that the distribution is Two-Tailed problem using F-Test hypothesis.

The Two-Tailed test is more conservative than a One-Tailed test because a Two-Tailed test takes a more extreme test statistic to reject the null hypothesis.

Table C : Essential parameters for rejecting the Null hypothesis	
Test Statistic:	9.9529
P-value	2.60585E-10
Significance Level:	0.0500

Table C: Essential Parameters for rejecting the Null Hypothesis for the given Population

- Here, the Null Hypothesis can be shown as, Ho: $\sigma 1^2/\sigma 2^2 = 1$ and the Alternative Hypothesis can be shown as,
- Ha: $\sigma 1^2/\sigma 2^2 \neq 1$. In this case, we have used the F-test since the variances of the two populations are equal. Based on the calculated Statistical parameters for the samples, we have determined the Test statistics and P value which are some of the significant parameters to make a decision in the Hypothesis testing.

- On the basis of decision of, Method 1: Reject H 0 if sample f ≥ Critical F* and

  Method 2: Reject H 0 if P value ≤ α ,
- We have also calculated the Critical value which is 1.9619 for the samples and using all the calculated parameters, we can conclude that we are rejecting the Null Hypothesis, Ho – Since the P value is less than the assumed significant value 0.05, also the critical value 1 is less than the Test statistic.

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.
3. In a left-tailed hypothesis test, we choose one direction for our alternative hypothesis: either hypothesize that the test statistic is "significantly big", or that the test statistic is "significantly small"
4. In a two-tailed hypothesis test, our alternative hypothesis encompasses both directions: we hypothesize that the test statistic is different from the predicted value
5. Using the Left-Tailed and the Two-Tailed Hypothesis, we have calculated the parameters which helps us to decide whether to accept or reject the null hypothesis

Hypothesis testing plays a crucial role in making business or strategic decisions. Suppose you are training your outside sales force, and want to know whether a specific sales technique results in a higher close ratio than the methods currently employed by your company. Your null hypothesis would be that the new technique has no effect on sales that isn't explained by random chance, while your alternative hypothesis would be that the method does have an effect, whether positive or negative. If you conclude that the technique has an effect, and it is positive, then you can implement the new method with confidence, knowing it is likely to bring you results.

In treatment of certain medical conditions, it helps us to decide if certain treatments have positive effects, 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. Bruce E. Trumbo California State University, Hayward *Journal of Statistics Education* v.3, n.2 (1995) Retrieved from https://www. http://jse.amstat.org/v3n2/trumbo.html
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