Revision Summary for Unit 1 Paper Version 2

Tyler Aman

The first major revision to this paper was done to all sections of the paper mainly a general overview to correct typos located within the paper to address Tim's concerns from the first submission. Next, some changes to captions and titles for the results section tables were made to clarify the tables usages and understanding for the reader at a glance, as also noted as a concern for Tim. After, small changes were again made to the introduction to make sure there was sufficient background information and attempted to mimic slightly a definition style introduction due to allowing readers to get the necessary background information on the subject of economic mobility and the more general concept of the "American Dream" and defining terms to allow easier reading later on, and introduction of the data being used as suggested by chapter 5 of H. J. Tichy's *Effective Writing for Engineers, Manager, Scientist* (Tichy, 1988). The discussion section was not changed after this due to it already discussing how both questions fared under analytical testing used in the paper as new guidelines suggests

Stat 485 Unit 1 Paper Version 2

Tyler Aman

Introduction

The “American Dream” is a concept that many immigrants, and those born in the United States have been chasing since the founding of this country and is a source of both pride and hope. This “Dream” or the ability to start from poor circumstances and improve your position in life is something Americans across all backgrounds strive to accomplish, also known as upward economic mobility. However, with so many people willing to risk everything to achieve this dream, the question becomes how feasible is the “American Dream” to to those who start with nothing? This paper looks to investigate the overall upwards economic mobility of the American populus using a subset of data gathered by the United States Department of Agriculture. This data will be used to answer the overall research question of what does economic upwards mobility in the United States look like.

This question will be approached in two parts: First, the researchers will look at what are the estimated upward mobility proportions of populations in 40 different cities across the United States, split across the four census region: Midwest, Northeast, South, and West. This will be done by showcasing the estimates of this proportion gained from previous analysis of this data, and the 95% confidence intervals of these estimates, to get an idea of what upward mobility looks like across the United States. The second part of this question will look to see if there are differences in upward mobility within certain regions as well as between the four regions. This

will be done through testing, using statistical methods, two hypotheses: First, is there a significant difference in upward mobility among the cities being sampled within a region. Second, that between the four regions there is no significant difference in regional upwards mobility. Some findings of interest to the reader possibly are that living in different cities, larger or smaller, could hurt or help a citizen's overall upward mobility chances. On the next scale up, another possibility is that living in a specific region as opposed to a singular city could affect your overall opportunity to move up economically.

Methods

The data being used in this study consist of samples of subject from 40 different commuting zones or local labor markets around the United States as defined by the United States Department of Agriculture ("Commuting Zones and Labor Market Areas"). The zones selected for use were selected using a probability proportional to size sampling scheme, where the largest zones were select with probability 1. Within these zones a subsample was selected from the subpopulation of the United States of individuals born between 1980 and 1982 living in that commuting zone at age 16. The selection process for these individuals to be subsampled was done uniformly at random using a simple random sample of this target population for each commuting zone. The sub-sampling size was completed such that which a given region each member of the cohort had an equal probability of being selected. Although, overall the sizes of the cohort in in the West and Midwest regions was similar, the South had a 50% larger cohort while the Northeast at a 20% smaller cohort. This leads our sample to accurately mimic the subpopulation in each region, however it leads to a skewing of the overall United States 1980 -1982 cohort.

Next, the income of all the individuals' families was for their 16th year, between the years 1996-1998, was recorded from federal tax records. Finally, The individuals personal income in 2010 was again determined using tax records. Besides name of zone, state, and census region, the data includes the overall sample size of each zone, the number of individuals who during their 16 year had household incomes in the lowest quantile between 1996-1998 (n.lowstart), and the proportion of these subsets of the sample who in that zone during 2010 had incomes in the highest quantile of individual incomes (p.upmover). Other variables added by the researchers include X which was calculated multiplying n.lowstart by p.upmover, to get the number of individuals who by 2010 had were in the highest quintile and labelled as such for ease of calculations later on in analysis. The data also included positions to hold the upper and lower bounds of the individual zones confidence intervals. Next, for testing purposes five proportions were calculated using the full data and subsets: a common proportion for each of the 4 census regions was done by subsetting the data, and then calculating the proportions using the sum of the X values listed for the region divided by the sum of all the values of n.lowstart, and an overall common proportion was calculated using the same method except using all data points not the subsets of each region. Finally, also for testing two vectors were created holding the four regions total X values by taking the sum of the X's in each region, and the total n.lowstart in each region by the same method.

To answer the first of the two parts of the question the researchers calculated 95% confidence intervals for the individual zones using the Clopper - Pearson Confidence interval method also known as the "exact" confidence interval (Brown, Cai, Dasgupta, 2001). This method was used as opposed to the standard Wald confidence interval due to its' lack of

restrictions on the data being used, accuracy in terms of coverage, defined as the actual rejection/acceptance area of the test for which confidence interval is the mathematical inverse of, as there is problems with the Wald interval in both of these areas, and its ease of calculation from a computing perspective, and it conservativeness in that its coverage probability being always greater than or equal to the nominal confidence level (95% in this case) (Brown, Cai, Dasgupta, 2001). Other arguments for use of this interval have been made by Dr. Ben Hansen of the University of Michigan in his January 18, 2018 lecture titled "Unit 1, 5th & last meeting" (Hansen, 2018).

To test the two hypotheses presented a Likelihood Ratio Test was used who's methods are described in detail in the appendix of this paper. For the individual regions test the null hypothesis assumed a common regional proportion for upward mobility. Then, for the overall test a common overall proportion for upward mobility. This method was used due to it popularity in statistics in general as well as it ease of calculation and quick computation. Another reasoning for using this test is its use of maximum likelihood estimation in its testing procedure as well as likelihood equations to be used in comparison.

Results

The results of the confidence interval portion of the 40 zones is displayed below in 5 tables:

Tables 1 - 5 - Clopper Pearson 95% confidence intervals for individual zones

	lower	p.upmover	upper
Phoenix	0.007	0.061	0.202
Bakersfield	0.016	0.075	0.204
Fresno	0.028	0.083	0.184
Sacramento	0.054	0.143	0.285
San Francisco	0.039	0.138	0.317
San Diego	0.007	0.054	0.182
Los Angeles	0.068	0.122	0.196
Portland	0.012	0.1	0.317

	lower	p.upmover	upper
Houma	0.051	0.133	0.268
Miami	0.005	0.037	0.127
Greenville	0	0	0.08
Chincoteague	0.017	0.053	0.119
Tulsa	0.021	0.075	0.182
Houston	0.001	0.022	0.115
Wichita Falls	0.064	0.167	0.328
Tyler	0.049	0.119	0.229

	lower	p.upmover	upper
Eugene	0.035	0.125	0.29
Seattle	0.063	0.185	0.381
Detroit	0.001	0.029	0.149
Saginaw	0	0	0.119
Lorain	0.01	0.083	0.27
Cleveland	0.024	0.115	0.302
Columbus	0.001	0.053	0.26
Iowa City	0.002	0.067	0.319

	lower	p.upmover	upper
Cedar Rapids	0.017	0.133	0.405
Peoria	0.001	0.056	0.273
Chicago	0.007	0.061	0.202
St. Louis	0.006	0.047	0.158
Erie	0.017	0.081	0.219
Oneonta	0.036	0.109	0.236
Scranton	0.011	0.087	0.28
Harrisburg	0	0	0.154

	lower	p.upmover	upper
New York	0.035	0.079	0.15
Newark	0.019	0.088	0.237
Philadelphia	0.015	0.07	0.191
Boston	0.001	0.032	0.167
Manchester	0.001	0.042	0.211
Bridgeport	0.001	0.053	0.26
Winston-Salem	0.02	0.094	0.25
Charlotte	0.001	0.023	0.12

Tables 1 - 5 each include the Clopper - Pearson 95% confidence interval bounds and the expected proportions for each zone. due to size limitations in R tables are not split directly across zones

The tables were seperated due to display limitations within the statistical program uses. Some will notice that a few entries have a lower bound of 0 which is also the same as the expected proportions of 0 this is due to the subsample finding zero individual who had moved upwards in the economic sense and that zero is the floor of the proportions limits. It is also interesting to note that some estimates and their respective confidence intervals for larger cities seem to be on the lower end of the spectrum with cities like Detroit and Houston having confidence intervals going from almost 0 to less than .15 possibly showcasing difficult mobility in those cities.

Next, five likelihood ratio tests were conducted to test the second part of the question. The results of the first four can be broken down here:

Table 6- Likelihood Ratio Tests results and confidence intervals

	LRT_Stat	p_values	lw_bound	prob.all	up_bound	df
Midwest	6.936	0.644	0.023	0.058	0.116	9
Northeast	5.828	0.757	0.035	0.071	0.123	9
South	20.292	0.016	NA	0.069	NA	9
West	5.619	0.777	0.065	0.108	0.165	9
Overall	6.313	0.043	NA	0.083	NA	2

Table 6 showcases the results of the five Likelihood ratio tests conducted with calculated 95% Likelihood Ratio Confidence intervals for those results which supported the null hypothesis of a common proportion and df of each test

For three of the four regions the results support the null hypothesis that over the 10 commuter zones for each region there appears to be a common proportion which I have included for convenience along with their calculated 95% confidence interval. The interesting finding is that the Southern region does not support the original research hypothesis that across commuter zones there is a common proportion of upward mobility for its region (LRT = 20.292, p value = .016, chi-squared df = 9) meaning that there are differences within the southern region of upward mobility which could be the result of Greenville, South Carolina being within that subset having a upward mobility proportion of 0. Next, excluding the Southern region a overall test was conducted to see if there was a common proportion for the three region: Midwest, Northeast and West. This test all showcased evidence against the original research hypothesis of a common upward mobility (LRT = 6.313, p value = .043, chi-squared df = 2) meaning that even among the regions that showcase a common proportion for upward mobility in the region, there is still differences among the regions in terms of upward mobility with the West region showing the highest proportion.

Discussion

The results of this study showcase interesting findings for the methods being used. For the first section of the research question, although criticized within the Brown, Cai, Dasgupta paper the Clopper-Pearson confidence interval showcases very good coverage for summarizing the upward economic mobility across the 40 commuter zones sampled in this analysis. Although, it is hard to discuss without looking at other confidence interval estimates for this data these estimates most likely showcase conservative estimates of these intervals compared to the true 95% confidence interval, which although this research values others may criticize or suggest other methods such as the Jeffery's interval as an alternative. However, the Jeffery's interval with proportions close to 0 or 1 still showcases more sporadic behavior while the Clopper-Pearson interval appears more steady when comparing graphs in the Brown, Cai, Dasgupta paper on Binomial confidence intervals(2001). Next, the likelihood ratio test produced mixed results about the second part of the research question being investigated. Although, most commuter zones within regions showcased similar proportions for upward mobility, the Southern region did not. Also, when comparing the three regions there appeared to be differences among the three in terms of economic mobility. This second result however, may be due to problems with sampling and not issues with the test because the Northeast region is under sampled in terms of the overall United States population therefore we could possibly be seeing a biased result when attempting to combine the three regions into one when compared to the true population of the United States. However, even with these shortcomings the Likelihood Ratio Test when used properly can be a powerful testing to when dealing with binomial data and the confidence interval used with some modification can also be a powerful summary tool.

References

Commuting Zones and Labor Market Areas. (n.d.). Retrieved January 19, 2018, from https://www.ers.usda.gov/data-products/commuting-zones-and-labor-market-areas/

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical science*, 101-117.

Hansen, B. (Y2018). *Unit 1, 5th & last meeting* [PDF presentation]. https://umich.instructure.com/courses/194528/files/folder/230-4section?preview=6875129

Tichy, H. J. (1988) *Effective Writing for Engineers, Manager, Scientist* New York CIty, New York: John Wiley & Sons.