North West Los Angeles' Average Price of Coffee in Licensed Establishments

By Courtney Engel, Natasha Ericta and Ray Luo
Statistics 201A Sample Project
Professor Xu

December 14, 2006

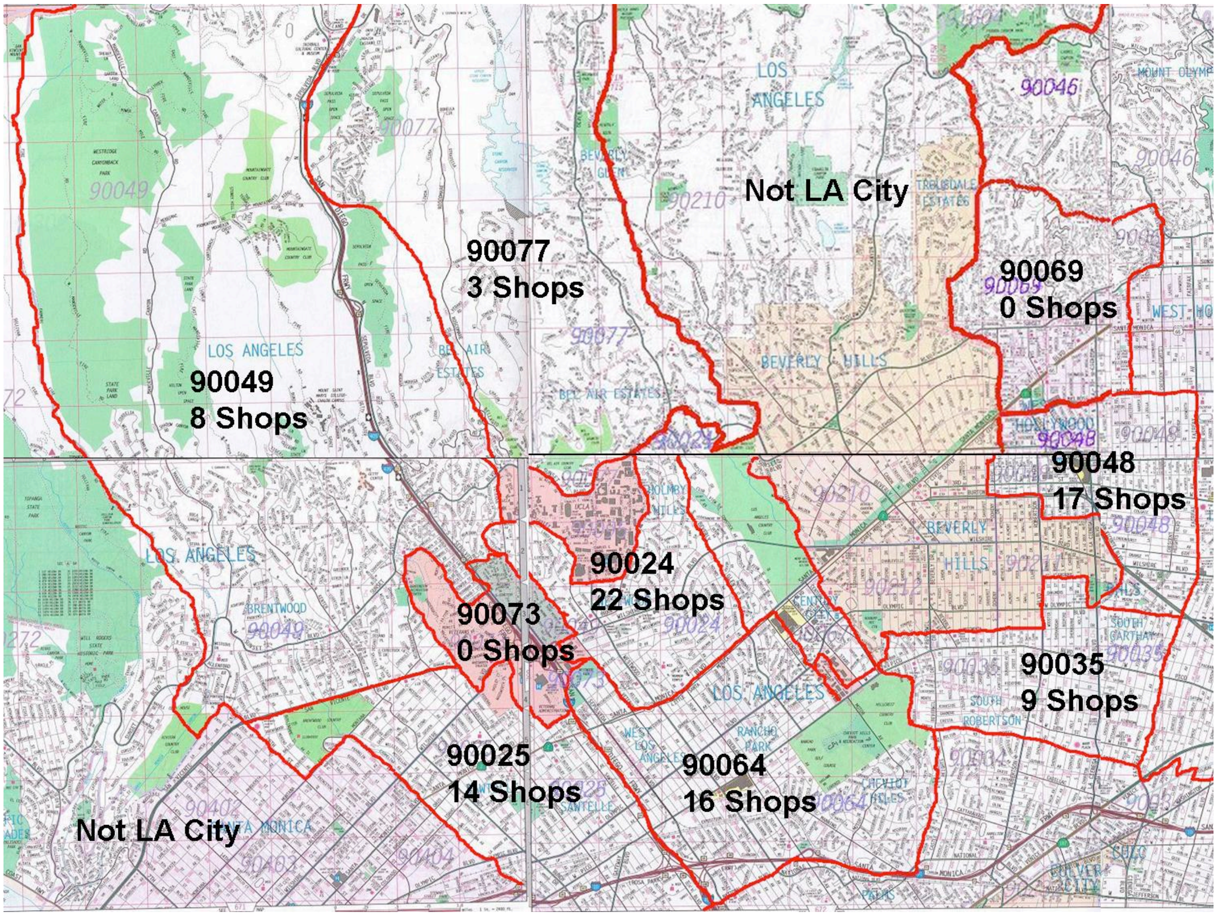
1 Background and Objectives

The purpose of this study focuses on the average price of coffee served in licensed establishments across the area considered as North West Los Angeles (NWLA). Coffee is a product of high demand and its consumption, a popular activity for many. LA City's community seems to view it as a necessity, commodity and leisure activity. Assuming its significant position in LA's marketplace, we wish to determine the average price of coffee and to apply this knowledge to real life choices. In other words, as students living in our chosen area, we hope to see how "expensive" or "inexpensive" an establishment in NWLA prices their coffee relative to the entire region.

2 Population and Multi-Stage Sampling

We determined that Multi-Stage sampling would be appropriate in our project since NWLA is partitioned into several zip codes. These zip codes are chosen to be the primary units since they best define the different areas of NWLA. Since no one primary unit can be representative of the entire region, multi-stage sampling not cluster sampling should be implemented here. Also, zip codes have different economic backgrounds and tax regulations which would make the difference in coffee prices between zip codes much different than within each zip codes.

Using a Thomas Guide map with zip code borders, we found which zip codes fulfilled our locations of interest. Then we contacted the LA City Tax office and received the records of all eateries in the specific zip codes. Each eatery is required to register with the government for tax purposes; therefore, this gave us the most complete list of possible units for the analysis. A map of the chosen locations can be seen below.



Notice that 90073 is the Veteran's Hospital and will not have any coffee shops.

The obtained report consisted of 756 eateries in PDF format. We, then, parsed the data using Python to aggregate a list of names, addresses and zip codes of each eatery. From here, we filtered through the list by researching online, phone calls or visits to each eatery to verify whether it falls under our description of a coffee establishment (i.e. café, coffeehouse, bakery, doughnut shop excluding all waited-on restaurants). The adjusted total τ consisted of 102 coffee-shop type places. Once we had a complete count, we divided the population into primary units / zip codes N. Each primary unit contains a total number of secondary units M_i as shown below:

Zip Codes	90024	90025	90035	90048	90049	90064	90067	90069	90073	90077
Units (M_i)	22	14	9	17	8	16	13	0	0	3

Method 1: From the $N = 8$ zip codes with coffee shops, we used simple random sample (SRS) without replacement to get $n = 5$ distinct primary units (PUs). We planned on using a total of $\sum_{i=1}^{n} m_i = 35$ total samples, because this is usually the sample size required for normality in basic SRS designs. This assumption needs to be checked when fitting a model in the future. Next, we use proportional allocation to decide how many samples to obtain at each PU. The idea is that we want to sample larger PUs more often. The selection of secondary units (SUs) still procedes by without replacement SRS. What we are doing here is assuming that the PUs are set given the primary stage selection, much like strata in stratified random sampling. Given these PUs, we allocate the $m_i = 35\frac{M_i}{M}$ proportionally amongst the PUs that were selected. The M_i, M are known from the complete data set. We used proportional allocation because the variances of the population needed for optimal allocation were not available. This is two-stage design with SRS at each stage (we excluded primary units which had no coffee shops-90069 and 90073). Prices of a 12oz

cup of house coffee and a 12oz cup of café latte are:

Zip Codes	90024	90025	90035	90048	90049	90064	90067	90069	90073	90077
Units (M_i)	22	14	9	17	8	16	13	0	0	3
Sample (m_i)	10	6	4	8	X	7	X	X	0	X

Method 2: We also wanted to compare the allocation strategy in method 1 with a probability proportional to size (PPS) sampling of PUs, followed by SRS of the selected PU. Thus the selection of the PUs procedes by the Hansen-Hurwitz method, with $p_i = \frac{M_i}{M}$. In this strategy, we take 35 independent samples of PUs with replacement using proportional to size selection. Then we take a single random sample from the PU selected. The hope is that we reduce the variance of the estimator within the PU by using an appropriate number of samples for larger PUs. Note that this approach differs from method 1 in that all PUs are equally represented across the two stages.

We first generate a random number uniformly from 0 to $\sum_{i=1}^{N} M_i = 102$. Then we assign PU_i using the cumulative intervals in the sum. For example, if the random number is greater than M_2 and less than M_3, then we select PU_3 (90035). Then we choose one sample at random from that zip code. Note that repeats are possible in this design, as seen in our data. We wanted to compare the PPS approach with the proportional allocation approach in the SRS design using the same number of total samples. We expect the PPS estimator to work well when the total $\hat{y}_i$ for each PU_i is proportional to M_i. The Hansen-Hurwitz estimator is known to be unbiased for this with-replacement scheme (Thompson, 2002). Note that the cost and sample size formulas from Thompson cannot be used in this case because the m_is are not constant, and the variances are not known.

Zip Codes	90024	90025	90035	90048	90049	90064	90067	90069	90073	90077
Units (M_i)	22	14	9	17	8	16	13	0	0	3
Sample (m_i)	8	4	7	4	2	6	3	0	0	1

The data for both methods are attached in Figures 1 and 3 in Appendix. The plots of the data is attached in Figure 2 and 4 of the Appendix.

3 Analysis and Estimation

When looking at the plot for the first method (Figure 2), the sample shows that the means and medians are very close with a slight outlier in 90064. We want to estimate μ and determine its standard error:

Unbiased Estimator:

Using SRS at both stages and proportional allocation of SUs, we find that the unbiased estimator gives a result of $\hat{\tau} = \$200.32 \pm 20.26$, and $\hat{\mu} = \$1.626 \pm 0.16$. The formulas are as follows.

$\hat{\mu} = \hat{\tau}/M$

$M = 102$

$\hat{\tau} = \frac{M_i}{m_i} \sum_{j \in s_i} y_{ij}$

$var(\hat{\mu}) = var(\hat{\tau})/M^2$

$v\hat{a}r(\hat{\tau}) = N(N-n)\frac{\sigma_\mu^2}{n} + \frac{N}{n} \sum_{i=1}^{N} M_i(M_i - m_i)\frac{\sigma_i^2}{m_i}$

Results:

$\hat{\tau} = 200.32$

$v\hat{a}r(\hat{\tau}) = 410.59$

$s_e = 20.26$

$\hat{\mu} = \$1.626$

$v\hat{a}r(\hat{\mu}) = .027$

$s_e = .164$

Ratio Estimator:

The ratio estimator gives a result of $\hat{\tau}_r = \$200.32 \pm 6.02$, and $\hat{\mu}_r = \$1.626 \pm 0.049$. We suspect that the standard error is so low because the m_i correlate with the y_i, because our proportional allocation assigns more units to the bigger PUs, leading to a bigger sum over SUs in those PUs. The formulas are as follows.

$\hat{\tau}_r = r \times M$, $r = \frac{\sum_{i \in s} \hat{y}_i}{\sum_{i \in s} M_i}$

$\hat{\mu}_r = \frac{\hat{\tau}_r}{M} = r$

$v\hat{a}r(\hat{\tau}_r) = N(N-n)\frac{\hat{\sigma}_r^2}{n} + \frac{N}{n}\sum_{i \in s} M_i(M_i - m_i)\frac{s_i^2}{m_i}$ where $\hat{\sigma}_r^2 = \frac{1}{n-1}\sum_{i \in s}(\hat{y}_i - rM_i)^2$

$v\hat{a}r(\hat{\mu}) = var(\hat{\tau}_r)/M^2$

Results:

$\hat{\tau}_r = 200.32$

$v\hat{a}r(\hat{\tau}_r) = 36.29$

$s_e = 6.02$

$\hat{\mu}_r = 1.626$

$v\hat{a}r(\hat{\mu}) = .00239$

$s_e = .049$

Here we see that the Ratio Estimator has a smaller standard error than the unbiased estimator. This shows that there is a correlation between the M_i and the y_i values.

When looking at the plot for the second method (Figure 4), the sample shows a lot more variation between zip codes and more difference in means and medians. It shows that prices for shops in the 90024 and 90035 area codes appear to have large variance. However, as the scatter plot shows, much of the variance may be explained by a single sample in

90024 that netted \$3. The means look to be highest for shops in 90064 and 90049, but we don't have enough samples to determine this without fitting a model.

Probability Proportional to Size:

Using PPS sampling of PUs and SRS of SUs, we used the Hansen-Hurwitz estimator to find $\hat{\tau}_p = \$161.69 \pm 6.08$, and $\hat{\mu}_p = \$1.59 \pm 0.060$. Note that this is a different sample from the above, although some of the same units are chosen after randomization based on what was sampled before. That is, we generate a random number to select randomly from each PU, but the SU selected chosen is based on an assignment of the samples we have, unless a new unit is needed, in which case we called (this happened five times total, three of them for the 90067 code, which didn't appear in the first method.) The PPS estimator also has low variance, due to the proportionality between p_i and y_i. The formulas are as follows.

$p_i = M_i/M \ p_i = .22, .16, .14, .09, .16$

$\hat{\tau}_p = \frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i} = \frac{M}{n}\sum_{i\in s}\frac{y_i}{M_i}$

$v\hat{a}r(\hat{\tau}_p) = \frac{1}{n(n-1)}\sum_{i\in s}(\frac{y_i}{p_i} - \hat{\tau}_p)^2 = \frac{M^2}{n(n-1)}\sum_{i\in s}M_i(\bar{y_i} - \hat{\mu}_p)^2$

$\hat{\mu}_p = \frac{\hat{\tau}_p}{M}$

$var(\hat{\mu}_p) = \frac{v\hat{a}r(\hat{\tau}_p)}{M^2}$

$\hat{\tau}_p = 161.68$

$v\hat{a}r(\hat{\tau})_p = 36.94$

$s_e = 6.08$

$\hat{\mu}_p = 1.59$

$s_e = .060$

Despite the larger difference in means and medians, the standard error for PPS is still smaller than the unbiased estimator. Both method results show that using a proportional method is more appropriate in the design than SRS. Both the Ratio Estimator for the SRS design and Hansen-Hurwitz estimator for the PPS design appear to have low variance. They work because they take advantage of the proportionality of the sum of values in each PU to the number of samples either allocated or selected in that PU. The PPS design samples all the zip codes, however, while the SRS design only samples an arbitrary number of PUs specified by the design. The slightly larger variance associated with the PPS estimator may reflect this sampling of two extra PUs. Purely based on design, the PPS estimator has greater intuitive appeal, because PUs are selected based on size, rather than having the size determined post sampling at the first stage. The purely SRS design assumes that the PUs determined are equal at the first stage before allocating at the second stage, which is not a good assumption in our case. On the other hand, the PPS design implements proportional to size sampling as opposed to allocation, taking advantage of the known structure of *all* PUs, not just the ones selected in the first stage.

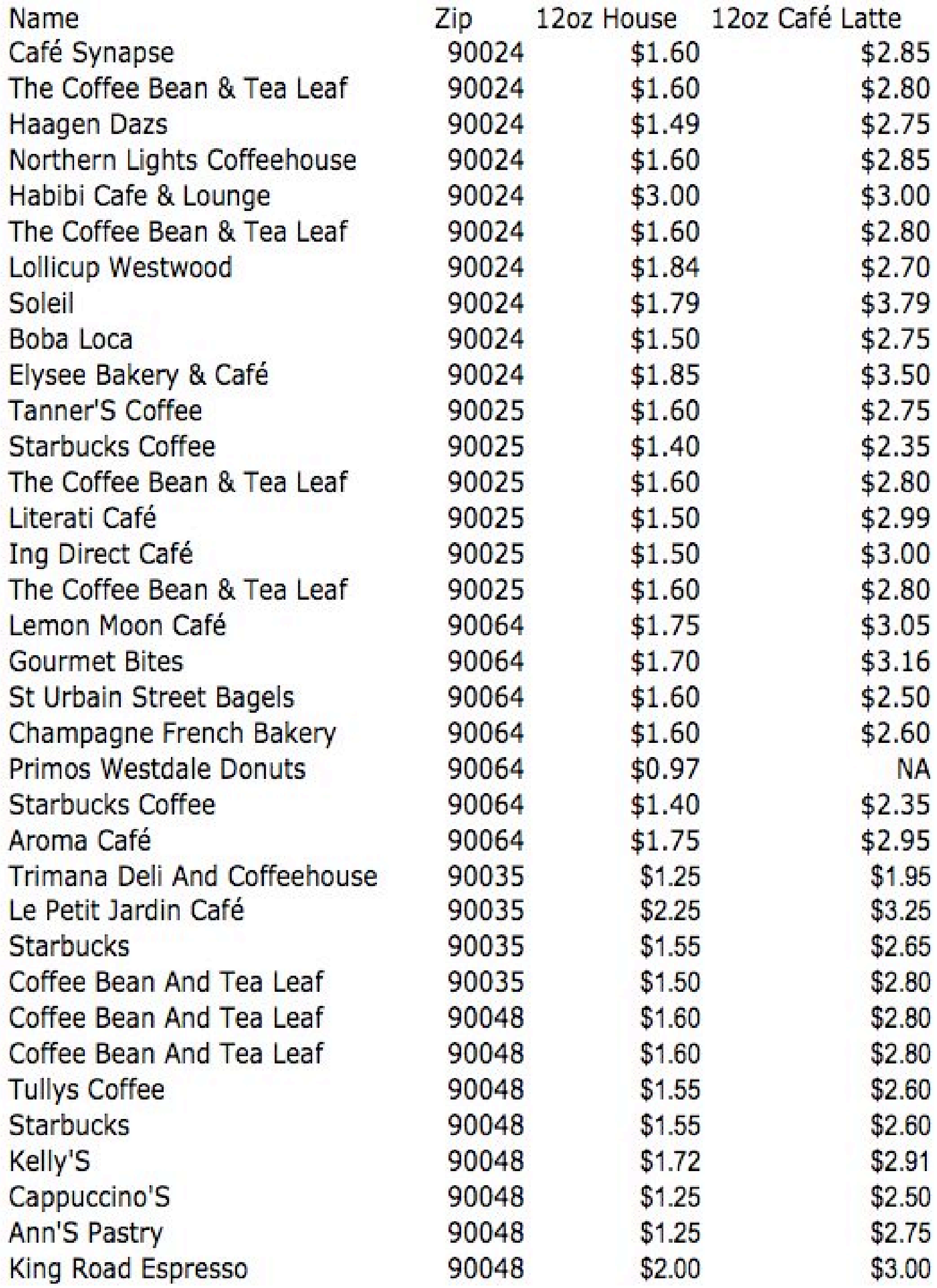
4 Conclusion

The estimated average price of a 12 oz cup of regular coffee in NWLA is according to the unbiased estimator with SRS of both PU and SU is \$1.63. Also, the average price is \$1.59 when using the Probability Proportional to size method. We can see that the estimates are not very different. In the first method, since we used proportional allocation in determining our m_i, the m_i are correlated to the M_i. If we assume that zip codes affect the price of coffee, then there is a correlation between the M_i and the y_i values. As a result, the m_i are also correlated to the y_i values; therefore the Ratio Estimator above has a smaller standard error compared to the unbiased estimator. In the second method, the result is similar, with a mean of $\hat{\mu}_p = \$1.59 \pm 0.060$. The design, however, has greater intuitive appeal, and requires one less parameter to be determined by the sampler, because the number of PUs are not determined a priori. Moreover, the estimator variance is small due to the proportionality of the sum of the prices in each zip code with the number of shops in each zip code used to determine p_i. One possible improvement on PPS is to use different numbers of samples in stage two to lower the cost of sampling. For example, if we take 40 samples, all independently selecting a PU followed by a SRS of one sample from the selected PU, then the cost is on the order of the 40 PU samples. If instead, we sample only 10 PUs, followed by SRS without replacement of 4 SUs in each PU selected, then the lower cost of the (more limited) second stage PU combined with the four fold reduction in the number of PUs selected will reduce the cost substantially. This scheme is still a PPS design, because $p_i = \frac{M_i}{M}$ still holds. However, it remains to be shown what the estimator variance of this strategy would be, because it would likely have a higher between PU variance but lower within PU variance.

When walking into a coffee shop, we can now compare the price of their 12 oz cup of regular coffee to our estimated averages. The estimates can be used as a benchmark for deciding whether that coffee shop suitably prices their coffee. This information is useful for price sensitive customers. Finally, the entire analysis can be used for various other drinks served in these establishments for other interested parties.

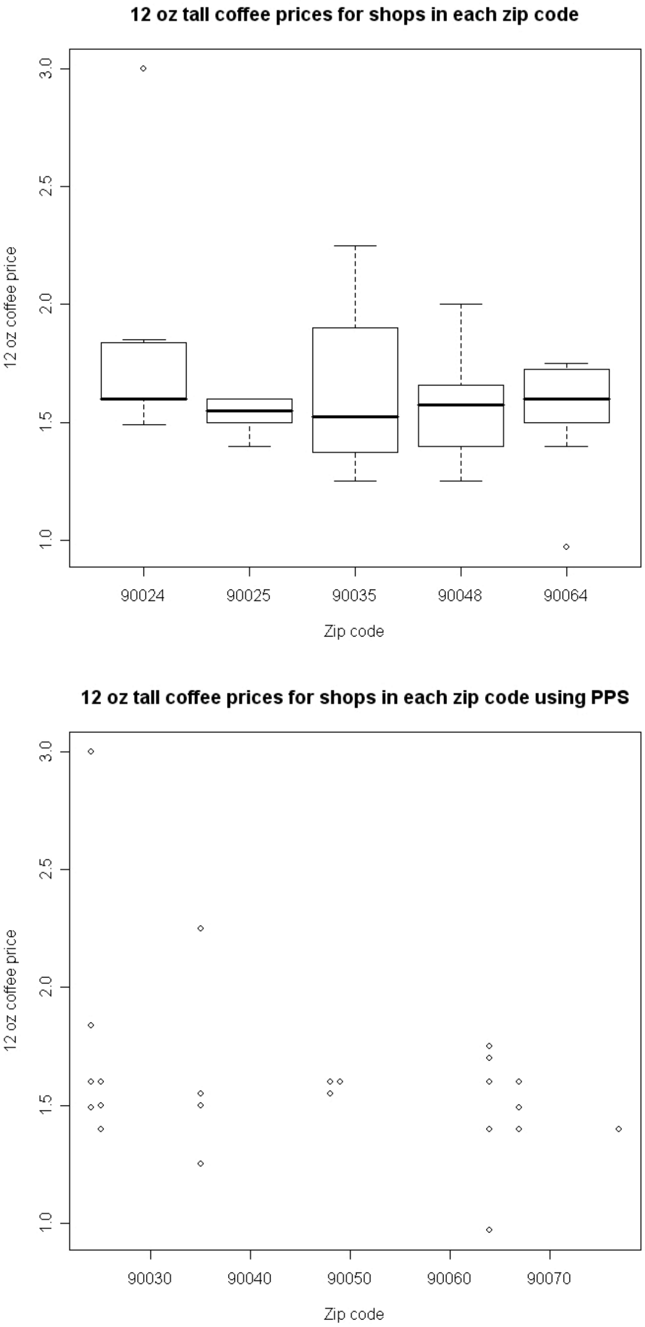
Some possible problems in our design are due to the subjectivity within our categorization. Because we did a manual check of the eateries, the category of "coffee-type" establishments is based on our opinion of what defines this category which may be different from what others consider as such an establishment. We also discovered that areas such as Beverly Hills and Pacific Palisades are considered part of this region but was excluded because they are not in LA city and have different tax regulations. If we had included these areas in our analysis, we believe that the data would be skewed due to prices being affected by tax. Also, we assume that different zip code have different economic backgrounds but the significance of these differences have yet to be determined.

Appendix



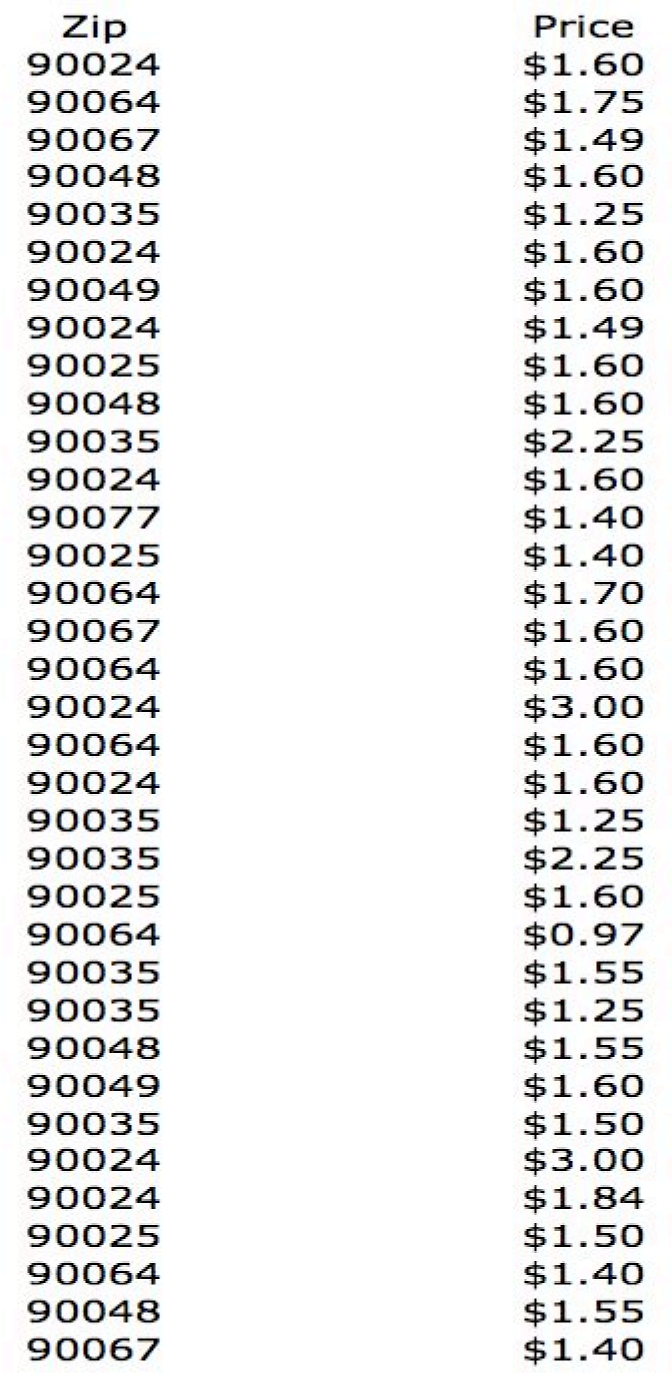
Name	Zip	12oz House	12oz Café Latte
Café Synapse	90024	$1.60	$2.85
The Coffee Bean & Tea Leaf	90024	$1.60	$2.80
Haagen Dazs	90024	$1.49	$2.75
Northern Lights Coffeehouse	90024	$1.60	$2.85
Habibi Cafe & Lounge	90024	$3.00	$3.00
The Coffee Bean & Tea Leaf	90024	$1.60	$2.80
Lollicup Westwood	90024	$1.84	$2.70
Soleil	90024	$1.79	$3.79
Boba Loca	90024	$1.50	$2.75
Elysee Bakery & Café	90024	$1.85	$3.50
Tanner'S Coffee	90025	$1.60	$2.75
Starbucks Coffee	90025	$1.40	$2.35
The Coffee Bean & Tea Leaf	90025	$1.60	$2.80
Literati Café	90025	$1.50	$2.99
Ing Direct Café	90025	$1.50	$3.00
The Coffee Bean & Tea Leaf	90025	$1.60	$2.80
Lemon Moon Café	90064	$1.75	$3.05
Gourmet Bites	90064	$1.70	$3.16
St Urbain Street Bagels	90064	$1.60	$2.50
Champagne French Bakery	90064	$1.60	$2.60
Primos Westdale Donuts	90064	$0.97	NA
Starbucks Coffee	90064	$1.40	$2.35
Aroma Café	90064	$1.75	$2.95
Trimana Deli And Coffeehouse	90035	$1.25	$1.95
Le Petit Jardin Café	90035	$2.25	$3.25
Starbucks	90035	$1.55	$2.65
Coffee Bean And Tea Leaf	90035	$1.50	$2.80
Coffee Bean And Tea Leaf	90048	$1.60	$2.80
Coffee Bean And Tea Leaf	90048	$1.60	$2.80
Tullys Coffee	90048	$1.55	$2.60
Starbucks	90048	$1.55	$2.60
Kelly'S	90048	$1.72	$2.91
Cappuccino'S	90048	$1.25	$2.50
Ann'S Pastry	90048	$1.25	$2.75
King Road Espresso	90048	$2.00	$3.00

Figure 1: Multi-Stage with SRS and Proportional Allocation


12 oz tall coffee prices for shops in each zip code

12 oz tall coffee prices for shops in each zip code using PPS

Figure 2: Multi-Stage with SRS and Proportional Allocation



Zip	Price
90024	$1.60
90064	$1.75
90067	$1.49
90048	$1.60
90035	$1.25
90024	$1.60
90049	$1.60
90024	$1.49
90025	$1.60
90048	$1.60
90035	$2.25
90024	$1.60
90077	$1.40
90025	$1.40
90064	$1.70
90067	$1.60
90064	$1.60
90024	$3.00
90064	$1.60
90024	$1.60
90035	$1.25
90035	$2.25
90025	$1.60
90064	$0.97
90035	$1.55
90035	$1.25
90048	$1.55
90049	$1.60
90035	$1.50
90024	$3.00
90024	$1.84
90025	$1.50
90064	$1.40
90048	$1.55
90067	$1.40

Figure 3: Multi-Stage with Probability Proportional to Size

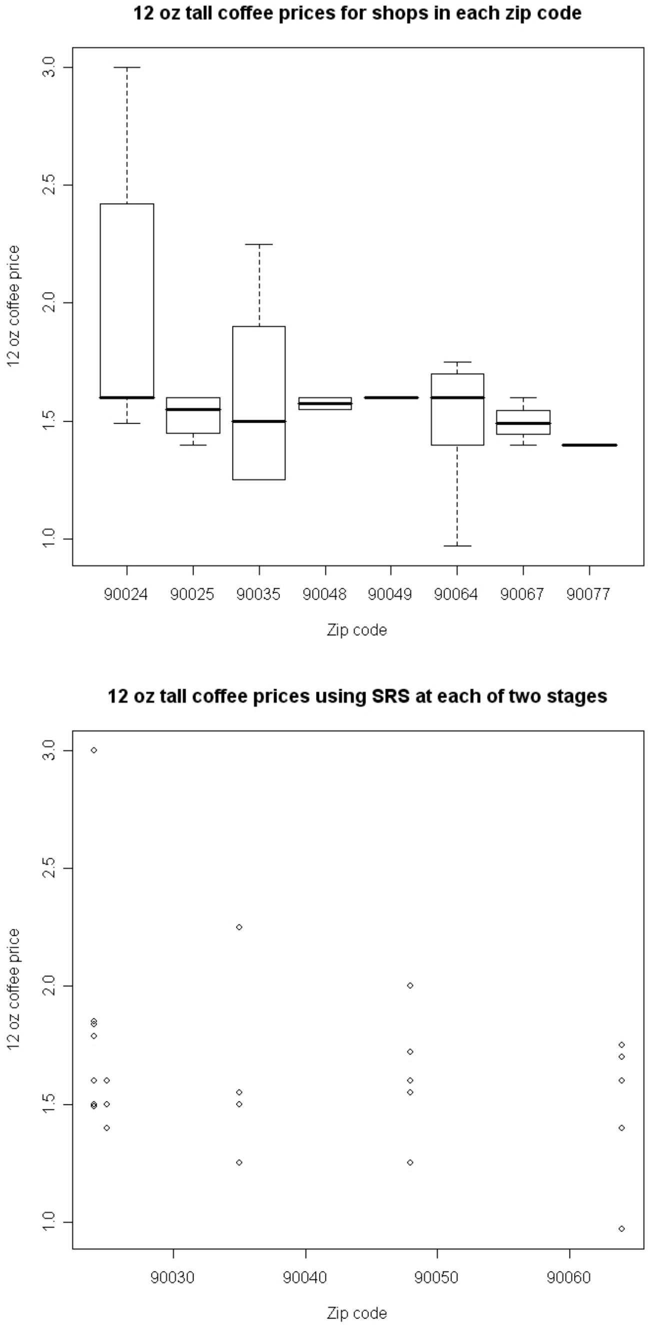


Figure 4: Multi-Stage with Probability Proportional to Size