Swaycha Goli, Maya Krishnamoorthy, Sophie Salem, Morgan Watson, and Maya Yin

BIOS 664 Project Findings - Group 7

Sample and Data Collection Experience

The goal of this research project was to estimate the proportion of available reservable study rooms in the most popular libraries on UNC's campus. We chose these libraries based on the three libraries on the UNC library website with the most reservable rooms. Specifically, the availability was determined during the hours of 8 am to 8 pm. The target population was reservable study rooms in Davis Library, the House Undergraduate Library (UL), and the Health Sciences Library (HSL). The sampling frame was chosen by generating a list of all possible reservable spaces, including study booths, in all three of the libraries for each hour between 8:00 am and 8:00 pm using the UNC Library Website. The sample was stratified by library and time of day. Time of day included 3 groups - morning (8:00 am - 12:00 pm), afternoon (12:00 pm - 4:00 pm), and evening (4:00 pm - 8:00 pm). Using the calculations from our initial design plan, we proportionately allocated a sample size to each of the libraries before stratifying on time interval. Each sampling unit was selected through an SRS of the study rooms included in the sampling frame for each time interval. Proportional allocation was utilized for the time interval stratification variable, so the sample chosen from each time interval was proportional to their stratum size. After completing sample size calculations in the project plan, the sample size for the project was determined to be 357. The sample was selected using SAS surveyselect.

The steps of sample selection did not differ from the originally proposed plan. The sample was stratified by library and time of day through proportional allocation and the ultimate sampling units were chosen through an SRS. The selection probability of each member of the sample was slightly different, ranging from 0.077 for HSL to 0.079. We expected the selection probabilities to be the same since a proportional allocation was used for the stratification resulting in an EPSEM design in the study. However, we believe that the slight difference in probabilities may be due to rounding our sample size values for each library to a whole number. Therefore, we can claim that our study design is approximately EPSEM. We did not encounter any problems in selecting the sample. The sample of reservable study spaces was split into 12 hourly time slots between 8:00 am and 8:00 pm. 24 hours before the time interval of interest, we logged into https://calendar.lib.unc.edu/reserve/ to check the availability of the study space (available = 1 and unavailable = 0). We manually entered this into our dataset in Excel.

When accounting for measurement error, we anticipated the problem of rooms being mostly unavailable if we checked on the day of interest, so we had decided in our original design plan to check the availability of rooms 24 hours before. On Fridays, all library reservations stop at 5:00 pm since all libraries close then, so we had to mark them as missing data. We decided to analyze them as missing rather than unavailable because we did not anticipate this in our sampling frame. Additionally, we had spelling errors in the time interval variable for "Afternoon," so we checked them before completing the weighting and analysis tasks.

We began our data collection with a pilot sample for Wednesday, March 20th. All of the group members collected the same sample for Wednesday at 5:00 pm to ensure consistency in the data collection process as well as in the input process into the dataset. The sampling frame was compiled into a dataset by all of the group members. Group members were paired up to enter all data points for each of the libraries (2 for Davis, 2 for UL, 2 for HSL) to ensure that our sampling frame was correct. Each library corresponds to a separate mini-sampling frame that was then used to select sampling units from each library in each time interval. The sampling task leader created the sample datasets by copying the sample dataset from SAS into Excel. Each of the members then entered the binary values 0 or 1 manually to confirm the availability of each sampled room for each day.

Weighting

Overall, the proportion of missing data was 0.0924. When looking at individual libraries, the proportions of missing data were 0.081, 0.118, and 0.088 for Davis, Health Sciences, and the Undergraduate libraries, respectively. Observing time intervals, the missing data proportions were 0.17, 0, and 0.11 for the morning, afternoon, and evening, respectively. Lastly, when we looked at missing observations for each hour interval, we noticed that 12-6 pm had no missing observations, but the other time intervals did. In summary, all variables showed different proportions in missing data.

We adjusted weights for non-response because we aimed to reduce bias due to nonresponse error, and we know the variables for both respondents and non-respondents. Since we had access to all library reservation data and chose a stratified-proportional allocation method in our sample selection, we are not as concerned about undercoverage. The underlying assumption is that nonrespondents have the same outcome distribution as respondents.

The nonresponse adjustment was done using the weighting class method. First, we assigned base weights to each data point based on their library as the base weights do not differ between time intervals. After finding the total weight in each weighting class, we calculated the weight of the respondents in each weighting class. The weighting classes were crosstabulations of the library and time interval. The non-response adjusted weights were then calculated using psi values. Psi was the proportion of weight from respondents to the total weight. The final step was merging the dataset including the nonresponse-adjusted weights with the original dataset, and calculating the final non-response adjusted weights from base weights and psi values.

Below are our weight distributions throughout the study. We aimed for an EPSEM design. However, due to a rounding error in the sample size calculation, the sampling weight ranges from 12.63 to 13.03.

Distribution of Base Weights for the Full Sample

Library	**Initial Base Weights (morning, afternoon, evening)**
Davis	13.03

UL	12.63
HSL	12.94
Total	4619.73

Distribution of Weights Without Missing Data

Library	Time of day	Weights before adjustments/unit	Weights before adjustments (%)	Adjusted Weights/unit	Adjusted Weights (%)
Davis	Morning	13.03	17.71	15.09	18.62
	Afternoon	13.03	20.51	13.03	18.62
	Evening	13.03	18.34	14.58	18.62
UL	Morning	12.63	4.82	15.00	5.19
	Afternoon	12.63	5.72	12.63	5.19
	Evening	12.63	5.12	14.12	5.19
HSL	Morning	12.94	8.02	16.92	9.52
	Afternoon	12.94	10.49	12.94	9.52
	Evening	12.94	9.26	14.67	9.52
Total		4192.82	100	4619.73	100

In comparing the sums of base weights for the full sample and adjusted weights for those without missing data, there is no difference. Both sum to 4619.73. The non-response adjusted weights are highest for Health Sciences Library in the morning because more missing data came from this stratum. It is the lowest in the afternoon because there was no missing data at these times.

Analysis

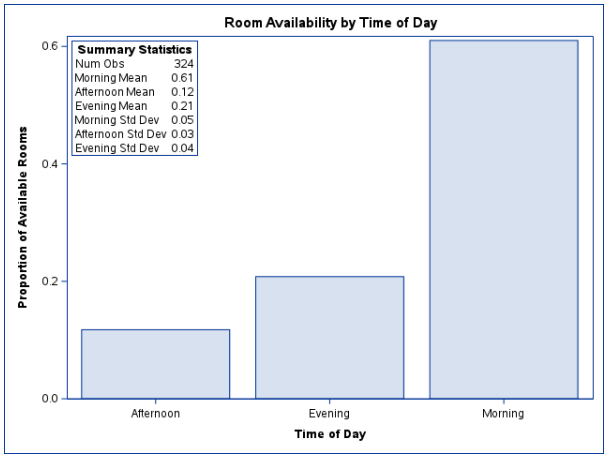
Our target population parameters were the proportion of library study rooms available, measured by reservation availability on the UNC library websites 24 hours beforehand. We found estimates for this in three libraries, reporting overall proportions, by-library proportions, and proportions by time of day. The mean availability of library rooms 24 hours beforehand is calculated in SAS as $\frac{\Sigma(w_i \times x_i)}{\Sigma w_i}$, where w is the non-response adjusted weight for each observation. To find the means, it is the same as using this formula: $\hat{\bar{y}}_{str} = \sum_{h=1}^{H} P_h \bar{y}_h$.

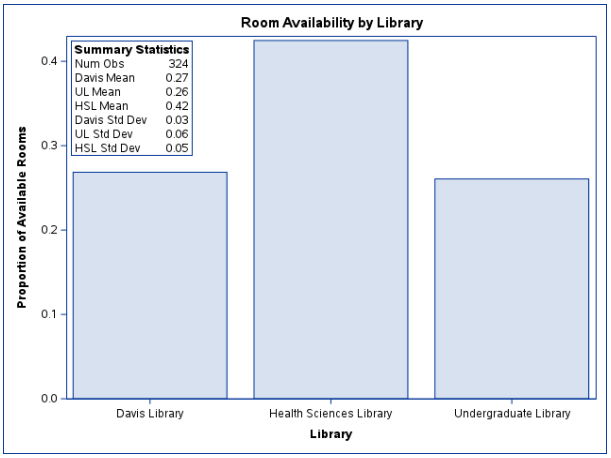
We calculated these estimates in SAS using the surveymeans procedure, including an N= statement, weight statement, and strata statement to account for our stratification design details. The associated variances were estimated using Taylor-Series Linearization on SAS. The code for this is provided in the appendix and was done through the surveymeans procedure. To account for the fpc, we included N=4620 in the surveymeans statement. To account for stratification, we added strata and domain statements to clarify what our strata are and to obtain the within-library variance estimates. We also added a weights statement to account for final weights after adjusting for nonresponse bias. The sample size is calculated based on controlling the margin of error at 0.05 assuming $z_{1-\frac{\alpha}{2}}$=1.96. From the output of the SAS program, the margin of error of the mean estimate, the halfwidth of the confidence interval, can be calculated as $z_{1-\frac{\alpha}{2}}\sqrt{V} = 1.96(0.023) = 0.045 \simeq 0.05$. The margin of error of the overall population mean estimate is successfully controlled within 0.05 using the planned sample size of 357. Below is a table of our point estimates with standard errors and 95% confidence intervals, along with a bar graph of our outcome.

Table 1: Point Estimates and Standard Errors for Proportion of Rooms Available

Stratification Variable		Mean	Std Error	95% CI
Overall		0.31	0.023	(0.27, 0.36)
Library				
	Davis	0.27	0.029	(0.21, 0.33)
	HSL	0.42	0.047	(0.33, 0.52)
	UL	0.26	0.059	(0.14, 0.38)
Time of Day				
	Morning	0.61	0.049	(0.51, 0.71)
	Afternoon	0.12	0.030	(0.06, 0.18)
	Evening	0.21	0.039	(0.13, 0.28)

Figure 1: Room availability by Time of Day and by Library





Final Conclusions

Our overall estimate found that when looking at reservable study rooms 24 hours in advance in the Health Sciences Library, the Undergraduate Library, and Davis Library, 31% of rooms are open to reserve. We found that this availability differs by the library and by time of day. The library with the highest proportion of availability was found to be the Health Sciences Library. One possible explanation for the availability of the HSL is its location. The HSL is centered around the UNC professional schools, and as such, might not receive as much utilization by undergraduate students. The time of day with the highest proportion of availability, nearly triple the availability of the next highest time interval, was the morning (8:00 am - 12:00 pm). This observation possibly occurred because individuals, especially students, are less likely to be on campus early in the day and would rather not wake up early to study.

To assess the non-sampling errors, we looked at our frame problems and measurement errors. When selecting the samples, each of the three libraries is assumed to be open from 8:00 am to 8:00 pm on weekdays. However, during the data collection process, we found out all libraries close at 5:00 pm on Friday, resulting in 19 missing samples in the evening time interval. One potential measurement error would stem from the data collection process. Since one person is responsible for recording the room availability for one day, they might have the library reservation website open throughout the day. People might fail to refresh the page when collecting new data, resulting in some discrepancy in the availability recorded and the true availability. In addition, as we are interested in the room exactly 24 hours beforehand, hypothetically we have to record the availability within a minute. However, our group members usually go back and forth between the website and the data collection sheet, resulting in some of the samples being collected out of the 1-minute window, although it is unlikely that the availability of the room changes within a minute. These measurement errors are minimized by carrying out our data collection training before the data collection process.

When looking at accuracy and precision, we noted that a majority of our missing data was from the HSL in the morning time interval, which could have possibly impacted the results

in which we found the most available library to be the HSL and the time with the most available rooms to be the morning period. Additionally, our estimates achieved close to our expected margin of error, which aids precision. Due to minimal non-sampling errors and quality measures, the accuracy of our estimates was thought to be high. However, we also recognize that rooms marked as reserved may be empty, and non-reserved rooms may be full. This is a potential error in the accuracy of our data.

Appendix

Sample Selection SAS Code:

```
*HSL;
libname sampling "/home/u59228083/BIOS664/Sampling Collection";
run;

data hslsrs;
   infile '/home/u59228083/BIOS664/Sampling Collection/HSL Sample Frame.csv' dlm=','
firstobs=2;
   length Day $10 Library $20 Room $20 Time $10 Time_Interval $20;
   input
      Day $
      Library $
      Room $
      Time $
      Time_Interval $;
run;

/*Step 1: sort HSL data by stratification variable (time interval)*/
proc sort data=hslsrs; by Time_Interval; run;

/*Step 2: Select PROPORTIONATE stratified sample for HSL (stratfied by 'time interval') */
proc surveyselect data=hslsrs method=srs n=34 out = hsl_sample seed = 012620241 stats;
         strata Time_Interval;
run;

*UL;
data ulsrs;
   infile '/home/u59228083/BIOS664/Sampling Collection/UL Sample Frame.csv' dlm=','
firstobs=2;
   length Day $10 Library $20 Room $20 Time $10 Time_Interval $20;
   input
      Day $
      Library $
      Room $
      Time $
      Time_Interval $;
run;
```

```
/*Step 1: sort UL data by stratification variable (time interval)*/
proc sort data=ulsrs; by Time_Interval; run;

/*Step 2: Select PROPORTIONATE stratified sample for UL (stratified by 'time interval') */
proc surveyselect data=ulsrs method=srs n=19 out = ul_sample seed = 012620241 stats;
        strata Time_Interval;
run;

*Davis;
data davissrs;
   infile '/home/u59228083/BIOS664/Sampling Collection/Davis Sample Frame.csv' dlm=','
firstobs=2;
   length Day $10 Library $20 Room $20 Time $10 Time_Interval $20;
   input
      Day $
      Library $
      Room $
      Time $
      Time_Interval $;
run;

/*Step 1: sort Davis data by stratification variable (time of interval)*/
proc sort data=davissrs; by Time_Interval; run;

/*Step 2: Select PROPORTIONATE stratified sample for Davis (stratfied by 'time interval') */
proc surveyselect data=davissrs method=srs n=66 out = davis_sample seed = 012620241 stats;
        strata Time_Interval;
run;
```

Weighting SAS Code:

```
libname mylib '/home/u62053865/BIOS 664';

data room;
 set mylib.groupfdata;
run;

proc freq data=room;
run;

data room_new;
```

```
	set room;
	if Availability=. then missing_flag=1;
run;

proc freq data=room_new;
	table missing_flag;
run;

/* percentage of missing for different libraries */
proc means data=room_new mean;
	var missing_flag;
	class Library;
run;

/* percentage of missing for different time intervals */
proc means data=room_new mean;
	var missing_flag;
	class Time_Interval;
run;

/* percentage of missing for different time */
proc means data=room_new mean;
	var missing_flag;
	class Time;
run;

/* drop the missing data */
/* add base weight */
/* 13.03 davis */
/* 12.94 HSL */
/* 12.63 UL */
data room_nomissing;
	set room_new;
	if missing_flag=1 then response = 0;
	else response = 1;
	if Library = "Davis Library" then SamplingWeight=13.03;
	else if Library = "Undergraduate Library" then SamplingWeight=12.63;
	else SamplingWeight=12.94;
run;
```

```
/* check response rate within each category */
proc means data=room_nomissing mean;
	var availability;
	class Library;
run;

proc means data=room_nomissing mean;
	var availability;
	class Time_Interval;
run;

/* check which one is correlated with missing data rate */
/* *** Method 1: Weighting Class Adjustment ; */
*calculate total weight in each weighting class;
proc freq data=room_nomissing ;
	weight SamplingWeight;
	tables Library*Time_Interval / out=nr_method1_all(drop=percent
rename=(count=total_wt));
run;

*calculate the weight of respondents in each weighting class;
proc freq data=room_nomissing;
 where response=1;
 weight SamplingWeight;
 tables Library*Time_Interval / out=nr_method1_resp(drop=percent rename=(count=resp_wt));
run;

*calculate psi;
data nr_method1_adj;
 merge nr_method1_all nr_method1_resp;
 by Library Time_Interval;
 psi=resp_wt/total_wt;
 NR_adjust=1/psi;
run;
proc print data=nr_method1_adj noobs; run;

*Merge onto dataset and calculate final NR weights for respondents;
proc sort data=room_nomissing; by Library Time_Interval; run;

data nr_method1_final;
```

```
 merge room_nomissing
      nr_method1_adj(keep=Library Time_Interval psi);
 by Library Time_Interval;
 if response=0 then delete; *remove non-respondents;
 NRWT=samplingweight*1/psi;
run;

*Check that weight sums before and after the adjustment match;
proc means data=room_nomissing sum; var samplingweight; run;
proc means data=nr_method1_final sum; var NRWT; run;

proc freq data=nr_method1_final;
tables library*time_interval*NRWT;
run;

proc freq data=nr_method1_final;
Weight NRWT;
tables library*time_interval;
run;
```

Analysis SAS Code:

```
proc surveymeans data=nr_method1_final N=4620 mean std clm plots=none;
   stratum Library Time_Interval;
   domain library time_interval;
   var Availability;
   weight NRWT;
   ods output domain = domainestimate;
run;
```