Team E   
36-303  
Project Proposals  
Tian Wu, Tim Higgins, Matt Belenky, Chao Wang, John Sperger  
  
Final Proposal (Items A-M)  
NOTE: As this is observational study without human subjects we do not need informed consent or to fill out an IRB application.  
Proposal 2—Accuracy of Pittsburgh Buses around Carnegie Mellon  
  
A.  
Why is the topic interesting?  
   Working efficiently and saving as much time as possible is key to raising one’s quality of life, especially to students and teachers of Carnegie Mellon, one of the most intense universities in the world. Many students and teachers are late for classes or important appointments due to the tardiness of public transportation. Waiting for a bus for too long not only disturbs people’s plans, but also negatively affects people’s health and mood. After measuring the actual lateness of buses we could potentially also attempt to measure individual perception of bus lateness.  
  
B.    
What questions do you propose to study?  
a)    Definition of “Being on time for a bus” ( [-2.5, 2.5])  
b)   How many buses are on time.  
c)    How much clustering are buses per day? (Three 61’s in a row)  
d)   Factors causing the lateness

i)    Weather

ii)   Special Events ( such as Super Bowl etc.)

iii) Rush Hour (both morning and evening)

iv) Traffic technical problems (broke down for a bus)

v)   Bus route

e)      Suggestions and solutions  
  
C.  
What previous research has been done?  
Here is our “Works Consulted” while producing this project. Each team member provided a source, and two were suggested by our professor.  
  
                                                  Works Cited  
  
Daskalakis, Nikolaos G. and Anthony Stathopoulos. “Users' Perceptive Evaluation of Bus Arrival Time Deviations in Stochastic Networks.*” Journal of Public Transportation, Vol. 11, No. 4*, 2008.<http://www.nctr.usf.edu/jpt/pdf/JPT11-4Daskalakis.pdf>.  
  
Farid, A. (2010). Trip productivity evaluation of bus service: Medan Kidd Bus Station. http://www.uniten.edu.my/newhome/uploaded/coe\_civil/mutrfc2010/001%20PG%20presentation.pdf. February 25, 2011.  
  
Macaffe, K. (2008).  Bus Punctuality Statistics GB: 2007. Great Britain: Department for Transport.  
http://www.dft.gov.uk/pgr/statistics/datatablespublications/public/buspunctuality/buspunctuality07. February 25, 2011  
  
McCord, Mark and Mishalani, Rabi. http://www.nctr.usf.edu/jpt/pdf/JPT%209-2%20Mishalani.pdf. January 23, 2011.  
  
Weikel, Dan. “MTA gets low marks for bus service in survey.” http://articles.latimes.com/2009/aug/29/local/me-bus-riders29. February 5, 2011.  
   
http://www.nctr.usf.edu/jpt/pdf/JPT12-3Caulfield.pdf. February 5, 2011.  
  
<http://www.dft.gov.uk/pgr/regional/buses/buspunctuality>. UK Dept. of Treasury. February 8, 2011.  
  
Weikel, Dan. “MTA gets low marks for bus service in survey.” http://articles.latimes.com/2009/aug/29/local/me-bus-riders29. February 5, 2011.   
  
D.

The sampling frame will be the list of port authority buses that CMU students use to get to and from school. The sample population will be all the buses that stop at the Forbes and Morewood bus stops. These two stops are adequate for our purposes for several reasons. First, the other Forbes stops (such as Hamburg and Beeler) are very close to the Forbes and Morewood stops. If the bus is late at Hamburg, then it will likely be just as late 100 feet down the road at Forbes and Morewood.

Second, the other stops close to campus (such as Fifth and Morewood, and the Wilkins stops) are used by a substantially smaller percentage of the student population and mostly for the purposes of shuttling to non residential locations. The areas in Shadyside in which students live are well serviced by the faster and more reliable loop buses.

Buses moving up and down fifth will also be independent from buses moving up and down Forbes because many travel through different neighborhoods and thus face different traffic patterns. This would lead to  trouble in data analysis because the independent routes could lead to a bimodal distribution, hurt the accuracy of any inference we would like to make about how late buses usually are, or  increase the number of man hours beyond what is feasible for our group.

E.

The target population is all of the buses which stop on campus that CMU students use to get to school. The target population is differs only slightly from the sampling frame for this study because port authority bus routes are set and all of the buses that stop outside of Hamburg Hall or across the street from it stop at the Forbes/Morewood bus stop. Still, there will be some coverage error because some students take buses from the 5th/Morewood bus stops which have different routes and don’t stop at Forbes/Morewood.

A possible source of sampling error for this study comes from choosing to randomly select hours to watch buses (e.g. 8-9AM) rather than randomly selecting individual buses (e.g. the 10:25PM 28X). While the hours will be randomly selected, this still means that the sampling method will be a combination of convenience (hour long blocks) and random selection. However, using individual buses as the sampling unit would be completely unfeasible given the manpower we have.

A potentially serious source of non-sampling bias is that the study will be conducted over a relatively brief period (2 weeks) in the winter. Seasonal differences in route deviations are likely due to issues such as weather that vary by season. In addition, special events in Pittsburgh could cause excessive delays that are not typical (e.g. post super bowl riots).

Bias due to seasonal differences – by measuring the weather each day during our sample we hope to control for the differences in delays due to weather. Using data about the frequency of inclement weather in Pittsburgh during other seasons the model we develop could theoretically account for seasonal differences. Alternatively, a follow up study could be run in subsequent months.

Error due to special events – if the events cause serious outliers we can just remove them from the data set as a form of post survey data adjustment.

F.

Our mode of data collection will be observational, we intend to observe the buses and note which buses arrives and when. This can then be used to determine if the buses are arriving on time by comparing it to a time table. For this proposal, an observational method is effectively the only usable one, but it is also the best since you cannot survey buses, etc.

In addition, to determine which buses to observe we will use the following method. We will assign every hourly increment in a period spanning approximately two weeks a numerical value. Using a random number generator will select a number of those hours. These will be the ones during which we take our measurements.

G.

   Our primary variable will be deviation from the listed arrival time in minutes. Secondary variables will be weather, time of day, rush hour (yes or no) and route. We will also observe other conditions and may include them as variables.

H.

Is our updated material above.

I.  
 Our project is an observational study without human subjects so no IRB review is required.  
  
J.  
 We do not have respondents.  
  
K.  
We decide to use a stratified random sampling over a 2-week period.  
We will stratify the observation time into 4 periods each day:  
7am-10am, 10am-1pm, 1pm-4pm, 4pm-7pm  
We will sample in 1hr period from these strata.  
  
The Sampling frame for our survey is the list of all Port Authority buses that stop at the Forbes and Morewood bus stops. Forbes and Morewood Out stop: 59U, 61A, 61B, 61C, 61D, 61F, 67, 67C, 67E, 69, 67J, 69A. Forbes and Morewood In stop: 13U, 28X, 59U, 61A, 61B, 61C, 61F, 67, 67E, 69, 67J, 69A  
  
Reason:  
A SRS of hour or half hour periods would increase the likelihood of the time of day negatively impacting the validity of our results, for this reason stratification allows us to compensate for certain problems (like rush hour) and ensure the robustness of our results.

Nonresponse concerns:

Our study involves observations of discrete units (the buses), the buses have no ability to not respond. However, because we are observing in chunks, it is possible that a bus may be so late that it shows up outside of our observational period. If this occurs then researchers are expected to stop recording other buses and wait up to an additional 15 minutes to record the arrival time of the missing bus. Buses that do not arrive in this additional interview will be labeled as too late/did not arrive.

L.  
Observational protocol:  
Weather condition (sunny, windy, rainy, snowy,cloudy…)  
Light/dark level (Day, Night, Dawn/Dusk  
Road condition (dry, wet, covered with ice…)  
Temperature (at beginning of measuring period as recorded by weather.com)  
Inbound/Outbound  
Date  
Day of the week & hour of Observation  
bus number/route  
When the bus is supposed to leave the bus stop  
When the bus actually leaves the bus stop  
Level of lateness (value of the difference between the scheduled and actual departure time, plus for a late bus, minus for an early bus)  
Show/No show status (No show is defined as being so late that arrival time is within five minutes of the next scheduled bus of that type. This is from the article that we refer to.)  
Highly unusual conditions (these will be written down and are meant to include notable and unusual circumstances like a broken water main, major traffic accident, etc. This may be used to throw data out as an outlier).  
Special events and other planned conditions.  
Rush hour (yes/no)  
Clustering (number of other buses of the same number that arrive within 2 minutes of each other).

M.  
We assume that there will be around 12 buses per hour at the bus stop.  
We set ME= 0.5 min  
SD=5 min  
(From the selective research, 5min seems to be a good starting point for the standard deviation of the bus lateness. )  
Z95%=1.96  
N=2wks\*7days/wk\*12hrs/day\*12buses/hr = 2016 buses  
We are doing SRS without replacement.  
n0=Z95%2\*SD2/ME2 = 1.962\*(52)/(0.52) = 384.16  
n=N\*n0/(N+n0) = 2016\*384.16/(2016+384.16)= 322.6729  
Number of hrs = 322.6728/12 =   26.8894  
Inflate the hours by 20% to account for clustering effects.  
è 26.8894\*1.2 = 32.26728  
Since we have 4 stratums, we will randomly sample roughly 8 hrs in each stratum.