Team E II5a

Tian Wu, Tim Higgins, Matt Belenky, Chao Wang, John Sperger

this all looks fine. But -- please see my remarks at the end about sample size calculation.

note that the PAT system is going on reduced service sometime in march; please be sure you complete your sampling either completely before, or completely after, the change in service.

-BJ

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.

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.

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 it's arrival time is within five minutes of the next scheduled bus of that type)

why 5? was this in one of the articles you looked at? (especially given your remarks about clustering below)

Team E II5a

Tian Wu, Tim Higgins, Matt Belenky, Chao Wang, John Sperger

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 set ME= 0.05

SD=5 min

(From the selective research, 5min seems to be a good starting point for the standard deviation of the bus lateness.)

$Z_{95\%}$=1.96

N=12(hrs/day)*7(days/wk)*2(wk)*60(min/hr)=168 * 60 = 10080 min

n0=Z^2*SD^2/ME^2=1.96^2* 5^2/ 0.05= 1920.8 min

n>= Nn0/(N+n0) = 10080*1920.8 / (10080+1920.8) = 1613.364 min = 26.88941 hrs

So we need to sample about 27 hours out of 168 hours over the 2-week period.

Here is another way to think about sample size:

Each hour is a cluster. Each bus is an observation within clusters.

Ignore clusters for a minute, calculate sample size for a SRS w/o replacement, and then inflate by 20% or so to get the sample size you need for clustered sampling.

Total number of buses (N) in two weeks:
I'll assume an average of 12 buses per hour (6 in each direction), but I'm sure you can find a more accurate average!

N = 2 weeks * 7 days/wk * 12 hours/day * 12 buses/hr = 2016 buses

Sample size SRS with replacement:
measuring how many minutes early or late each bus is
SD = 5 min (your estimate)
ME = 0.5 min = 30 sec (seems more achievable than 0.05 min = 1/20 min = 3 seconds!)
Z = 2

n0 = z^2*SD^2/ME^2 = 2^2*(5^2)/(0.5^2) = 400

SRS w/o replacement

n = N*n0/(N + n0) = 2016*400/(2016 + 400) = 334

Number of hours: 334/12 = 28 hrs [again using my crude 12 buses/hr guess]

Now inflate that by about 20% to account for clustering effects... 28*(1.20) = 33 or 34 hours.

You would want to then sample these 33 or 34 hours across the strata, probably 8 or 9 hours randomly chosen in each stratum