Data Collection Techniques and Program Design

Outline

1. Summary of Current Practice
2. Data Collection Program Design Process
3. Data Needs
4. Manual Data Collection Techniques
5. Sampling
6. Special Considerations for Surveying

Summary of Transit Data Collection Practice

- Transition from manual to automated data collection systems
- Variation in techniques used: automated, manual, mixed
- Statistical validity of sampling approach is often weak
- Inefficient use of data often limits use of analytic planning methods
- ADCS presents major opportunity for strengthening data to support decision-making

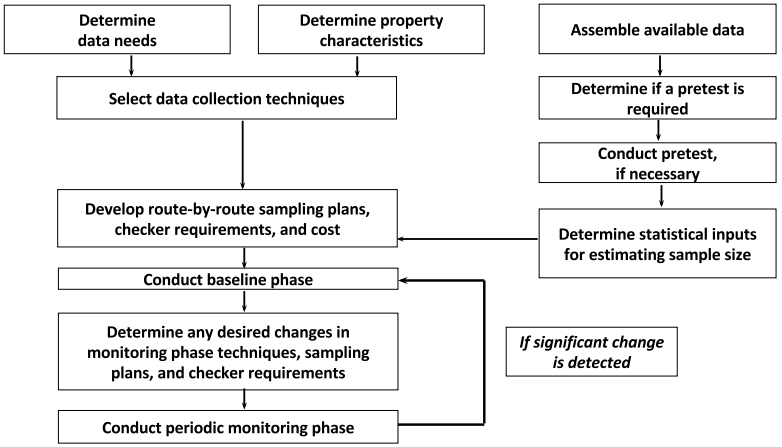
Manual vs. Automatic Data Collection

Manual	Automatic
low capital cost	higher capital cost
high marginal cost	low marginal cost
small sample sizes	large sample sizes
aggregate/disaggregate	aggregate/disaggregate
quantitative/qualitative	quantitative
unreliable	errors and biases can be estimated and corrected
limited spatially and temporally	ubiquitous
not immediately available	available in real-time or quasi-real-time

Key Automated Data Collection Systems

- Automatic Fare Collection Systems (AFC)
  - increasingly based on contactless smart cards with unique ID
  - provides entry (exit) information (spatially and temporally) at the individual passenger level
  - traditionally not available in real-time
- Automatic Vehicle Location Systems (AVL)
  - bus location based on GPS
  - train tracking based on track circuit occupancy
  - real-time availability of data
- Automatic Passenger Counting Systems (APC)
  - bus systems based on sensors in doors with channelized passenger movements
  - passenger boarding (alighting) counts for stops/stations with fare barriers
  - train load weight systems can estimate number of passengers on board
  - traditionally not available in real-time

Data Collection Program Design Process



Data Needs

- Route (Route Segment) Level
  - Load (at peak point -- other key points)
  - Running time
  - Schedule adherence
  - Total boardings (i.e., passenger-trips)
  - Revenue
  - Boardings (or revenue) by fare category
  - Passenger boarding and alighting by stop
  - Transfer rates between routes
  - Passenger characteristics and attitudes
  - Passenger travel patterns
- System Level
  - Unlinked Passenger Trips
  - Passenger-miles
  - Linked Passenger trips

Data Inference

Auxiliary Data Item	**Inferred Data Item**
AFC Boardings APC Counts	Total Boardings
AFC Boardings APC Counts	Passenger Miles
AFC Boardings APC Counts	Peak Point Load

Passenger Counting Techniques

- Manual
  - Checker (with handheld device)
    - ride check (on/off counts and running time)
    - point check (load on board and headway)
- Automated
  - Fare System
    - passenger counts
    - transaction data
  - Automatic Passenger Counters

Manual Data Collection Roles

- If ADCS systems are implemented with analysis and inference tools, the main role for manual data collection is for survey data used to supplement the count-based ADCS information
  - traditional survey methods
  - web-based surveys

Sampling Strategies

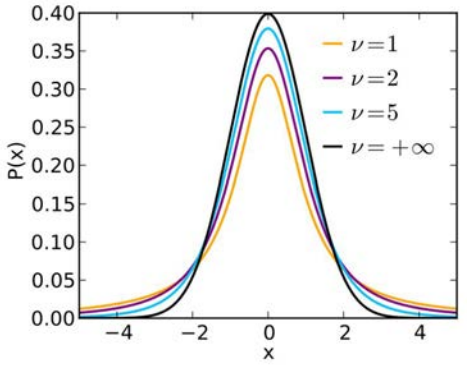
- Simple random sampling
  - Every trip has equal likelihood of being included in sample
- Systematic sampling
  - Sample every 6th day – quasi-random, but smooths data collection load
- Cluster sampling
  - Identify natural clusters in advance, select among them at random
  - With passenger surveys, bus trip = cluster of passengers
    - Example: on-board survey
    - Example: sample round trips, or clusters of 4 trips
- Ratio estimation/Conversion factors
  - Estimate relations between data items from comprehensive (baseline) data set
  - Take advantage of complete or less expensive data sources
    - Example: convert AFC boardings to pass.-mi
    - Example: convert load at checkpoint to load elsewhere
- Stratified sampling
  - Separate sample for each stratum
    - Example: long vs. short routes for average trip length

Tolerance and Confidence Level

- Accuracy of an estimate has two dimensions
  - "Mean boardings per trip is 33.1."
  - Exactly 33.1?
    - relative tolerance
      - "Mean boardings per trip is 33.1, plus or minus 10%"
    - absolute tolerance
      - "Mean boardings per trip is 33.1, plus or minus 3.3"
      - "Mean percentage of students is 23%, plus or minus 5%"
  - Are you sure?
    - tolerance and confidence level
      - "I'm 95% confident that mean boardings per trip is 33.1, plus or minus 10%"
- To simplify matters
  - hold confidence level fixed (90% or 95%)
  - vary tolerance to reflect different levels of accuracy
  - National Transit Database specification for annual boardings, pass-miles: ± 10% relative tolerance at 95% confidence level

t probability distribution

- Arises when estimating the mean of a normally distributed population with unknown mean and variance, from a sample of size *n*



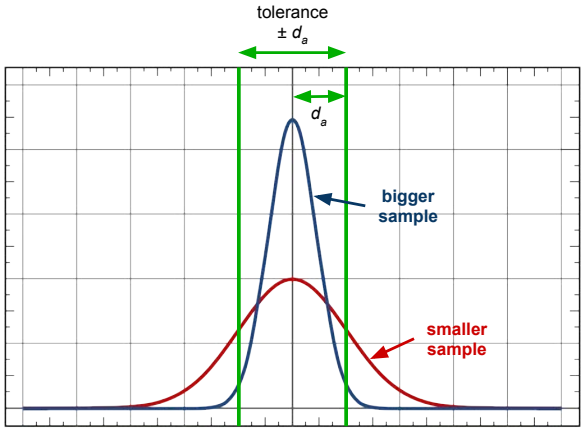
sample mean $\bar{x} = \frac{\sum x_i}{n}$

sample variance $\sigma_x^2 \approx \frac{\sum (x_i - \bar{x})^2}{n-1}$

variance of the distribution of the mean $\sigma_{\bar{x}}^2 \approx \frac{\sigma_x^2}{n}$

degrees of freedom $\nu = n - 1$

Tolerance and Confidence Level



$$d_a = t\sigma_{\bar{x}} = t\frac{\sigma_x}{\sqrt{n}}$$

absolute tolerance

t-statistic: number of standard deviations from the mean

Calculating Required Sample Size

When an absolute tolerance is specified:

$$d_a = t\sigma_{\bar{x}} = t\frac{\sigma_x}{\sqrt{n}}$$

$$n = t^2\frac{\sigma_x^2}{d_a^2}$$

When a relative tolerance is specified:

$$d_r = \frac{d_a}{\bar{x}}$$

$$n = t^2\frac{\sigma_x^2}{d_r^2\bar{x}^2} = \frac{t^2}{d_r^2}\underbrace{\left(\frac{\sigma_x}{\bar{x}}\right)^2}$$

coefficient of variation

Estimating Averages

- Relative tolerances are typically used for averages
  - 5720 boardings ± 5%
    - in this case equivalent to an absolute tolerance of ± 286
- the coefficient of variation is typically easier to guess than the mean and variance separately
  - e.g. with a relative tolerance of ± 5%, a confidence level of 95%, a coefficient of variation of 0.3, and assuming a large sample,

$$n = \frac{t^2}{d_r^2}\left(\frac{\sigma_x}{\bar{x}}\right)^2 = \frac{1.96^2}{0.05^2}(0.3)^2 \approx 140$$

  - sample is not that large, with d.f. = 139, n = 140.732, so let n = 150

Estimating Averages

Simple Random Sample: $n = \frac{3.24v^2}{d^2}$ $d = \frac{1.8v}{\sqrt{n}}$ 90% confidence level and $n > 11$

Where n = sample size (number of trips)
d = tolerance (e.g. d = .05 means ± 5% tolerance)
v = coefficient of variation (i.e. ratio of standard deviation to mean)

Required Sample Size for Estimating Averages

	d = tolerance									
v	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
0.1	13	4	2	1	1	1	1	1	1	1
0.2	52	13	6	4	3	2	2	1	1	1
0.3	117	30	13	8	5	4	3	2	2	2
0.4	208	52	24	13	9	6	5	4	3	3
0.5	324	82	36	21	13	10	7	6	5	4
0.6	467	117	52	30	19	13	10	8	6	5
0.7	636	159	71	40	26	18	13	10	8	7
0.8	830	208	93	52	34	24	17	13	11	9
0.9	1050	263	117	66	42	30	22	17	13	11
1.0	1296	325	144	82	52	37	27	21	17	13
1.25	2025	507	225	127	82	57	42	32	25	21
1.5	2917	730	324	183	117	82	60	46	37	30

Estimating Proportions

- Consider sampling a group of passengers to estimate the proportion of passengers who are students.
- Each observation is a Bernoulli trial, with probability *p* that the passenger is a student.
- We are estimating probability *p* of the Bernoulli distribution
  - variance $\sigma^2 = p(1-p)$ from 0 to 0.25
- Tolerance is typically specified in absolute terms
  - 23% ± 5%

$$n = t^2 \frac{\sigma_x^2}{d_a^2} = \frac{t^2}{d_a^2} p(1-p)$$

Estimating Proportions Using Absolute Equivalent Tolerance

- Collecting data to estimate *p* to calculate sample size can be expensive, so often *p* is initially assumed
  - worst-case scenario: $p = 0.5$, $\sigma^2 = 0.25$
- Appropriate tolerance depends on *p*
  - smaller tolerances and larger samples are typically needed for extreme *p*
    - 32% ± 5% probably OK
    - 1.2% ± 5% probably not OK
- Conveniently, these two factors work in opposite directions
  - if *p* is extreme, tolerance must be small but variance will also be small
  - if *p* is closer to 0.5, variance will be larger, but tolerance can be larger
- Practical solution: assume $p = 0.5$ and work with *absolute equivalent tolerance*.

Estimating Proportions Using Absolute Equivalent Tolerance

- *Absolute Equivalent Tolerance* d_e is the tolerance you'd get if the proportion were 50%.

Expected proportion *p*	Tolerance corresponding to ±5% AET
50%	5.0%
60% or 40%	4.9%
70% or 30%	4.6%
80% or 20%	4.0%
90% or 10%	3.0%
95% or 5%	2.2%

$$d_e = \frac{0.5 d_a}{\sqrt{p(1-p)}}$$

$$n = \frac{t^2 \sigma_x^2}{d_a^2} = \frac{t^2 p(1-p)}{d_a^2} = \frac{0.25 t^2}{d_e^2}$$

Example: 95% confidence, large sample

$$n = \frac{0.25 \cdot 1.96^2}{d_e^2} = \frac{0.96}{d_e^2} \approx \frac{1}{d_e^2}$$

with d_e = 5%,

$$n = \frac{1}{0.05^2} = 400$$

Sample Size for Proportions

- A large sample size is needed to estimate a proportion accurately

Tolerance achieved by AET and sample size

n	600	267	150	96
AET	4%	6%	8%	10%
p = 50%	4%	6%	8%	10%
p = 60%	3.9%	5.9%	7.8%	9.8%
p = 70%	3.7%	5.5%	7.3%	9.2%
p = 80%	3.2%	4.8%	6.4%	8.0%
p = 90%	2.4%	3.6%	4.8%	6.0%
p = 95%	1.7%	2.6%	3.5%	4.4%

Sample Size for Passenger Surveys

- Determine needed sample size for proportion
  - e.g., proportion of passengers who are pleased, who own a car, etc.
- Multiply sample sizes if proportions are desired for various strata
  - e.g., proportion of passengers car-owning passengers who are pleased
- Multiply by "clustering effect"
  - e.g., in on-board survey, 4 responses from same bus may be equivalent to 1 response from a randomly selected rider; clustering effect depends on question
  - if so, expand sample size by 4
- For origin-destination matrix,
  - sample size = 20 * number of cells (rule of thumb)
  - level of detail determined number of cells
- Expand by 1/(response rate)
- Be prepared to revise your expectations when you see how large the needed sample is!

Response Rate

- Along with getting correct answers, your primary concern should be getting a high response rate
- Cost: lower response rate means more surveying to get the needed number of responses
- Non-response bias: non-responders may be different from responders, and you'll never know!

Non-Response Bias

- Non-responders may be different from responders
- Some non-response bias is predictable and insidious
  - standees are less likely to respond, making close-in origins underrepresenting
  - low literacy, teens, and non-native population respond less
  - predictable biases can be modeled and corrected by numerical procedures
- Ways to improve response rate
  - shorten the questionnaire
  - quick oral survey: "What station are you going to?"
  - get information from counts whenever possible (e.g. fare type)
  - distribution method, surveyor training, supervision

Suggested Tolerances

- Peak Load (also boardings)
  - Routes with 1-3 buses ±30%
  - Routes with 4-7 buses ±20%
  - Routes with 8-15 buses ±10%
  - Routes with >15 buses ± 5%
- Vehicle trip time
  - Routes with trip time ≤ 20 mins ±10%
  - Routes with trip time ≥ 20 mins ± 5%
- On-time performance
  - ±10% AET

Default Values for Coefficient of Variation of Key Data Items

Data Item	Time Period	Route Classification	Default Value
Maximum Load	Peak	< 35 pass./trip	0.50
		≥ 35 pass./trip	0.35
	Off-Peak	< 35 pass./trip	0.60
		35-55 pass./trip	0.45
		> 55 pass./trip	0.35
	Evening	All	0.75
	Owl	All	1.00
	Saturday	All	0.60
	Sunday	All	0.75
Running Time	All	short (≤ 20 min.)	0.16
		long (> 20 min.)	0.10

Step-by-Step Manual Data Collection Program Design

1. Determine data needs and acceptable tolerances based on uses of data
2. Select statistical inputs (i.e. coefficient of variation) based on preliminary data analysis and/or default values.
3. Select data collection techniques based on data needs and efficiency of each technique for property.
   a. Baseline: ridechecks + supplementary point checks
   b. Monitoring: pointchecks
   c. Update: ride checks
4. Determine constraining sample sizes for each technique by route and time period by applying formula.
5. Determine detailed checker requirements for each route and time period.
6. Estimate ratios (e.g. average fare, trip length, peak load/total passengers) using baseline data for possible use in monitoring.
7. Revise monitoring plan (techniques and sample sizes) based on data analysis.

Conversion Factor Equations

Compute conversion factor and its coefficient of variation: $R = \frac{\bar{y}}{\bar{x}}$

where R = conversion factor

$\bar{y}$ = average of inferred data item (e.g. boardings) in paired sample

$\bar{x}$ = average of auxiliary data item (e.g. load) in paired sample

$$v_R^2 = \frac{1}{n - 1.7}\left(v_x^2 + v_y^2 - 2 v_x v_y r_{xy}\right)$$

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

$$S_{xy} = \mathrm{Cov}(x, y) = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{n - 1}$$

where v_R = coefficient of variation of conversion factor

v_x = coefficient of variation of auxiliary item (e.g. load)

v_y = coefficient of variation of inferred item (e.g. boardings)

r_{xy} = correlation coefficient between auxiliary and inferred items

n = number of paired observations in sample

Determine Sample Size in Monitoring Phase

$$n_2 = \frac{v_x^2\left(1 + v_R^2\right)}{0.31\, d_m^2 - v_R^2}$$ (90% confidence level assumed)

where n_2 = sample size of auxiliary item in monitoring phase

d_m^2 = desired tolerance of the inferred data item

Desired tolerance of the inferred data item = ±10%

V_x \ V_R^2	.0001	.0005	.0010	.0015	.0020	.00225	.0025	.00275
0.10	4	4	5	7	10	12	17	29
0.20	14	16	20	26	37	48	67	115
0.30	31	35	43	57	82	107	151	258
0.40	54	62	77	101	146	189	268	459
0.50	84	97	120	157	228	295	418	717
0.60	121	139	172	226	328	425	602	1032
0.70	16	189	234	307	447	578	819	1404
0.80	214	247	306	401	583	755	1070	1834

1.258J / 11.541J Public Transportation Systems
Spring 2017