B. Collection of Information Employing Statistical Methods

1. Describe the potential respondent universe and sampling method used.

As defined in Section 19–8 of the Department's Economic Regulations (14 CFR § 241), the Origin and Destination (O&D) data samples revenue passenger trips moving in whole or in part on domestic and/or international scheduled air carrier services. The carriers that will report data, called Reporting Carriers, shall include all certificated air carriers conducting scheduled passenger services that sell passenger tickets (except helicopter carriers). Small certificated air carriers and commuter air carriers that do not participate in franchise code-share agreement would only report tickets sold on scheduled service flights conducted by the carrier. Carriers that participate in code-share agreements, or contracts with carriers to provide contract lift will not be required to report, unless they issue tickets for their own branded services. The reporting carriers are to collect and report data in accordance with the Instructions and supplemental O&D Directives that may be issued periodically.

As new carriers begin service that they sell, they will be required to file O&D data. These carriers will not be added to the reporting carrier list automatically, but will be added as soon as administratively possible. The other reporting carriers will then be notified with the updates to the reporting carrier list at least one month in advance.
The O&D data are collected from a passenger's ticket. The data includes the passenger's routing, i.e. origin to destination and all connecting points. The data sampling reported is based on the final, right-most digit of the standard ticket document number. Reporting carriers using a random numbering system with the right-most digit equal to "0" (zero), "2" (two), "7" (seven) or "9" nine would report the complete itinerary into the data submission. Group tickets with more than 10 passengers are to be reported regardless of ticket number. Any reporting carrier that does not assign a ticket number in which a final, right-most digit is not randomly assigned must develop an alternative method of creating a valid 40% sampling of data. The alternative method must be approved by the Office of Airline Information and Statistics Director 60 days before reporting data.

2. Description of procedures for the collecting information, including statistical methodology for stratification and sample selection, estimation procedures, degree of accuracy needed.

Reporting Carriers will examine each flight coupon received. Each passenger ticket contains at least one flight coupon, known as a flight coupon stage. Upon presentation of the first flight coupon to be flown from the itinerary at the boarding gate, a Reporting Event occurs. This event will notify the carrier that issued the ticket that a lift has occurred for transportation. Based on the reporting event and the sampling process, the issuing carrier becomes the reporting carrier and this carrier will report the ticket to the DOT.

The information reported will include each flight coupon. A flight coupon lists the year and month of travel, air carrier that will be operating the route, air carrier that issued (sold) the ticket to the passenger, marketing carrier(s) of other segments in the ticket (if applicable), total amount of the ticket including taxes and fees, external entity fees and taxes collected for the ticket, amount of time spent between the arrival at one airport and the departure from that same airport (dwell time) (if applicable) and the point of stopover/connecting point of the passenger on a "direct" or "through" flight where the flight number does not change, known as a via point (if applicable). On subsequent coupons, the travel flight year and month will be listed, if applicable, and this is to be reported. A unique record identification number will be included with the submitted ticket information. This number will allow for easier identification for correcting tickets with errors. Please see *Appendix C* for descriptions of each data element.

All tickets issued by a reporting carrier, regardless of market size, carrier size or size of aircraft the carrier operates are to be reported.

The following paragraphs are the statistical methodologies for the stratification and sample selection and the estimation procedures for collecting the data:

Estimation and variance estimation procedures are as follows:

Consider a quarterly O&D dataset. Let n_g denote the observed number of passengers in the *g*-th OD pair using the O&D data and N_g the true number of passengers. Here $g = 1, 2, ...G,$ where G is the number of possible OD pairs. Note that with the current sample design of the O&D data set, $f_g = \frac{n_g}{N_g}$ could be zero, especially for smaller markets, in some quarters. Let y_{gi} denote the price paid by the *i*-th passenger in the sample and Y_{gi} the price paid by the *i*-th passenger in the quarter's population of passengers for the *g*-th OD pair. Then, an estimate of the average ticket price:

$$\bar{Y}_g = \frac{1}{N_g} \sum^{N_g} Y_{gi}$$

is given by:

$$\bar{y}_g = \frac{1}{n_g} \sum^{n_g} y_{gi}$$

It can be seen that:

$$E(\bar{y}_g \mid n_g) = \bar{Y}_g$$

So, this estimator is conditionally unbiased. Also,

$$E(\bar{y}_g) = EE(\bar{y}_g \mid n_g) = \bar{Y}_g$$

and so the unconditional bias of this estimator is zero. The conditional variance of this estimator is:

$$V(\bar{y}_g \mid n_g) = \left(\frac{1}{n_g} - \frac{1}{N_g} \right) S_g^2$$

Where:

$$S_g^2 = \frac{1}{N_g - 1} \sum_{i=1}^{N_g} (Y_{gi} - \bar{Y}_g)^2$$

The unconditional variance is obtained as follows:

$$V(\bar{y}_g) = EV(\bar{y}_g \mid n_g) + VE(\bar{y}_g \mid n_g)$$

The second term equals 0 and:

$$V(\bar{y}_g) = EV(\bar{y}_g \mid n_g) = E\left(\frac{1}{n_g} \right) S_g^2 - \frac{1}{N_g} S_g^2$$

Since N_g is unknown, an unbiased estimator of the variance is:

$$\frac{1}{n_g}\left(1 - \frac{n}{N}\right) s_g^2 = (0.6)\frac{s_g^2}{n_g}$$

Where:

$$s_g^2 = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (y_{gi} - \bar{y}_g)^2$$

3. **Describe the methods to maximize response rates, and describe how the Department deals with non-responses**.

The Department contacts delinquency carriers when a report is late filed. The contact may be a telephone call or an email transmission. If no response is forthcoming, then a warning letter is sent to the carrier requesting the data be submitted within the next five business days. If the reports are not received within the 5-day period, the matter is referred to the Assistant General Counsel for Aviation Enforcement and Proceedings. DOT has the authority to fine carriers for each day that the data report is late without just cause. However, fines and penalties are generally used as a last resort. Overall, the airline industry has an outstanding record for complying with O&D data reporting obligations. Occasionally, there may be a delayed response due to a carrier strike or bankruptcy. When a delayed response does occur, the Department will place a notice on the reporting status internet page to alert users that a carrier's data are not included because of the delay.

4. **Describe any tests of procedures or methods undertaken**.

Carrier reports are electronically reviewed for conformance to instructions, traffic volumes and for various other relationships. Major problems discovered in this review or in later stages of processing are taken up with the carrier and resolved.

Reported data are subjected to many computerized edits. The first set of computerized edits are preliminary. These checks occur at the initial submission. Checks that occur at this stage are, (1) file format is verified (file is in .csv format) and (2) file name is correct (two letter airline code, reporting year and quarter, i.e. AA202003). Checks are also done to ensure that a transmittal letter has been submitted and the year and quarter of the letter is correct. If there are any issues at this stage, the Data Analyst will send the file back to the reporting carrier for correcting and re-submission.

At the next stage, the data analyst 'processes' the raw data input. Computerized edit checks occur at this stage to detect input format problems. Some of the checks performed by the system are, reporting year/quarter are not null, itinerary fares are not negative or null, reporting/operating/issuing carriers are not null, and the number of passengers is not null. Airport and carrier codes on each flight coupon stage of the itinerary are tested for validity against the Official Airline Guide electronic files and are IATA/DOT issued codes, the itinerary is not incomplete, surface-transportation portions at the beginning or end of ticket itineraries are removed, and the operating carrier on each flight coupon-stage is tested to

determine if it serves the airport of the flight-coupon origin and destination. If the carrier does not service the airport, the record is placed on the "Deletions Report" for the Data Analyst to review. If the Data Analyst is not able to resolve the issue after research (i.e. contacting the carrier), the record not included in the final product. The passenger volume on dropped records is a fraction of one percent of the total number of sample passengers reported by each carrier.

To illustrate:

Edit for Alaska Airlines for 2nd Quarter 2020

Tickets with invalid fare codes	0
Tickets with invalid point codes	100
Tickets with surface at start or end	0
Tickets with invalid carrier on flight coupon	14
Tickets with invalid output format	0
Tickets requiring modification	69
Percent of tickets requiring modification	0.08 (where 1.00 = 1 percent)
Number of tickets in	209,315
Number of passengers in	515,247
Number of tickets deleted	0
Number of passengers deleted	0
Number of tickets out	208,156
Number of passengers out	514,447
Percent of tickets passing edit	99.77
Percent of passengers passing edit	99.77
Average flight coupons per ticket	1.83

As can be seen above, less than one tenth of one percent of the flight coupons submitted were incorrect. City-pair passengers from the O&D data are normalized for comparison to the carrier's T-100 traffic reports as a further check. Significant discrepancies are discussed with the relevant carrier for correction.

Each carrier is responsible for developing edit procedures and internal controls over its data entry and processing procedures so that valid and reliable data are captured in the O&D data inputs. Since the carriers have many different statistical systems, it is not practicable for the Department to prescribe specific controls in this area. Each carrier is responsible for developing the appropriate internal control procedures to ensure integrity and accuracy of the data.

5. Provide the name and telephone number of individuals consulted on the statistical aspects of the design and the name of the agency unit, contractor grantee, or other persons who will actually collect and/or analyze the information for the agency.

Mr. James Bouse is the contact person for the O&D data. He can be reached at (202) 366-4876.

MISSION STATEMENT

The Department of Transportation (DOT) depends on the financial data reported on Form 41 to fulfill its strategic plan to monitor and study the movement of aircraft and passengers. Further, the DOT has adopted an agency-wide, coordinated effort together with the Office of the Secretary, the Federal Aviation Administration, the Bureau of Transportation Statistics (BTS), and Office of the Inspector General to advance consumer satisfaction.

BTS continually strives to improve the quality, reliability and accessibility of transportation-related information. BTS is also mindful to mitigate the paperwork burden imposed on the air transportation industry and the public: in part by advancing the precepts of the Clinger-Cohen Act and the Paperwork Reduction Act by re-engineering its data processing system.

APPENDIX

APPENDIX A

The Origin and Destination (O&D) data collection began in 1968 and was manually filed. As time moved on, electronic submissions replaced the manual submissions. The last rule change occurred on January 1, 1998.

APPENDIX B

All U.S. Certificated Air Carriers, regardless of carrier size or size of aircraft, will be required to report to the O&D data collection. Carriers that will report data will be known as Reporting Carriers. All reporting carriers will appear on the Reporting Carriers List, which will be maintained by the Office of Airline Information.

Reporting Carriers will report revenue passenger tickets that they issue that meet the criteria to be included in the sample and have flown lift usage. Tickets issued by a reporting carrier will be known as *Category One* tickets. The reporting carrier will report the itinerary of their issued tickets at the moment they process the first lifted flight coupon, regardless of which carrier lifted the coupon or where the coupon falls in the sequence of all the coupons in the

ticketed itinerary. The first lift is known as the reporting event. The tickets that will be reported are called reportable tickets. Reporting carriers will always examine the coupons that were issued by carriers that do not appear on the Reporting Carriers List to determine whether they have a duty to report that ticket.

Tickets issued by a carrier that is not a reporting carrier will be known as Category Two tickets. When examining coupons of tickets issued by carriers not on the Reporting Carriers List, reporting carriers will employ long ago established "first reporting carrier" rules as the criteria for deciding whether the tickets should be reported. In the current 19.7 collection, there is a section called, "B. *Selection of Reportable Flight Coupons.* The flight coupons identified above are to be examined to isolate the reportable flight coupons, i.e. coupons from which data are to be recorded. Flight coupon data are reported only by the first reporting carrier. Such carriers shall report the required data for the entire ticketed itinerary. If another reporting carrier has preceded an examining carrier on any stage in the trip itinerary, including any stage in a conjunction itinerary and any stage in a re-issued ticket (either before or after re-issue) that coupon is not reportable.

If tickets eligible to be reported were re-issued, the original ticket and the re-issued ticket(s) are to be reported, only if the original ticket meets the reportable ticket qualification. A re-issued ticket is considered an exchanged ticket. An exchanged ticket indicator is to be applied to the reported exchanged ticket. The value of the ticket applied to the exchanged ticket is to be reported, as well as the new itinerary of the passenger.

No adjustment is made in the O&D data for alterations or changes in the trip itinerary subsequent to the stage covered by the reportable coupon.

APPENDIX C

Data Elements to be collected:

Destination: An airport in the ticket sequence of travel where the passenger deplanes from a flight stage. IATA/DOT airport codes are to be submitted.

Dwell Time: Period of time passenger spends on the ground after the previous flight segment has been completed and before departure of the next flight segment. In the 19.7, the origins and destinations are provided for each segment. It is not known, however, the amount of time that the passenger spends at each mid-point en-route to their destination. In 19.8, the reporting of dwell time will remove the expense and error associated with deciding when a passenger has reached a destination and when the passenger is simply waiting for a connecting flight to the intended destination. Reporting of this element will also enable better alignment with the T-100 monthly data. As the reporting carrier knows the flight dates and flight times of a ticket's itinerary, the DOT proposes that the reporting carriers report in one hour increments the number of hours elapsed between a passenger's arrival and the

passenger's departure from an Airport. If the number of hours is greater than 24, use the value of '99'.

Exchanged Ticket Indicator: In 19.8, an indicator will appear on those selected tickets in which the remaining value of a previously issued has been applied to a newly issued ticket. In 19.7 this data element does not exist. This field will enable analysts to determine when a reported fare may not comport with the original itinerary.

Issuing Carrier: The IATA/DOT identifier of the airline carrier that issued the ticket.

Marketing Carrier: Under a code-share agreement, the IATA/DOT air carrier code that markets the seat on the aircraft, whether it operates the flight segment or not.

Operating Carrier: Under a code-share agreement, the air carrier whose aircraft and flight crew are used to perform a flight segment. IATA/DOT carrier codes are to be submitted.

Origin: The first point in the itinerary and the point where the passenger first boards a carrier at the beginning of the itinerary. IATA/DOT airport codes are to be submitted.

Reporting Carrier: The carrier in a given itinerary which has lifted the reportable flight coupon in that itinerary and which carrier is required to record the O&D data for that itinerary for the report to the Department. IATA/DOT carrier codes are to be submitted.

Reporting Month: Month in which a coupon in a ticket is used for air transportation for the first time.

Reporting Year: Year in which a coupon in a ticket is used for air transportation for the first time.

Record Identification Number: A unique Record identifier submitted by the Reporting Carrier.

Tax Amount: The aggregate of fees and taxes imposed by external entities (e.g. airport operating authorities and government jurisdictions). Examples of taxes include the Passenger Facility Charge, International Departure and Arrival Tax, and Flight Segment Tax.

Total Amount: In 19.7, the data element of "Total Dollar Value of Ticket" represents the fare paid by the passenger plus taxes. Being one value, there is uncertainty as to the value of the fare itself and the value of the taxes. In 19.8, this data element will be re-named to the industry standard term of "Total Amount".

> The reporting carrier would report the total amount collected for the purchase of the ticket that allows the passenger to board the aircraft. The total amount would include all fees and taxes imposed, including carrier-imposed surcharges that are identified as fuel

charges and other descriptions, as well as the amount of non-airline imposed taxes and fees for the ticket.

The reported total value of the ticket will include any fee(s) associated with the purchase of a passenger ticket in order for the passenger to board the aircraft. These fees include, but not limited to, the purchase of a ticket either through the carrier's website or a third-party vendor's website (include only if the third party vendor's fee is included in the ticket price), the purchase of a ticket either through the carrier's phone reservation system or at the airport ticket counter, the purchase of a ticket through a travel agency (include only if the travel agency's fee is included in the ticket price), the purchase of a frequent flier ticket when using miles for travel, check-in fees, the printing of a boarding pass from the carrier's website, the selection of a seat on the aircraft, and the fee charged for the use of a credit card to purchase a ticket (if the credit card fee is included in the fare). The total value of the ticket is as of the time of the ticket purchase.

Differentiating the amount of tax collected from the amount of total fare collected removes uncertainty in determining the actual passenger revenue retained by the airlines. This amount would not include ancillary charges, such as baggage fees or ticket change fees.

Travel Flight Month: The scheduled month of each flight coupon stage in the itinerary.

Travel Flight Year: The scheduled flight year of each flight coupon stage in the itinerary.

Via points: Points in which an aircraft lands and departs with the same flight number at a planned point of stopover. These flights are considered "through flights". These "via points" are in the carrier's reservation systems, but have not been included in the itineraries reported to DOT under 19.7. The focus on 19.7 has been on the "Origin" and "Destination" and not the intermediate points (unless a transfer occurred to a different flight number in the itinerary). Reporting all cities in the itinerary will better align the O&D data with the T-100 data, removing the effort and cost of market validation analysis. This will allow the T-100 to facilitate validation of O&D data submissions.

APPENDIX D

Glossary of Terms:

Bilateral agreement: Prior to an air carrier operating flights to another country, both countries must negotiate a treaty level agreement. This agreement is called a bilateral agreement.

Commuter Air Carrier: An air carrier that operates small aircraft which consists of 60 seats or fewer and/or performs scheduled passenger service of at least five round trip flights per week between two or more scheduled points.

Connecting point: An intermediate point in a sequence of travel at which the passenger deplanes from one flight and boards another flight, either on the same carrier or from the flight of one carrier to a flight of another carrier, for the continuation of the journey.

Coupon/Coupon Stages: See *Flight Coupon Stage*.

Data Analyst: The individual(s) in the Office of Airline Information responsible for processing the reported the O&D data.

Examining Carrier: A carrier that examines a ticket to determine if it is to be reported.

First Reporting Carrier Rule: Rule applied during the reporting event evaluation. The first reporting carrier in the sequence of travel for a Category Two ticket is designated as the carrier responsible for reporting the ticket.

Flight Coupon: See *Flight Coupon Stage*.

Flight Coupon Stage: A defined origin and destination for a single stage of flight provided by a single operating carrier. Tickets are composed of one or more flight coupon stages (also known as coupons and coupon stages).

Group Tickets: A single ticket issued to two or more passengers.

Intraline: An agreement that is not in place between air carriers to coordinate passengers with itineraries that encompass multiple airlines. Passengers traveling on intraline carriers have to check-in with the carrier of their next segment of travel when they land at their stopover point.

Interline: An agreement between air carriers to coordinate passengers with an itinerary that encompasses multiple airlines to not have the passengers check-in again or have to deal with their luggage again at the point of stopover.

Office of Airline Information: The department in the US Department of Transportation, Office of the Secretary of Research and Technology (OST-R), Bureau of Transportation Statistics division responsible for collecting, processing and disseminating the O&D data.

Points of stopover: See *Via points.*

Re-issued Ticket: A ticket issued in exchange for all or part of the unused portion of a previously issued ticket. A re-issued ticket is also considered an exchanged ticket.

Reportable Ticket: The combination of flown flight usage, sampling process criteria, and the Category One and Category Two ticket evaluation determines if a ticket is reportable.

Reporting Event: The occurrence of a Reporting Carrier recognizing that a ticket has been flown and evaluating the ticket to determine if it should be reported to the O&D data.

Revenue passenger: See Title 14 Code of Federal Regulations Section § 241 Section 03 – Definitions for Purposes of this System of Accounts and Reports – Passenger, revenue.

Routing: The sequence of travel for each flight coupon stage including all intermediate points of routing, stopover, or connection (interline or intraline) in the movement of the passenger from the first airport in the sequence of travel to the last airport in the sequence of travel for the ticket.

Scheduled Service: Transport service operated on a Certificated Air Carrier or Commuter Air Carrier's routes pursuant to published flight schedules, including extra sections of scheduled flights.

Small Certificated Carrier: An air carrier holding a certificate issued under Section 41102 of the Statute that provides scheduled passenger air service within and between only the 50 States of the United States, District of Columbia, the Commonwealth of Puerto Rico, and the U.S. Virgin Islands with small aircraft as defined in this section.

Ticket: A legal contract between an Issuing Carrier and a revenue passenger.

Yield: Passenger revenue per revenue passenger-mile.

APPENDIX E

October 5, 2015
The Honorable Susan Kurland
Assistant Secretary for Aviation and International Affairs
U.S. Department of Transportation
1200 New Jersey Avenue SE
Washington, DC 20590

Dear Ms. Kurland:

As you know, A4A and our members have been working with your staff, in particular Todd Homan and the Office of Aviation Analysis team, to develop mutually acceptable changes to DOT's Passenger Origin and Destination (O&D) data set incorporated in 14 CFR Part 241. The goal of this "modernization" exercise has been to improve the integrity and reliability of the data DOT collects, reduce the administrative burden on carriers, and avoid unnecessary

reporting of sensitive sales and related information. Together, we have worked diligently on this project over the past few years.

I am pleased to inform you that we recently completed that effort, as outlined in the attached document entitled "A4A-DOT Aviation Data Modernization Solution Understanding: July 2015." Although the outline substantially expands airline data reporting, A4A members support all elements in this document. The Office of Aviation Analysis likewise has advised that this solution satisfactorily addresses its goals to improve the information generated by the O&D data set.

The attached solution understanding reflects a tremendous amount of effort and time taken by A4A and its member carriers, in consultation and collaboration with DOT experts, to achieve consensus. It has been carefully crafted to address all stakeholder needs and concerns. For this reason, any deviation from the document would be problematic and undermine our support. Accordingly, we look forward to the Department initiating a rulemaking that reflects the changes contemplated in the document.

Finally, given that the changes are fairly extensive and include, on balance, the reporting of a greater volume and scope of information at more granular level, A4A members anticipate needing up to two years to make and test the one-time changes to their systems and procedures. One of the biggest programming changes is the required marriage of schedules with revenue accounting data. Another is the determination of international taxes.

We would happy to meet with you or your staff as needed in advance of an NPRM. Please feel free to contact me with any questions.

Sincerely,

Sharon L. Pinkerton

APPENDIX F

Computing the Sample Rates Necessary for BTS O&D Average Fare Estimates for Small Airports

Michael D. Wittman[a,b]

[a]Inter*VISTAS* Consulting, 125 High Street, Suite 2104, Boston, MA 02110 [b]Massachusetts Institute of Technology, International Center for Air Transportation 77 Massachusetts Avenue, Building 35-217, Cambridge, MA 02139

Introduction

The Passenger Origin and Destination (O&D) data, produced by the Bureau of Transportation Statistics (BTS), is a widely-used sample of domestic airline ticket data. The O&D is highly cited in government, academic, industry, and media reviews of domestic airfares. Currently, the O&D collects a 10% sample of tickets sold by U.S. carriers that fit certain requirements. However, this sample rate is likely too low to accurately reflect changes in average fares and passenger volumes at very small airports.

The Department of Transportation has expressed an interest in changing the sample rate[1] of domestic airline tickets to improve data quality in the O&D data. This document uses statistical techniques to calculate O&D sample rates that are sufficient to reflect changes in average fares at small airports to a certain degree of accuracy. The sample rate necessary will depend on DOT's preferences for margins of error and confidence levels; this document assumes standard statistical confidence intervals, but these parameters can be changed with DOT input.

Problem Statement

Consider a small airport that enplanes 500 one-way passengers per quarter.[2] We will compute the sample rate necessary to accurately predict average fares at this airport with a certain margin of error—for instance, +/-10%. To solve for this sample rate, we will need to split the problem into two subproblems:

1. First, we will compute the sample size necessary to predict average fares at the airport by +/-10%. That is, for an airport with 500 enplaned passengers per quarter, how many tickets would need to be sample to predict average fares at that airport with a certain degree of confidence?

2. After the sample size is determined, we will then compute the sample rate needed to select that many tickets out of the entire population of tickets. That is, we will compute the percent of tickets necessary to ensure that the sample will contain at least as many tickets as are necessary to accurately compute average fare for that market.

These two problems will be solved sequentially. We will then examine how the required sample rate will change based for airports of various sizes and how these rates will change as a function of the level of confidence required in both the average fare calculation and the sample generation.

Computing Sample Sizes

First, we wish to calculate the sample size necessary to estimate the average fare of the population of tickets from some small airport with a certain degree of accuracy. We can use a com-mon and well-known sample size formula (e.g., from Section 7.2.2.2 of NIST's Engineering Statistics Handbook, at http://www.itl.nist.gov/div898/handbook/prc/section2/prc222.htm):

$$n = \left(t_{df,\alpha/2} \cdot \frac{s}{E} \right)^2$$

where n is the required sample size, $t_{df,\alpha/2}$ is the critical value from Student's t distribution for a distribution with degrees of freedom df and confidence level α, s is the standard deviation of the sample, and E is the desired margin of error.

[1]In this document, *sample rate* will refer to the percent of total tickets sampled by the DOT. In the current O&D data set, the sample rate is 10%. [2]This represents approximately 500*2/91.25 ≈ 11 passengers per day both ways.

When the required sample size (and hence the degrees of freedom) is unknown, we can use critical values from the z-distribution instead of the Student's t distribution to approximate n, so long as n is sufficiently large. Then, the formula becomes:

$$n = \left(z_{\alpha/2} \cdot \frac{s}{E} \right)^2$$

Now, let's use this formula to compute sample sizes necessary for a variety of airports.

A Practical Example of DVL

Consider a very small Essential Air Service airport, such as Devils Lake, ND (DVL). For DVL, T-100 data estimates 175 total onboards for 1Q2014. Assuming 1.0 passengers per ticket[3], this would lead to a total population of 175 tickets. Note that BTS estimates 313,564 passengers flew out of North Dakota airports in 1Q2014, so DVL tickets represent just 0.05%, or about 1 in 2,000, of the total ticket sample for North Dakota alone.

For DVL, only three ticket observations were recorded in the 1Q2014 O&D.4 The average fares for these three observations had a mean of $979 and a standard deviation of $431.50, reflecting the significant uncertainty regarding this sample.

Suppose for a moment the average fare from this airport truly is $979, and that from our sample we wish to compute the average fare within 10% accuracy with a 95% confidence level. That is, 95% of the samples we select would have an average fare of within 10% of the true average fare of the airport, or within $97.90. Then, using our formula, we would have:

$$n = \left(t_{df,\alpha/2} \frac{s}{E} \right)^2 \approx \left(z_{\alpha/2} \frac{s}{E} \right)^2 = \left(1.96 \cdot \frac{\$431.50}{\$97.9} \right)^2 = 74.62 \approx 75 \text{ tickets} \qquad (3)$$

The refore, we would need a sample size of 75 tickets to predict the average fare from DVL +/- 10%, with a 95% confidence level. With an estimated 175 DVL tickets in the total

population, a sample size of 75 means that 42.9% of DVL tickets would need to be included in the final sample, compared to just three in the actual 1Q14 sample.

Sample Sizes for Generic Airports

However, since the average fare and standard deviation will vary across airports, using a single airport to decide sample size is unlikely to produce robust results. A generic example may thus be more useful to DOT when setting policy for sample sizes across many different airports.

Let's consider a generic example. Suppose we have an airport with a true average fare of $500, that we wish to predict that airport's average fare within 10% (or within $50), and that the observed standard deviation of the average fare is $300.[5] Then, to have the sample average fare be within $50 of the true average fare 95% of the time, we would need a sample size of:

$$n = \left(t_{df,\alpha/2}\frac{s}{E}\right)^2 \approx \left(z_{\alpha/2}\frac{s}{E}\right)^2 = \left(1.96 \cdot \frac{300}{50}\right)^2 = 138.29 \approx 139 \text{ tickets}$$

[3]Based on DOT estimates.

[4]This highlights how the current 10% sample rate is likely too small—we would have expected 17.5 tickets to be present, about six times as many as were actually sampled.

If we were satisfied with a 90% level of confidence, the sample size would be

$$n = \left(t_{df,\alpha/2}\frac{s}{E}\right)^2 \approx \left(z_{\alpha/2}\frac{s}{E}\right)^2 = \left(1.645 \cdot \frac{300}{50}\right)^2 = 97.42 \approx 98 \text{ tickets} \qquad (5)$$

[5]The average standard deviation of the 1Q2014 O&D sample for North Dakota airports was $305.

Now that these generic sample sizes have been calculated, we can move to the next phase of the problem: calculating the sample rates necessary to draw at least this many tickets from a large sample of tickets.

Computing Necessary Sample Rates

In the previous section, we found that in order to accurately estimate average fare from an airport with a $500 true average fare within +/-$50 with a 90% level of confidence, we would need to sample 98 tickets from this airport. Now, we must compute what percentage of tickets need to be sampled from the entire pool of available tickets to ensure that at least 98 tickets from the airport in question are drawn. This sample rate will depend on the total number of tickets from this airport that are available in the pool.

In statistics, binomial distributions are often used to model repeated draws from a sample population. These draws are often called "Bernoulli trials." In a Bernoulli trial, a draw can result in one of two outcomes: success or failure. In this case, "success" would represent a random ticket drawn from the entire pool of ticket that matches the target airport. A failure would be a ticket drawn from any other airport. Suppose that in a pool of 250,000 itineraries, 500 of those tickets originate from our small airport. In this case, the probability of success would be 500/250000 = 0.002

With a binomial distribution, the probability that exactly k successes will be obtained after n Bernoulli trails with success rate p is:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

However, for very large values of n, computing $\binom{n}{k}$ is computationally inefficient. In cases with large n and small p, such as our example, a Poisson distribution can be used instead to model the binomial distribution. This is also known as a Poisson approximation, or the Poisson Limit Theorem. In this case, we can use the following formula:

$$P(X = k; \lambda) = \frac{e^{-\lambda}\lambda}{X\,X!}$$

where $\lambda = n\pi$ is a parameter that represents the expected number of successes from a sample size of *n*. Since π, the probability of successfully drawing a ticket from our small airport, is known, we will solve for the sample rate *n* necessary *for the entire sample* to draw at least 98 tickets from our small airport.

That is, we wish to solve for *n* in the following equation:

$$P(X \geq 98; \lambda) = 1 - \sum_{x=1}^{98} \frac{e^{-\lambda}\lambda x}{x!}$$

This equation can be solved in Excel using its Solver functionality. The resulting sample rates necessary to draw 98 tickets (sufficient to accurately compute average fares 10% with a 90% confidence level) for small airports of various sizes are shown below in Table 1, assuming a $500 average fare and a $300 standard deviation.

	Airport Size			
Quarterly Passengers	200	500	1,000	5,000
Estimated Total # of Tickets	200	500	1,000	5,000
90% Confidence	55.4%	22.2%	11.1%	2.2%

95% Confidence	57.4%	23.0%	11.5%	2.3%
98% Confidence	59.7%	23.9%	11.9%	2.4%

Table 1: Sample Rates to Draw at least 98 tickets from a large population of tickets. Assumptions: Average Fare = $500, Standard Deviation = $300, Pax/Ticket = 1.0.

Note that we now have two separate confidence values. The rows of Table 1 indicate the levels of confidence that we will draw at last 98 tickets from a given airport out of the total pool of available tickets. In turn, drawing at least 98 tickets is sufficient to predict average fare (10%) 90% of the time. Then, given these two confidence levels, what is the true probability that we will be successful in predicting the average fare at a given airport?

The key is the realization that these two events are *statistically dependent*. That is, we need to draw at least 98 tickets from the pool of available tickets to ensure that the resulting sample can estimate the true average fare (within acceptable bounds) 90% of the time. In other words, if we do not draw at least 98 tickets from the target airport in our sample, then the probability that the resulting sample's average fare is a good estimate of the actual average fare is no longer 90%.

We can estimate the probability of two statistically dependent events as follows. Let $p(A)$ represent the probability that event A occurs, and $p(B|A)$ represent the probability that event B occurs *given* event A. Then, the probability that both events A and B occur is:

$$p(A \wedge B) = p(A) \cdot p(B|A)$$

For our example, let Event A be "The sample contains at least 98 tickets from our given airport" and Event B be "The sample predicts the average fare at that airport within ±10%." Then, for an airport with 500 quarterly passengers and a 23% sample rate, $p(A) = 95\%$ (from Table 1) and $p(B|A) = 90\%$ (from Equation 5). Hence, the probability that this sample rate will accurately predict average fare at that airport is $90\% \cdot 95\% = 85.5\%$.

DOT has indicated that they wish to accurately predict average fares at airports with 90% confidence. One way to achieve this goal would be to find the sample rates necessary to draw at least 139 tickets[6] from the target airport at least 95% of the time. Then, both $p(A) = p(B|A) = 95\%$, and the probability that this sample would provide a good estimate of average fares for that airport would be $95\% \cdot 95\% = 90.25\%$.

Table 2 shows the sample rates necessary to draw at least 139 tickets for small airports of various sizes with 95% confidence. These sample rates would thus be sufficient to predict average fares for a small airport of that size within ± 10% at least 90.25% of the time.

	Airport Size			
Quarterly Passengers	200	500	1,000	5,000
Estimated Total # of Tickets	200	500	1,000	5,000
Sample Rate	79.5%	31.8%	15.9%	3.2%

Table 2: Sample Rates to Predict Average Fares within ± 10% With 90% Confidence. Assumptions: Average Fare = $500, Standard Deviation = $300, Pax/Ticket = 1.0.

As Table 2 shows, a sample rate of about 32% would be sufficient to predict average fares 10% at small airports with 500 quarterly passengers with 90% confidence. For airports with 1,000 quarterly passengers, a sample rate of 16% would be sufficient.

These sample rates could be reduced if the acceptable average fare margin of error for small airports was different than that for larger airports. For instance, if the margin of error for fares was relaxed to 20%, or $100, a 22.6% sample rate would be sufficient for airports like DVL with about 200 quarterly passengers at 90% confidence.

Performance of DOT's Current 10% Ticket Sample Rate

Assuming a $500 average fare, a $300 standard deviation of fare, and an average passengers/ticket ratio of 1.0, DOT's current sample rate of 10% will generate average fares within 10% of the true mean for airports with about 1,550 quarterly passengers[7] at a confidence level of 90%. Sample estimates of average fare for airports with less than 1,550 quarterly passengers will not be within 10% of the true sample mean with 90% confidence.

Of the 692 airports for which data is currently reported in the T-100, 520 airports (75%) had at least 500 quarterly passengers in 1Q2014, and 387 airports (56%) had at least 1,550 quarterly passengers. That is, the current 10% DOT ticket sample rate for O&D is estimated to be sufficient to compute average fares for 56% of U.S. airports with a 90% level of confidence. Increasing this sample rate to about 32% would result be sufficient to compute average fares at 75% of U.S. airports at the same confidence level. However, given the uncertainty in the data and the high standard deviation of average fares, a near-100% level of sampling would be necessary to compute sufficiently-confident average fares for all U.S. airports.

It is possible, since the selection of OD markets served from each small airport is fairly limited, that the standard deviation in fares for very small airports would in fact be lesser than the standard deviation for large airports, in which many types of short-haul and long-haul itineraries may exist for business and leisure passengers. If the standard deviation for smaller airports was in fact $200 instead of $300, then a sample size of just 15.1% would be sufficient to compute average fares at airports with 500 quarterly passengers with a 90% confidence level. Further research would be necessary to better estimate the standard deviation in average fare samples for small airports as opposed to larger airports.

Conclusions and Next Steps

Using a simple formula to compute sample size, we found that the DOT's current 10% sample rate is sufficient to estimate average fares within $50 for 387 U.S. airports with

at least 1,550 quarterly passengers with a 90% level of confidence. Increasing the sample rate to 32% would allow DOT to compute average fares within $50 for an additional 133 airports with at least 500 quarterly passengers-reflecting 75% of all U.S. airports. Additional research is required to confirm that average fare and standard deviation accurately reflects O&D data, and additional DOT input is required to specify the scope of precision necessary to improve O&D reporting of average fares in the future.

[6]We calculated in Equation 4 that a sample size of 139 tickets would be necessary to predict average fare within 10% with 95% confidence.

[7]About $\frac{1550 \cdot 2}{91.2} \approx 34$ passengers per day both ways.

APPENDIX G

Below are comments by Bureau of Transportation Statistics, Survey Programs Director Ms. Chou-Lin in regards to the analysis performed by Michael Wittman: "Computing the Sample Rates Necessary for BTS O&D Average Fare Estimates for Small Airports", by Michael D. Wittman. This paper reviews the current sampling rate for DOT's Origin and Destination (O&D) data set and computes the sampling rate necessary for improving the quality of the airfare estimates for small airports.

One key estimate of the O&D data is to calculate the average airfare at the national level and the airport level. The current method selects a large, simple random sample of 10% of tickets sold by US carriers, and hopefully a large enough sample will be selected from each airport. It appears that the resulting sample size is not large enough to produce a valid estimation for small airports. An alternative method employed to improve the estimates for small airports uses a standard formula to calculate the sample size necessary to estimate average airfare for tickets of each airport. Then it computes the necessary sampling rate to achieve the desired sample size for the airport in question.

The paper concludes that "the current 10% sample rate is sufficient to estimate average fares within $50 for 387 US airports with at least 1,550 quarterly passengers with a 90% level of confidence. Increasing the sample rate to 32% would allow DOT to compute average fares within $50 for an additional 133 airports with at least 500 quarter passengers- reflecting 75% of all airports".

The advantage of this method is simplicity: it's easy for the airlines to carry out the sample selection, resulting in a simple random sample, and the analysis is less complicated. The disadvantage of this method is that the sample size for each airport is a random number and some of the small airports may not have enough sampled cases. There are other methods to select the desired number of sample cases that will produce an estimate with a pre-specified confidence level. One common customized sampling method is stratified sampling.

Assume that BTS wants to accurately predict average fares with, say a10% margin of error at the airport level with 90% confidence. The following steps are used to allocate the samples.

First, we should calculate the sample size by airport. The following formula can be used to estimate sample size for airport a:

$$n_a = \left(\frac{z_{1-0.05} s_a}{0.1 \hat{\mu}_a} \right)^2$$

Here $z_{1-0.05} = ¿$ 1.645, s_a^2 is the standard deviation of airfare for airport a estimated from historical data, $\hat{\mu}_a$ is the average airfare estimated from historical data. Ideally, we should first use historical data to estimate s_a^2 and $\hat{\mu}_a$, then we can calculate the customized sample size n_a for each airport a.

To specify the sampling parameters for the airlines, we also need estimates of the historical total itineraries of each airport - $\widehat{N}_a$ so we can estimate the airport sampling rate: $n_a/\widehat{N}_a$. For example, if $n_a/\widehat{N}_a = 0.05$, we can tell the airlines to sample the itineraries using the last two digits of the ticket number, e.g. 10-14.

Since this method incurs unequal selection probabilities, a design weight should be calculated: $w_a = ¿\ N_a/n_a$ for all selected cases in airport a. Because of this, specialized software is needed for estimation.

This method can reduce the sample size dramatically, but it makes the sampling and estimation more complicated. Consequently, the airlines need to specify different sampling rates for different airports (origins). Also, the resulting sample is no longer a simple random sample. This may cause confusion for end data users.

Summary

If the cost associated with a large sample size is not a problem, the universal sample allocation (current method) is a better method. To determine the new sampling rate for the future data selection and to access its impact on the estimates, it would need further confirmation of the data used in this paper and specify the desired level of precision for the estimates.

Chou-Lin Chen
March 25, 2019