How Close is Close Enough? Statistical Equivalence of Onboard versus Online Surveys of Transit Customers

by

Ben Cummins
Research Intern, Resource Systems Group, Inc.
M.R.P. Candidate, Dept. of City and Regional Planning, Cornell University
319 S. Albany St.
Ithaca, NY 14850
520.245.5414
bhc44@cornell.edu
(Corresponding author)

Greg Spitz
Director, Resource Systems Group, Inc.
55 Railroad Row
White River Junction, Vermont 05001
802.295.4999
gspitz@rsginc.com

Tara P. O'Malley
Market Research Coordinator, Chicago Transit Authority
567 West Lake St.
Chicago, IL 60680
312.681.4249
tomalley@transitchicago.com

Margaret Campbell
Senior Consultant, Resource Systems Group
55 Railroad Row
White River Junction, Vermont 05001
802.295.4999
margaret.campbell@rsginc.com

Submitted November 15, 2012, to the 2013 Transportation Research Board Annual Meeting

Word count: 3,921 words + 14 figures and tables = 7,421 words

ABSTRACT

Stagnant budgets and growing rates of internet access have increased the appeal of substituting traditional survey methods for electronic ones. Online surveys are particularly appealing in the public transit industry, due to the expense and logistical difficulty of surveying customers onboard buses and trains. It is therefore critical to understand, quantify, and test the differences between onboard versus online transit survey data.

Traditional hypothesis tests are designed to show that two sample statistics likely come from different populations. However, failing to find a difference cannot be interpreted as evidence that there is no difference. Furthermore, a difference may be statistically significant, but so small as to provide no practical insight (which often happens when working with large sample sizes). Statistical Equivalence Testing (SET) provides an analytical framework with which to evaluate whether two datasets are similar enough to be interchangeable (i.e., statistically equivalent).

The paper describes statistical equivalence tests conducted on customer satisfaction data collected onboard transit systems and data collected electronically via email lists from users of the same systems. We compare proportions of satisfied customers across various economic and travel behavior characteristics between these datasets. Within our chosen threshold of .05 (statistics within five percentage points of one another), one of the two datasets evaluated shows strong evidence of equivalence between onboard and online survey methods, while the other dataset shows strong evidence of nonequivalence. Findings support the idea that, at least in some cases, online surveys can substitute for onboard ones.

Introduction

Market researchers in transportation and other fields often find themselves faced with a choice between two survey methods: a more expensive and ostensibly more probabilistic approach, and a cheaper, perhaps less probabilistic one. This issue is becoming more and more common as electronic communication (e.g. email and the internet) and devices become ubiquitous (*1*). This study is rooted in evidence that significant coverage error can be relatively easily avoided when sampling transit users online, as provided in previous research by Spitz and Smith (*2*).

Substituting more expensive forms of surveying with less expensive forms is ever more appealing, especially if the difference between the two methods can be demonstrated to be negligible. Therefore, it is critical to understand, quantify, and test the differences between directly recruited respondent datasets (onboard recruitment in the case of transit) and online survey datasets via an email recruit method or via other methods (e.g. twitter, web page, Facebook, etc.)

Understanding whether or not these datasets are statistically equivalent is particularly helpful in cases where studies are performed on a regular (e.g. annual) basis. By comparing paper and electronic methods in a base year, agencies may be able to perform comparable annual surveys without incurring the full expenditure of conducting a major in-field survey each time. Instead, these surveys could be conducted completely online via email recruiting and/or other low cost recruiting methods and produce data that is comparable to much higher cost field efforts. The equivalency will need to be tested over time to ensure it maintains itself (assuming it was present in the first place), but it is most likely that as even more people gain access to electronic communication, equivalency will only improve.

Statistical Equivalence Testing (SET) provides an analytical framework with which to evaluate whether two datasets are similar enough to be interchangeable (i.e. statistically equivalent) (*3*). Sample statistics (e.g. means or proportions) may be used to compare different populations using traditional hypothesis testing. However, even in the case of a statistically significant difference, that difference could be so small as to provide no practical insight to the researcher. For example, imagine that a large study finds a sample mean income of $100k on one transit line, and a sample mean income of $95k on another line. This difference is found to be significantly different at α=.05 (something that could be done with a sample size of just a few hundred). However, is this difference actionable for the transit agency? Likely not. Instead, they may be more interested in knowing whether the populations are within $10k of one another.

Additionally, failing to reject the null hypothesis in a traditional test (i.e. finding no statistically significant difference between the two groups) cannot be interpreted as evidence that the two groups are the same (*4*). This is simply a reiteration of the saying that “the absence of evidence is not evidence of absence” (*5*). This paper demonstrates how, by specifying a threshold of indifference, we can construct hypothesis tests to determine whether two population parameters are likely to be similar enough for our research purposes.

We perform statistical equivalence tests on customer satisfaction data collected onboard transit systems and data collected via email from users of the same systems. Survey data was collected onboard Chicago Transit Authority (CTA) trains and buses, and is compared to data collected through emails to CTA customers. Separate data was collected onboard Metra commuter rail, and is compared to email surveys of Metra customers.

Background on Statistical Equivalence Testing

In all social science research, null hypothesis statistical testing (NHST) is the primary tool employed when testing whether two samples come from different populations. However, by the very nature of the test, a researcher can never arrive at the conclusion that two samples come from the same population. As mentioned above, failing to reject the null hypothesis does not provide any evidence that it’s true. Furthermore, with a large enough sample size, NHST will reveal even the smallest difference between two populations to be statistically significant (*6*). In cases involving modest sample sizes, it may be appropriate to use NHST to add validity, say, to the sign of a coefficient. However, if we are looking for evidence that two populations are satisfactorily similar, it will not suffice.

To solve this problem, we look to a methodology that was developed in biostatistics as a way to ensure an alternative treatment is as effective as the known treatment. This methodology, namely SET, has since been applied in several other fields, though as far as we know, not yet market research (7). SET shifts the question of NHST. While the NHST test for the difference between two means or proportions asks:

- What is the probability of having a difference this large assuming the populations are the same?

The SET test for the difference between two means or proportions instead asks:

- What is the probability of having a difference this small assuming the populations are non-trivially (in the eyes of the researcher) different?

As structured, SET can provide a researcher with evidence that two samples indeed come from equivalent populations (*3*).

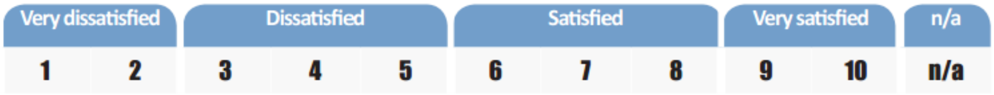
To make this possible, it is up to the researcher to define a threshold— two parameters whose difference is within this threshold are considered equivalent, and parameters whose difference is greater than the threshold are not equivalent. This threshold is arbitrary; there is no mathematical justification, just as there is no justification of using a whole host of statistical rules of thumb; these include 95% confidence intervals, plus or minus 5% error, sample sizes of 400, etc. (*3*). This threshold should be equal to the maximum difference between the two parameters of interest (e.g. means or proportions) that the researcher would consider trivial, and should be set prior to beginning any analysis (*3*). For example, a consumer products company manufactures toothpaste at two different plants. Rather than prove that the two plants produce tubes of toothpaste with mean net weights that are identical to one another (a technically impossible task), they may prefer to make sure that the means at the two plants are likely to be within 1 mg of one another.

Just as failing to reject the null hypothesis in a traditional hypothesis test cannot be taken as evidence that the samples come from the *same* population, likewise failing to demonstrate equivalence is not the same as demonstrating *nonequivalence* (*8*). This is especially important to remember when dealing with modest sample sizes. Larger sample sizes increase the power of an equivalence test (i.e. its probability of demonstrating equivalence where it exists), and demonstrating equivalence requires a larger sample size, generally, than showing a difference in a traditional test (*5*).

Analysis of Two Case Studies

In this paper, we look at the proportions of customers who report being satisfied with the overall service for two Chicago transit agencies: Chicago Transit Authority (CTA), which provides rail (rapid rail/subway) and bus service; and Metra which provides commuter rail throughout greater Chicagoland. Satisfaction was indicated by respondents on questionnaires that used a 10-point scale, but is analyzed as a binary variable with scores of 1-5 constituting "dissatisfied" and 6-10 constituting "satisfied" (SET need not be performed with binary proportions, means or other statistics can also be used). This corresponds with labels on the questionnaire, as shown in Figure 1.

Figure 1: Questionnaire scale



Data were collected using two distinct survey methods. The first group was surveyed directly onboard trains and buses, with the option to complete the survey immediately, mail it in, or complete it online. The second group was contacted through email and completed the surveys online.

The response rates for CTA were approximately 20% onboard the trains and buses and 19% through email, yielding 5,443 onboard responses and 7,668 email responses. On Metra, response rates were roughly 42% for onboard customers and 8% by email, generating 11,698 surveys from customers onboard and 4,287 through email. This information is presented in Table 1.

Table 1: Survey response

	Response Rate	Surveys Collected
CTA		
onboard	20%	5,443
email	19%	7,668
METRA		
onboard	42%	11,698
email	8%	4,287

CTA customers' email addresses were provided by customers who agreed to participate in future market research studies when they registered to receive a Chicago Card Plus, a widely-used transit payment card. Metra customers' email addresses were taken from a voluntary service alert list and a marketing database.

Statistical Methods to test for Equivalency

The null and alternative hypotheses for a two-sided test for equivalence of two proportions are as follows:

H_0: $|\pi_1 - \pi_2| > \delta$

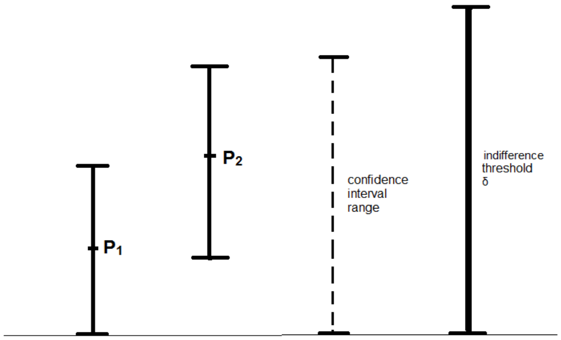
H_A: $|\pi_1 - \pi_2| \leq \delta$

where π_1 and π_2 represent proportions of the same question from samples 1 and 2, respectively, and δ is the equivalency threshold selected by the researcher.

As can be seen in the null hypothesis, if the difference between the two population proportions is greater than the threshold, then they are not statistically equivalent. If the differences are less than the threshold, then they are equivalent (i.e. the alternative hypothesis holds true).

The simplest way to conduct the test is to construct confidence intervals around the sample proportions. If the combined range of the two confidence intervals is less than the indifference threshold, the two sample proportions are equivalent. Figure 2 illustrates this.

Figure 2: Equivalence demonstrated through confidence intervals



Researchers may prefer to formalize the calculation into a hypothesis test, as we have done in this paper. The test statistic can be expressed as:

$$t = \frac{|P_1 - P_2| - \delta}{s_{P_1 - P_2}} \quad (5)$$

where $S_{P_1-P_2}$ is the pooled standard error of the two proportions:

$$S_{P_1-P_2} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$$

where n_1 and n_2 are the sample sizes of samples 1 and 2, respectively.

The two-tailed test for equivalency of two proportions is constructed as two simultaneous one-sided tests. Since the null hypothesis depends on an absolute value, we can divide it into two separate hypotheses, namely:

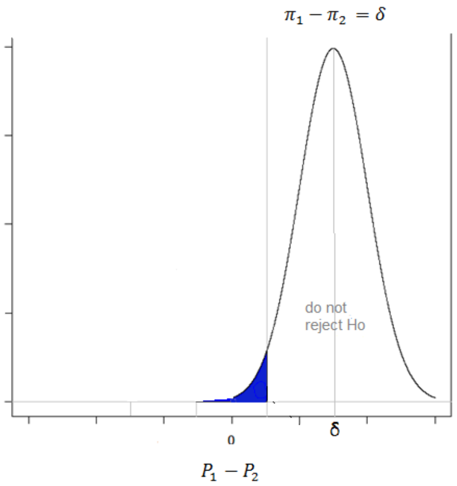
$$\pi_1 - \pi_2 > \delta$$

or

$$\pi_1 - \pi_2 < -\delta$$

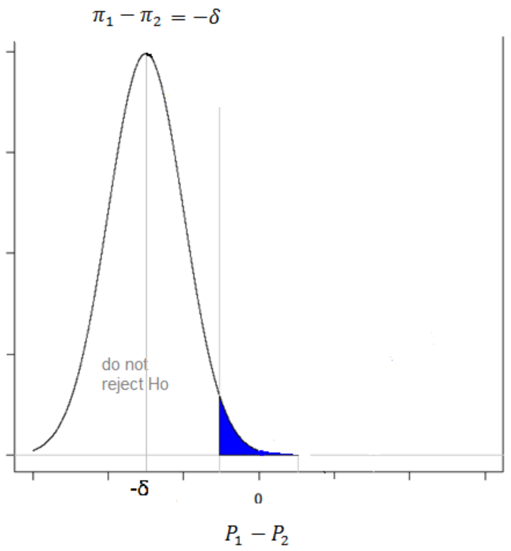
We start with "half" of the null hypothesis, and for the purposes of the test, we assume that $\pi_1 - \pi_2 = \delta$. We then construct a distribution of sample means centered on δ. If the difference P_1-P_2 falls in the left-sided tail of this distribution, we can say that it is unlikely that the true difference $\pi_1 - \pi_2$ is δ or greater. This is illustrated in Figure 3.

Figure 3: Right side of equivalency test



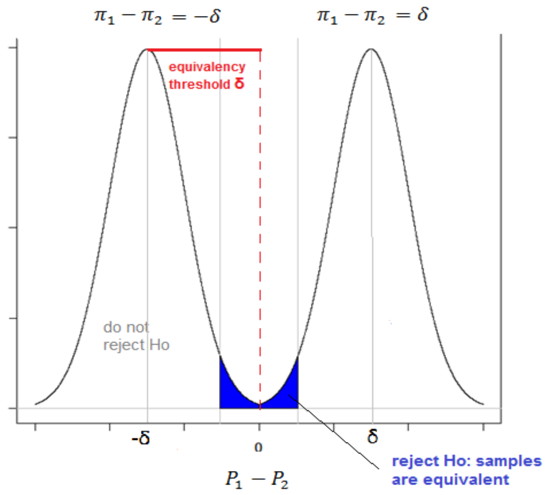
Likewise, if the difference between the sample proportions falls in the right tail of the distribution of sample means centered on –δ, we can say it is unlikely that the true difference $\pi_1 - \pi_2$ is –δ or less. This is illustrated in Figure 4.

Figure 4: Left side of equivalency test



If both of these conditions are satisfied, we conclude that the true difference is likely within our threshold of indifference (i.e. the sample proportions are statistically equivalent). The complete test is illustrated in Figure 5.

Figure 5: Two-tailed test for equivalence of two proportions



In the event that the absolute difference between the two sample proportions is greater than the indifference threshold δ, the samples are considered to be nonequivalent and no test is performed, as it is impractical and unnecessary. For example, if P_1 =.60 and P_2=.65 and the threshold is .04, the difference of .5 falls further from the mean than δ, which will always be in the "do not reject H_0" zone. However, if the threshold were .07, the test may or may not reject H_0, depending upon sample sizes and the corresponding standard errors.

For this study, we have selected an indifference threshold δ of .05. While selecting a threshold is necessarily arbitrary, this is based on the belief that a maximum difference of 5 percentage points is tolerable to most transit agencies when conducting customer satisfaction surveys. This threshold could, of course, be changed to suit a variety of needs. However, finding evidence for equivalence with smaller thresholds requires larger sample sizes. In the next section, we discuss the considerably large sample size required for a statistically powerful test.

Example Minimum Sample Size Calculation

As an example, let's consider a scenario, loosely based on our data, in which the online and onboard populations have proportions of 80% and 82.5% satisfied customers, respectively. We are seeking the commonly accepted statistical power of 80% (i.e. an 80% chance of accepting the alternative hypothesis of equivalency when the parameters are, in fact, equivalent). The indifference threshold is 5% (again, five percentage points, not a relative difference of 5%).

The equation for minimum sample size for a given level of power can be written as:

$$\frac{(Z_\alpha+Z_{\frac{\beta}{2}})^2[P_1(1-P_1)+P_2(1-P_2)]}{[(P_1-P_2)-\delta]^2} \quad (5)$$

where:

the power of the test is 1 – β

Z_α is the z-score associated with a significance level α

and $Z_{\frac{\beta}{2}}$ is the z-score associate with a significance level $\frac{\beta}{2}$

In our hypothetical, the recommended sample size at a significance level α of .05 comes out to 6,347 per sample group. In market research, this may be easily achieved for a very broadly segmented sample. However, dividing the sample into several categories (e.g. many levels of income) might require impractically large samples. Collecting enough data to demonstrate equivalence could prove as costly as doing the onboard study twice, thereby reducing the benefit of undertaking the study. Researchers should carefully consider to what extent they need to prove equivalence, and what the monetary costs may be before undertaking an equivalence study.

This study is conclusive, despite often falling short of this sample size. This is easily explained; the differences between sample proportions in the results discussed below are either much smaller or much larger than the parameters in our hypothetical.

RESULTS - CTA

For the CTA dataset, we have segmented the sample by household income, primary mode of travel (bus v. rail), and a five-level variable for why the customer chooses to use CTA services. Tests for equivalence across the two survey methods (onboard v. email) were conducted for each segment. Reported in the tables below are the sample sizes and sample proportions of satisfied customers for each segment and survey method. Below that, the difference between the two sample proportions (onboard v. email) for each segment is reported, along with the standard error of that difference. Finally, the student's t test statistic and corresponding probability value are reported.

δ = .05

RED denotes equivalence at α=.05

YELLOW denotes equivalence at α=.10

Table 2: CTA overall satisfaction by household income

		<$25k	$25k-	$60k-	>$100k
			$60k	$99k	
ONBOARD	N	1209	1158	735	598
	P(sat)	0.797	0.825	0.865	0.858
EMAIL	N	589	1955	2156	2588
	P(sat)	0.803	0.820	0.840	0.847
	difference	**0.007**	**0.005**	**0.025**	**0.011**
	SE of diff	0.020	0.014	0.015	0.016
	t stat	**2.167**	**3.197**	**1.692**	**2.430**
	p=	**0.038**	**0.002**	**0.095**	**0.021**

Table 3: CTA overall satisfaction by primary mode

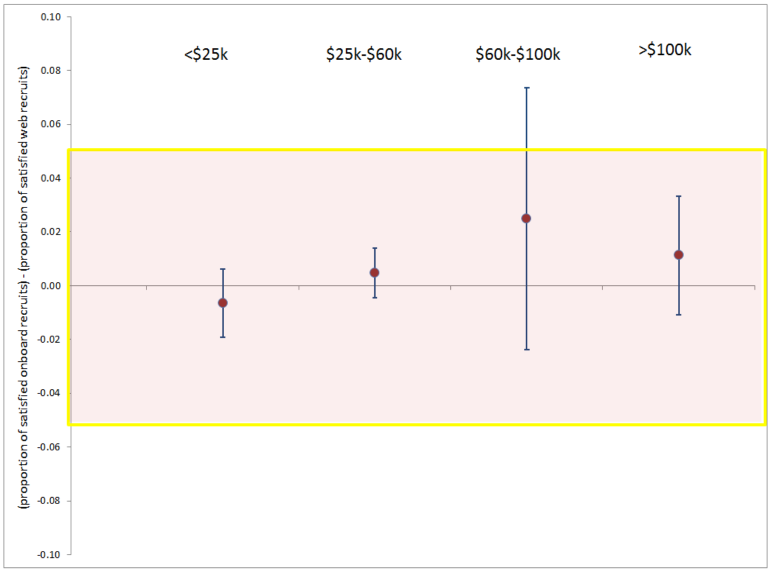
		CTA	CTA
		bus	**rail**
ONBOARD	N	2386	1675
	P(sat)	0.819	0.846
EMAIL	N	2052	2727
	P(sat)	0.818	0.829
	difference	**0.001**	**0.017**
	SE of diff	0.012	0.011
	t stat	**4.237**	**2.910**
	p=	**0.000**	**0.006**

Table 4: CTA overall satisfaction by reason for riding CTA

		I can’t drive/don’t have a car available	I don't have a car because I prefer to take the bus or train for all of my trips	I have a car available but prefer to take the bus or train for some purposes	I prefer to take the bus or rail for most purposes, but have or use a car for special trips or emergencies
ONBOARD	N	1798	637	948	743
	P(sat)	0.780	0.887	0.859	0.857
EMAIL	N	1176	1165	3053	1881
	P(sat)	0.764	0.868	0.841	0.857
	difference	0.015	**0.019**	**0.017**	**0.000**
	SE of diff	0.016	0.016	0.013	0.015
	t stat	**2.201**	**1.928**	**2.504**	**3.276**
	p=	**0.035**	**0.062**	**0.017**	**0.002**

Most proportions are equivalent at α=.05, and those that are not are close. Figure 6 below shows the confidence intervals surrounding the difference between the two proportions for the four household income segments. The shaded area represents the indifference threshold. As part of the confidence interval for the $60k-100k range lies within the equivalence range and part lies outside of it, this result is considered ambiguous. A more powerful test (i.e. larger sample size) would be required to demonstrate equivalence or nonequivalence. However, given that most of the confidence interval lies within the equivalence threshold, we can say it is likely that the two sample segments are equivalent.

Figure 6: Confidence intervals for the difference between CTA onboard and email sample proportions, by household income



With this evidence, we can conclude that these two methods are probably equivalent within our specified tolerance.

RESULTS – Metra

The surveys for CTA and Metra were not identical, so it was not possible to segment the samples by the exact same variables. For the Metra data, we have instead segmented the sample by household income, number of monthly trips, and automobile availability. As above, the sample sizes, proportions of satisfied customers, differences between those proportions for the two methods, standard errors of those differences, test statistics, and p-values are presented in the figures below

Table 5: Metra overall satisfaction by household income

		<$25k	**$25k-**	**$60k-**	**>$100k**
			$60k	**$99k**	
ONBOARD	N	360	1333	2286	4229
	P(sat)	0.900	0.894	0.887	0.899
EMAIL	N	215	703	1357	2011
	P(sat)	0.805	0.829	0.836	0.832
	difference	0.095	0.065	0.051	0.066
	SE of diff	0.031	0.017	0.012	0.010
	t stat	**N/E**	**N/E**	**N/E**	**N/E**
	p=				

Table 6: Metra overall satisfaction by number of monthly trips

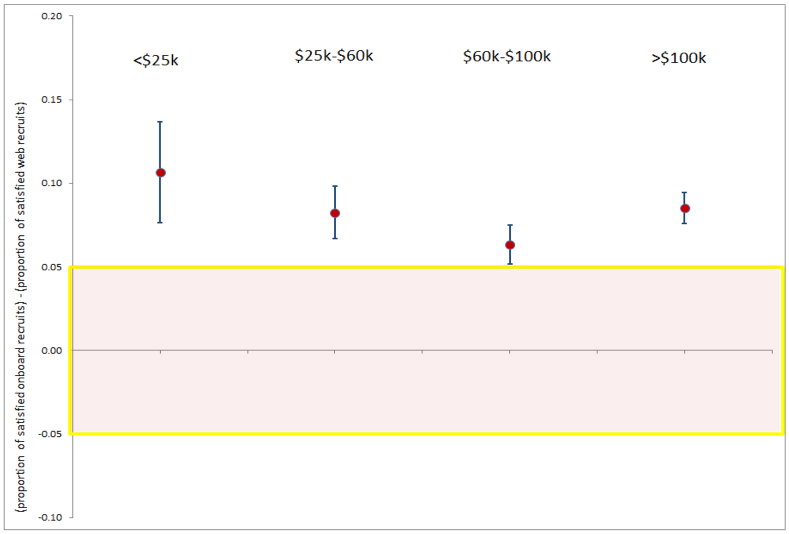
		Less	**10-19**	**20-29**	**30-39**	**40+**
		than 10				
ONBOARD	N	756	657	1678	1056	4647
	P(sat)	0.923	0.915	0.889	0.907	0.873897
EMAIL	N	525	279	643	412	2035
	P(sat)	0.946667	0.845878	0.790047	0.832524	0.790172
	difference	-0.023	0.069	0.099	0.075	0.084
	SE of diff	0.013779	0.024206	0.017796	0.020449	0.010256
	t stat	**1.931453**	**N/E**	**N/E**	**N/E**	**N/E**
	p=	**0.061779**				

Table 7: Metra overall satisfaction by automobile availability

		Yes	No
ONBOARD	N	7874	1366
	P(sat)	0.888	0.903
EMAIL	N	3814	472
	P(sat)	0.830	0.843
	difference	0.058	0.060
	SE of diff	0.007	0.019
	t stat	**N/E**	**N/E**
	p=		

Unlike with the CTA data, the Metra onboard and email samples do not appear to be taken from equivalent populations. None of the segments yielded equivalent proportions of satisfied customers at α=.05. Comparing Figure 6 to Figure 7 nicely illustrates the difference between the CTA and Metra results. In the Metra case, as the entire confidence interval for each segment falls outside the indifference threshold, we can say with confidence that the samples are nonequivalent.

Figure 7: Confidence intervals for the difference between Metra onboard and email sample proportions, by household income



Discussion/Conclusion

It has been clearly demonstrated that the first of these two datasets generally satisfies tests for equivalence between onboard on online samples. This is a finding of some significance, as it demonstrates that CTA can substitute efficient and inexpensive surveys for onboard surveys by simply emailing their Chicago Card Plus list. This may save considerable resources. We recognize that politically this will not be easy to do, and more data needs to be collected to continue to check and make sure the equivalency holds true, but it is an important fact for CTA that their Chicago Card Plus dataset appears to be as representative as sampling onboard transit vehicles, assuming an indifference threshold of 5%.

The second of the two datasets showed equally clear nonequivalence, which leads us to conclude that the way an online customer list is compiled can significantly affect whether or not it is representative of the riding population as a whole. Finding cost-effective, ethical, and representative methods of online recruitment is a critical area of future research. However, our findings support the idea that, at least in some cases, online surveys can substitute for onboard ones.

This paper is intended to be a first step in applying the principles of statistical equivalence in the field of transportation market research. Methods of equivalence testing that are being applied in other fields must be further studied and considered for this new application. Based on this paper, however, we are confident that equivalence testing can have an important place in transportation market research as a way to understand how to most cost effectively sample transportation customers, especially as more and more customers obtain internet access. Confidently substituting economical, online survey methods for traditional methods may ultimately prove to be both practical and necessary, as the need for customer data increases in spite of stagnant or decreasing budgets.

REFERENCES

1. Pew Research Center. *Pew Internet and American Life Project Surveys: Device Ownership Chart.* http://pewinternet.org/Static-Pages/Trend-Data-(Adults)/Device-Ownership.aspx. Accessed August 15, 2012.
2. Smith, C. and Spitz, G. "Internet Access: Is Everyone Online Yet and can we Survey them There?" In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2176, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 35-41.
3. Streiner, D.L. "Unicorns Do Exist: A Tutorial on 'Proving' the Null Hypothesis." Canadian Journal of Psychiatry Vol 48, No 11, 2003, pp. 756–761
4. Cohen, J. "The Earth is Round (p<.05)"*, American Psychologist*, Vol.49. No. 12, 1994, pp. 997-1003
5. Wellek, Stefan. "Testing Statistical Hypotheses of Equivalence and Noniferiority." CRC Press, Boca Raton, FL, 2010
6. Cohen, J. "Things I Have Learned (So Far)." American Psychologist, Vol. 45, No. 12, 1990, pp. 1304-1312
7. Limentani, G. B.; Ringo, M. C.; Feng, Y.; Bergquist, M. L.; MCSorley, E. O. "Beyond the t-test: Statistical Equivalence Testing." Analytical Chemistry, Vol. 77, 2012, pp. 221A–226A
8. Tryon, W.W. "Evaluating Statistical Difference, Equivalence, and Indeterminacy Using Inferential Confidence Intervals: An Integrated Alternative Method of Conducting Null Hypothesis Statistical Tests." Psychological Methods Vol.6, No. 4, 2001, pp. 371-386

LIST OF FIGURES AND TABLES