Listening Analysis of Personally-Recruited Panelists using Wilcoxon Tests

Abstract

The two-sample Wilcoxon test is commonly used as a nonparametric alternative to the two-sample t-test, particularly in the case of skewed distributions or when any other non-normal distribution is evident. The test is utilized in a modality analysis for a city-wide panel sampling.

Introduction

Arbitron is a media research company whose core business is the production of ratings for radio broadcast stations across the United States. To achieve this end, the company enlists households from the larger media markets to join a *radio panel*. Each household member over age 6 becomes a *panelist* and agrees to wear a small electronic device, or meter, that can detect an encoded signal broadcast by local radio stations. The meter data are analyzed and used to produce ratings for the stations in each market.

The Portable Person Meter (PPM) is a small apparatus, the size of a pager, which transmits a cellular signal containing the listening and motion data each night to Arbitron. The listening data is tabulated into the ratings based on whether the panelist actually wore their meter long enough during the course of the day. Panelists agree to wear the meter for a period not exceeding two years.

Households are contacted to join the PPM panel through a combination of mailings, phone calls, and finally in-person recruitment (IPR) for a subset of non-respondents. Each market is subdivided and a stratified sample is selected from an address-based frame. Most households are recruited through a mixture of phone and mail contact, but not every selected household can be reached in this manner, even after repeated attempts. Through non-response studies, we found that such households tended to contain younger and more Black and Hispanic persons. In order to ensure a more representative panel, Arbitron was compelled to initiate *in-person recruitment* (IPR).

IPR is an effort to improve representation of younger and more ethnic households We wish to investigate, using the two-sample Wilcoxon test, whether the recruitment method (IPR vs. non-IPR or *Other*) has any impact on panelists' listening. Furthermore, we would like to control for a number of socio-demographic factors known to influence radio listening. Once installed, the meters collect listening data identically, regardless of the method in which the household was contacted to join the panel. Therefore, our assumption is that radio listening is independent of the recruitment type (IPR vs. Other).

Many surveys rely on multiple methods to select respondents and gather data from the population. The 2007 Health Information National Trends Survey (HINTS) used a combination of RDD respondents that were interviewed over the phone, in addition to mailed surveys. If certain important questions are left blank in the Current Population Survey (CPS), a representative is sent to the address to personally interview the respondent. Even the Census Bureau, whose surveys are required by law to be completed by households, implements in-person follow-up operations for non-respondents. Thus, multiple modes of data collection are prevalent in many surveys as a means to improve non-response and obtain a sample that better represents the population. A central question is whether there exists a modality effect on respondent behavior for a particular survey.

We will investigate the existence of a modality effect in terms of radio listening based on the recruitment type.

The standard unit of measurement in the radio industry is the *quarter hour* (QH). Every broadcast hour is subdivided into 4 quarter hours of listening. In practice, a panelist only needs to listen to a station for 5 minutes (not necessarily consecutive) in order for the station to gain credit for that quarter hour. This analysis will examine the quarter hours tabulated from IPR panelists compared to panelists recruited using mail/phone methods. (Note in the analysis we refer to the mail/phone method of recruitment as *Other*). Currently, IPR recruitment has been implemented in all 48 markets (metropolitan areas) measured by the PPM service. The initial waves of recruitment corresponded to high density ethnic areas, HDA's, and it is with respect to these areas which we will aim to assess the differences in the QH listening.

Before discussing the analysis, we make the important distinction that we are not conducting *descriptive* inference on the *finite population*, but rather making inferences on the *sampled population*. In general, care must be taken when applying common statistical procedures to survey data as it must account for the complex sampling. In such cases, the actual estimates may be accurate (i.e., means, percentages, regression parameters), but the variance of those estimates are not properly computed so that anything computation based on the variance estimators are erroneous (i.e., confidence intervals and statistical tests). In our case, we are considering tests on the panelists *themselves* rather than any segments of the population they are meant to represent. Thus, we are allowed to make use of a wider variety of statistical tools and are not restricted to computational methods in the SAS/SUDAAN/STATA survey procedures.

Background on the Wilcoxon Rank Sum Test

We wish to measure whether there is a difference between the listening habits of IPR panelists and non-IPR panelists (Other) by comparing the mean number of quarter hours of listening between the two groups. As such, the usual avenue would be to use the unbalanced version of the t-tests. Although it's possible that the equal variance assumption may be met, the usual Welch t-test (SAS uses the Satterthwaite approximation for the degrees of freedom) would usually suffice whether or not the variance assumption holds. However, we will see that the radio listening data follows a skewed distribution, which would violate the normality assumption inherent in all of the various t-test methods. Thus we will instead appeal to the non-parametric alternative: the two-sample rank sum test.

Assuming the same distributional form in two independent samples, the Wilcoxon two-sample rank sum test is commonly used to detect a shift in location between the populations. In such cases, the Wilcoxon test enjoys some nice properties. Even when the underlying distributions are normal with common variance, the Wilcoxon test is *nearly* as powerful as the appropriate t-test (the Wilcoxon has a relative efficiency of $\frac{3}{\pi}$ in comparison to the t-test). Despite the slightly lower efficiency, the Wilcoxon test is known to be pretty robust in comparison to the t-test (whose efficiency and optimality properties are decidedly *non-robust*). For the one-sided tests in particular, the Wilcoxon test is unbiased and is actually UMP for logistic distributions. [5].

The two-sample rank sum test is generally attributed to Wilcoxon [1] due to his 1945 publication, although an equivalent technique was developed in Imperial Germany by Gustav Deuchler some thirty years prior in 1914. The work was largely neglected, perhaps due to the turmoil and aftermath of the First World War. Apparently, no less than 7 independent proposals of the rank sum test were published from 1914 to 1952, a test which Kruskal deems to be quite "natural" [2]. Due to the complexity of the statistical tables involved, an equivalent approach was developed by Mann and Whitney in 1947 [3] that simplified the computations involved and sometimes the Wilcoxon test is referred to as the Mann-Whitney U Test or the Wilcoxon-Mann-Whitney test. For a comprehensive account of the use of Wilcoxon statistics, see [6].

We consider the following non-parametric setting. Let X and Y be two random variables from the same family of distributions, differing only by a location parameter (i.e., mean). That is, $F_X(t) = F_Y(t-\theta)$ for some fixed θ. Our null hypothesis is that the distributions are stochastically equivalent, H_0: $F_X(t) = F_Y(t)$, or that $\theta=0$. Unlike the t-test, there are no distributional assumptions on the forms of F_X and F_Y.

There exist several equivalent formulations of the Wilcoxon test, including the Mann-Whitney version (which differs from the Wilcoxon statistics only by a constant and thus presents an equivalent test). In fact, the test can be considered as a specific case within the class of linear rank statistics [5]. The following treatise is taken from [4]. Now let $X_1, X_2,\ldots, X_{n1}$ and $Y_1, Y_2,\ldots, Y_{n2}$ be two independent and identically-distributed samples from X and Y. We now combine the two samples and place them in ascending order. Define R_i for $1 \leq i \leq n_1$ to be the rank of X_i in the combined sample ordering. Thus, the ranks of the X_i's will range from 1 to $n=n_1+n_2$. Now let…

Formula 1

$$T = \sum_{i=1}^{n_1} R_i$$

The intuition behind the test is that if the two samples are really from identical distributions, then the ranks of the X_i's should be regularly dispersed within the combined sample. It can be shown that under the null hypothesis, there is a function of T that has an approximately normally distribution (for larger samples) [4], and this is used to define the thresholds for large-sample Wilcoxon tests. To illustrate this, the Mann-Whitney U statistic can be alternatively defined as

Formula 2

$$U = T - \frac{n(n+1)}{2}$$

…where we recognize the second term as the sum of all the ranks of the total combined sample, $n=n_1+n_2$. It can be shown, with a fair amount of difficulty [4], that U has an approximately normal distribution with mean $\frac{n_1 n_2}{2}$ and variance $\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$, or equivalently…

Formula 3

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \sim N(0,1)$$

If n_1 and n_2 depict the sizes of the two independent samples, then one may define a *large* sample as $n_1 \geq 8$ and $n_2 \geq 8$. In that case, Mann and Whitney noted in 1947 that "the distribution is almost normal" [3]. For smaller samples, the exact distribution of U may be used – Wilcoxon tabulated the values in this case for up to 8 samples on each side. The exact null distribution of the Wilcoxon test is just based on the discrete distribution of the total number of combinations of possible rankings of the X_i's – which can become unwieldy for larger samples. So the joint distribution of the X_i's is given by

Formula 4

$$\Pr(X_1 = x_1, X_2 = x_2, ..., X_{n_1} = x_m) = \binom{n}{n_1}^{-1}$$

The two-sided Wilcoxon test rejects the null hypothesis (H_0: $\theta = 0$ vs. H_1: $\theta \neq 0$) when U is too large or too small. SAS provides p-values based on the both the exact distribution and the normal approximation. A test based on the t distribution is also provided; in most cases this will resemble the normal test and some distinction may arise in the case of samples not much larger than 8 apiece. In larger samples the selection is academic: the p-values will be (nearly) identical.

Data Description

Our analysis consists of data compiled from three months in 2012: January, February, and March 2012 within twelve Arbitron media markets, or *metros*. These "months" are really 28 day-periods that loosely correspond to calendar months, with an additional *holiday* month that begins some time in December. The *Arbitron media markets* define a collection of counties within major metropolitan areas that are used to categorize radio ratings. These closely resemble the more well-known *designated market areas*, or DMA's, that are similarly defined by the Nielsen Company to categorize television ratings.

Table 1 illustrates the number of panelists, by recruitment method Other or IPR, in the metros by the survey period.

Table 1

Recruitment Counts by Market	All Levels	MONTHYR					
		JAN12		FEB12		MAR12	
		Other	IPR	Other	IPR	Other	IPR
	Total	N	N	N	N	N	N
All Markets	34,745	9,159	2,419	9,077	2,474	9,138	2,478
New York	5,243	1,382	376	1,341	442	1,271	431
Los Angeles	5,529	1,365	436	1,426	429	1,418	455
Chicago	2,867	758	207	745	200	748	209
Philadelphia	1,573	445	97	414	107	404	106
San Francisco	2,419	607	171	615	182	644	200
Detroit	1,438	413	69	392	65	430	69
Washington, DC	1,646	452	98	435	112	428	121
St. Louis	902	230	68	229	60	259	56
Dallas-Ft. Worth	2,226	629	152	594	143	573	135
Atlanta	1,869	518	106	519	105	524	97
Phoenix	1,664	454	113	433	113	446	105
San Diego	1,476	399	108	376	101	389	103
Tampa-St. Petersburg-Clearwater	693	177	55	182	53	173	53
Miami-Ft. Lauderdale-Hollywood	5,200	1,330	363	1,376	362	1,431	338

Each panelist is a respondent in a daily survey for radio listening. When panelists wear their meter, as gauged by the amount of daily motion recorded by the PPM, they are admitted into a daily survey and assigned a daily weight, based on their demographic information and geographic location within the metro. The number of panelists participating by market is located in Table 1, while Table 2 lists the distribution of the sample with respect to important socio-demographic characteristics.

Table 2 illustrates the sample distribution with respect to recruitment method and select weighting variables.

Table 2

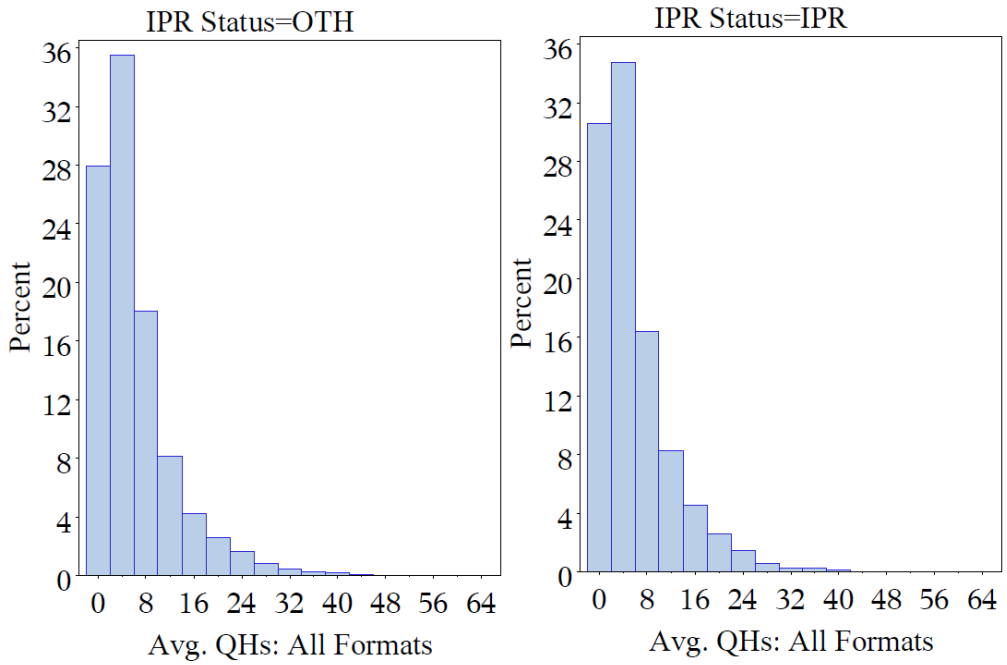
Panel Characteristics	All Levels	JAN12				FEB12				MAR12			
		Other		IPR		Other		IPR		Other		IPR	
	Total	N	Pct	N	Pct	N	Pct	N	Pct	N	Pct	N	Pct
GENDER_CODE													
Female	18,232	4,857	53.0	1,202	49.7	4,825	53.2	1,248	50.4	4,847	53.0	1,253	50.6
Male	16,513	4,302	47.0	1,217	50.3	4,252	46.8	1,226	49.6	4,291	47.0	1,225	49.4
Age Group													
6-17	6,644	1,701	18.6	534	22.1	1,657	18.3	542	21.9	1,676	18.3	534	21.5
18-24	4,164	1,013	11.1	350	14.5	1,042	11.5	356	14.4	1,050	11.5	353	14.2
25-34	4,612	1,174	12.8	381	15.8	1,133	12.5	393	15.9	1,131	12.4	400	16.1
35-44	5,216	1,367	14.9	384	15.9	1,341	14.8	386	15.6	1,355	14.8	383	15.5
45-54	5,904	1,560	17.0	409	16.9	1,545	17.0	417	16.9	1,549	17.0	424	17.1
55-64	4,606	1,301	14.2	226	9.3	1,292	14.2	240	9.7	1,303	14.3	244	9.8
65+	3,599	1,043	11.4	135	5.6	1,067	11.8	140	5.7	1,074	11.8	140	5.6
Race/Ethnicity/Language													
NH Black	11,141	3,003	32.8	772	31.9	2,875	31.7	798	32.3	2,902	31.8	791	31.9
Hispanic/English	5,431	1,423	15.5	395	16.3	1,428	15.7	392	15.8	1,394	15.3	399	16.1
Hispanic/Spanish	10,433	2,543	27.8	923	38.2	2,518	27.7	947	38.3	2,550	27.9	952	38.4
NH Other	7,740	2,190	23.9	329	13.6	2,256	24.9	337	13.6	2,292	25.1	336	13.6
Employment Status													
Full time	12,467	3,306	36.1	861	35.6	3,271	36.0	869	35.1	3,284	35.9	876	35.4
Other	15,634	4,152	45.3	1,024	42.3	4,149	45.7	1,063	43.0	4,178	45.7	1,068	43.1
<18 Years Old	6,644	1,701	18.6	534	22.1	1,657	18.3	542	21.9	1,676	18.3	534	21.5
Presence of Children Y/N													
Yes	18,409	4,637	50.6	1,478	61.1	4,612	50.8	1,514	61.2	4,675	51.2	1,493	60.3
No	16,336	4,522	49.4	941	38.9	4,465	49.2	960	38.8	4,463	48.8	985	39.7

Results of the Listening Analysis using Wilcoxon Rank Sum Tests

Before discussing the Wilcoxon tests, we first investigate some of the necessary assumptions to substantiate their validity. We consider the histograms of the listening distributions for IPR panelists and Other panelists. Note the skewness of the listening for

both – which violates the normality assumption of the t-test which hastens us to make use of the more robust non-parametric Wilcoxon test. Recall that the listening shown here is the average number of credited quarter hours per tabulated day. The Y-axis is the percentage of panelists, while the X-axis depicted the average listening. Note that the shapes of both distributions closely mirror one other. We may conclude that the listening between the different recruitment methods follows the same distributional shape. Our purpose, of course, is to verify whether the average (mean) listening differs and it would appear that we have satisfied the conditions for using the test.

Table 3 – Average Listening for IPR Panelists vs. Other Panelists (Two scales shown)



Recall that panelists are recruited at the household level, therefore the panelist-level listening does exhibit some dependency. We make some stipulations by considering only panelists ages 18+ to adjust for child-parent listening to an extent, although this is not entirely satisfactory. Another option would be to consider household level listening, but this would also require adjusting for the number of panelists within the household, which adds more complexity. We posit that although there is some dependence between panelists within a household, the effects are negligible. This assertion is based on prior knowledge of panelists and household listening.

At the macro level, it would appear from the charts above that the mean listening for both recruitment methods is identical. However, under large sample sizes, as is the current case, a number of statistical tests will reject similar null hypothesis because the sheer magnitude of the observations makes the test powerful enough to pick minute differences between populations. Nevertheless, we begin by running the Wilcoxon test on the full population. We run this separately for each month, and find that the test does not reject the null hypothesis for any of the three months. In the output below, note that the *Sum of Scores* for Other and IPR are equal to the sum of the ranks of a sample of size 9,343. That is, $1+2+\ldots+9343 = \frac{9343*9344}{2} = 43{,}650{,}496 = 34{,}897{,}452 + 8{,}753{,}044$.

Table 4 – Wilcoxon Tests for JAN12

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable total Classified by Variable iprStatus					
iprStatus	**N**	**Sum of Scores**	**Expected Under H0**	**Std Dev Under H0**	**Mean Score**
OTH	7458	34897452.0	34843776.0	104625.335	4679.19710
IPR	1885	8753044.0	8806720.0	104625.335	4643.52467
Average scores were used for ties.					

Wilcoxon Two-Sample Test	
Statistic	8753044.0000
Normal Approximation	
Z	-0.5130
One-Sided Pr < Z	0.3040
Two-Sided Pr > |Z|	0.6079
t Approximation	
One-Sided Pr < Z	0.3040
Two-Sided Pr > |Z|	0.6079
Z includes a continuity correction of 0.5.	

Thus, we find that based on the aggregated monthly estimates, the recruitment method has no significant impact on radio listening. Our main interest, of course, is the collection of tests stratified by Age, Gender, and Race/Ethnicity – factors which are known to considerably impact radio listening by station format. These factors (Age, Gender, Race/Ethnicity) are especially important to Arbitron's customers since stations specifically target certain groups defined by these factors. We repeat our Wilcoxon tests stratified by these important demographic characteristics, in addition to the survey month. Given the numbers in our sample, we can rely on the normal or t approximations rather than the exact values – which need to be specified in the PROC NPAR1WAY and are more computationally intensive.

Tables 5A and 5B depict the counts and listening averages for each of the cells in the Age, Gender, Race/Ethnicity crossings:

Tables 5A and 5B: Sample Counts and Average Listening Levels, by Age, Race/Ethnicity, and Gender (All Months Combined)

TABLE 5A Sample Counts	Black				Hispanic				Other			
	F		M		F		M		F		M	
	Oth	IPR	Oth	IPR	Oth	IPR	Oth	IPR	Oth	IPR	Oth	IPR
Age Group												
18-24	499	154	537	193	713	268	773	334	279	51	304	59
25-34	554	180	407	135	845	330	818	364	423	91	391	74
35-44	749	201	466	120	1,004	314	882	363	478	75	484	80
45-54	959	249	636	188	970	320	874	276	654	99	561	118
55-64	853	143	529	132	557	170	601	145	671	62	685	58
65+	613	76	354	59	611	86	485	97	565	64	556	33

Table 5B Listening Averages	Black				Hispanic				Other			
	F		M		F		M		F		M	
	Oth	IPR	Oth	IPR	Oth	IPR	Oth	IPR	Oth	IPR	Oth	IPR
Age Group												
18-24	5.0	5.2	4.6	3.6	5.9	6.1	5.3	5.8	4.8	4.7	4.4	3.2
25-34	6.6	4.3	5.1	6.8	6.4	5.9	8.0	7.3	4.5	5.3	5.3	4.7
35-44	6.7	6.3	8.2	8.2	6.7	6.7	7.7	7.6	5.5	7.0	6.2	6.7
45-54	7.1	5.9	8.4	9.7	6.9	6.6	9.7	9.6	6.2	7.5	7.9	8.9
55-64	6.4	5.3	8.5	8.3	6.5	6.8	8.6	8.2	5.5	4.6	7.7	6.3
65+	5.7	5.6	8.2	6.3	6.0	5.2	8.8	5.4	5.4	4.1	6.5	6.3

Stratified Wilcoxon Test Results

The results from the demographic tests lend greater evidence towards the assertion that listening is unaffected by recruitment method. Of the 108 tests that were run, there were only 4 Wilcoxon tests with a p-value below 0.05, a rate of 4%, which yielded evidence of listening differences between recruitment methods. But this is about the number of tests that we would expect to randomly pass using the alpha threshold of 5%. We must point out, however, that these tests are not independent, as many of the same panelists will be in the panel during the full three month period.

Table 6: Stratified Wilcoxon Tests (Age x Race/Ethnicity x Gender x Month)

formatType	pvalues	Frequency	Percent
Total	<0.05	4	3.70
Total	>0.05	104	96.30

Based on the results of the Wilcoxon tests, we conclude that Arbitron's in-person recruitment of non-response households does not produce any observable differences in the total listening within the radio panel. This is an important question, as it was unclear whether people would be more or less prone to consume radio if a panel service representative were to visit the home and give personal instruction on the installation and maintenance of the PPM's.

Conclusion

One of Arbitron's primary business objectives is to obtain household participation in Arbitron's PPM panel to measure media consumption in major media markets. During the household selection process, Arbitron undergoes a variety of quality assurance tests to ensure the sampled populations reflect the population of radio listeners and non-listeners. One such test was to verify whether radio ratings were artificially attenuated by the company's new recruitment practice – that of visiting households and providing personal instruction in joining the radio panel (IPR.) Based on a stratified modality analysis using the Wilcoxon two-sample rank sum test, we found that the number of monthly panelist strata based on gender, age group, and race/ethnicity that exhibited listening differences between the IPR and Other groups was consistent with the number of positive tests we'd expect to see due to pure chance. Therefore, we conclude that the new recruitment practice implemented by Arbitron will have limited effect on the listening levels in the radio panels within the major media markets.

References

[1] Wilcoxon, Frank (1945). "Individual comparisons by ranking methods". *Biometrics Bulletin* **1** (6): 80–83.

[2] Kruskal, William H. (September 1957). "Historical Notes on the Wilcoxon Unpaired Two-Sample Test". *Journal of the American Statistical Association* **52** (279): 356–360.

[3] Mann, Henry B.; Whitney, Donald R. (1947). "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". *Annals of Mathematical Statistics* **18** (1): 50–60.

[4] Craig, Allen T.; Hogg, Robert V., *Introduction to Mathematical Statistics,* 5th ed., Prentice-Hall, Inc., New Jersey, 1995

[5] Lehmann, Erich L.; Romano, Joseph P., *Testing Statistical Hypothesis,* 3rd ed., Springer Science+Business Media, LLC., New York, 2005

[6] Bellera, Catherine A.; Marilyse, Julien., "Normal Approximations to the Distributions of the Wilcoxon Statistics: Accurate to What N? Graphical Insights". *Journal of Statistics Education*, **18** (2), (2010)

Thanks

William Waldron
Statistician
Arbitron
William.Waldron@arbitron.com

Kelly Dixon
Statistician
Arbitron
Kelly.Dixon@arbitron.com