Estimating attendance at the Powell Street Festival 2015

1. Introduction

The Powell Street Festival (PSF) is a community event in early August of each year. For the most part, it is a non-ticketed event, so it is difficult to obtain direct attendance estimates. In 2014, the PSF conducted a two-component survey to estimate attendance on 2/3 August 2014; the survey was repeated in 2015 on 1/2 August 2015.

An "attendance" is defined as a person visiting the festival on a day. If a person visits the festival on both days, the person would be counted twice. If a person visited the festival in the morning, left for lunch, and then returned, it is also counted as two visits.

2. Survey design and estimates.

The survey has two components. In the first component, the festival was divided into discrete areas and a head count was obtained on each day at 13:00, 15:00, and 17:00 (2014) and 13:30, 15:30, and 17:30 (2015) (Table 1). These head counts are assumed to be complete.

These counts, by them selves, do not provide sufficient information to estimate attendance. For example, if the three counts were 2500, 2500, 2500 in a day, it is impossible to know if the same 2500 people were present for all three head counts, or if the 2500 people present at the first count left and were replaced by a second set of 2500 people for the second count etc. Additional information is needed that is provided by the second survey component.

In the second component, surveyors wandered around the festival and approached approximately every 50th person they encountered. The exact value of 50 is not important, but this skip sampling ensures that more people are approached when the festival is busy and less when the festival is not rather than a fixed quota each hour. Each person that was approached was asked to identify their party size, when they arrived at the Festival and their planned departure time from the Festival. No person was interviewed more than once. A summary of the interviews conducted is found in Table 2. A plot of the arrival and departure times is shown in Figure 1 along with when the head counts were conducted. This plot shows that interviews took place through out the day as expected with no major gaps.

The estimated attendance for each day (Table 3) is then found as

$$\hat{N} = \frac{\text{Number of interviews}}{\text{Total active interviews}} \times \text{Total head counts}$$

For example, the estimated attendance on Saturday is found as

$$\hat{N}_{Saturday} = \frac{95}{(36+43+42)} \times (2442+3252+2448) = 6392$$

The number of people in each party appears in both the numerator and denominator in the derivation of this formula and so cancels in the final form above. An interview is defined as being "active" if the arrival and departure time spanned the time of a head count.

The intuitive idea behind the estimator is as follows. Suppose that the average party size was 3 (for convenience). The average head count (2442 + 3252 + 2448)/3 = 2714 divided by the average party size (3) implies that on average 905 parties were present at the head counts. The average of the active interviewees (36+43+42)/3= 40.3 implies the active interviewees comprised 40.3/905 = .045 = 4.5% of the total party counts. Because we interviewed over the entire day, the 95 interviews must represent 4.5% of the total parties for the day and so there must be 95/.045 = 2132 parties present. Finally, the average party size of 3 implies that 2132 x 3 = 6392 people attended the festival. You can see that the average party size divides one number and multiplies another number and so cancels in the overall operations.

A similar estimate can be obtained for Sunday, and the same formula can be used for both days by combining the data together. Because of the form of the formula, the estimated total for both days may not match exactly the sum of the individual days, but it will be close.

Of course this is just an estimate of the attendance and some measure of uncertainty must be provided. Following the methods of Dauk and Schwarz (2001), margins of error and 95% confidence intervals are presented in Table 3. The margin of error is the size of potential difference from the actual attendance 19 times out of 20. The 95% confidence interval (computed as the estimate ± margin of error) indicates a range of plausible actual attendances given the data that was collected.

It is not possible to obtain an estimate of the number of unique people that attended the events over the two days with this sampling method.

Key assumptions for this estimator to be valid.
There are two key assumptions for this estimator to be valid. First the head counts are accurate at each time. If these counts are biased upwards, then the estimate of attendance will also be biased upwards. Second, the interviews select parties with equal probability through out the day. That is why every n^{th} party was selected throughout the day rather than having fixed quotas every hour. The latter would tend to give a lower probability of selection to parties when the festival is busy.

3. Comparison of estimated attendance in the two years.
The instantaneous head count in 2015 were typically larger than the head counts conducted in 2014. The party size measured in the interviews was similar in both years, but the average length of stay appears to be shorter in 2015. Because the average length of stay is shorter, there would be larger turnover of parties during the event.

Consequently, it is not surprising that the estimated attendance appears to be larger in 2015 than in 2014 (Table 3 and Figure 2).

Appendix A – Technical details

This is a variant of an aerial-access creel survey design as outlined in Dauk and Schwarz (2001). In this case, the head counts are the equivalent of the aerial survey; the interviews the equivalent of the roving-surveys; and a count of 1 was used in place of the catch from the angling party. An aerial-access estimator was used because the variation in length of stay among interviewed parties indicated that the effects of length-biased sampling could be ignored.

Dauk, P.C. and C.J. Schwarz. 2001. Catch estimation with restricted randomization in the effort survey. Biometrics 57: 461-468.

Table 1. Head counts on each day.		
2014		
	Saturday 2014-08-02	Sunday 2014-08-03
13:00	2442	3123
15:00	3252	3119
17:00	2448	1819
2015		
	Saturday 2015-08-01	Sunday 2015-08-02
13.30	3314	3445
15:30	3688	3508
17:30	2361	2300

Table 2. Summary statistics about interviews on each day.		
2014		
	Saturday 2014-08-02	Sunday 2014-08-03
Number of interviews	95	80
Average party size[1,2]	3.3	3.2
Average planned length of stay (hours)[2]	3.0	3.3
Parties active during 13:00 headcount	36	39
Parties active during 15:00 headcount	43	51
Parties active during 17:00 headcount	42	39
2015		
	Saturday 2015-08-01	Sunday 2015-08-02
Number of interviews	84	100
Average party size	3.0	3.1
Average planned length of stay (hours)	2.6	2.6
Parties active during 13:30 headcount	39	43
Parties active during 15:30 headcount	34	39
Parties active during 17:30 headcount	34	30

[1] A party was deemed to be active during a headcount if the time of the headcount was between the reported time of arrival and estimated time of departure for that party.

[2] Not adjusted for potential length of stay bias.

Table 3. Estimated attendance at the PSF			
2014			
	Saturday 2014-08-02	Sunday 2014-08-03	Both days
Estimated attendance	6392	4999	11,342
Margin of error	1000	588	1,116
95% confidence interval	5411 -> 7375	4422 -> 5576	10,253 -> 12,530
2015			
	Saturday 2014-08-01	Sunday 2014-08-02	Both days
Estimated attendance	7584	7350	14,935
Margin of error	1050	1105	1,530
95% confidence interval	6551 -> 8617	6260 -> 8439	13,433 -> 16,436

The margin of error is a measure of uncertainty in the estimate. The actual attendance will be within the margin of error from the estimated attendance 19 times out of 20. The 95% confidence interval reports the range of plausible attendances.

The attendance for both days is different than the sum of the individual attendances. This is an artifact of the statistical method.
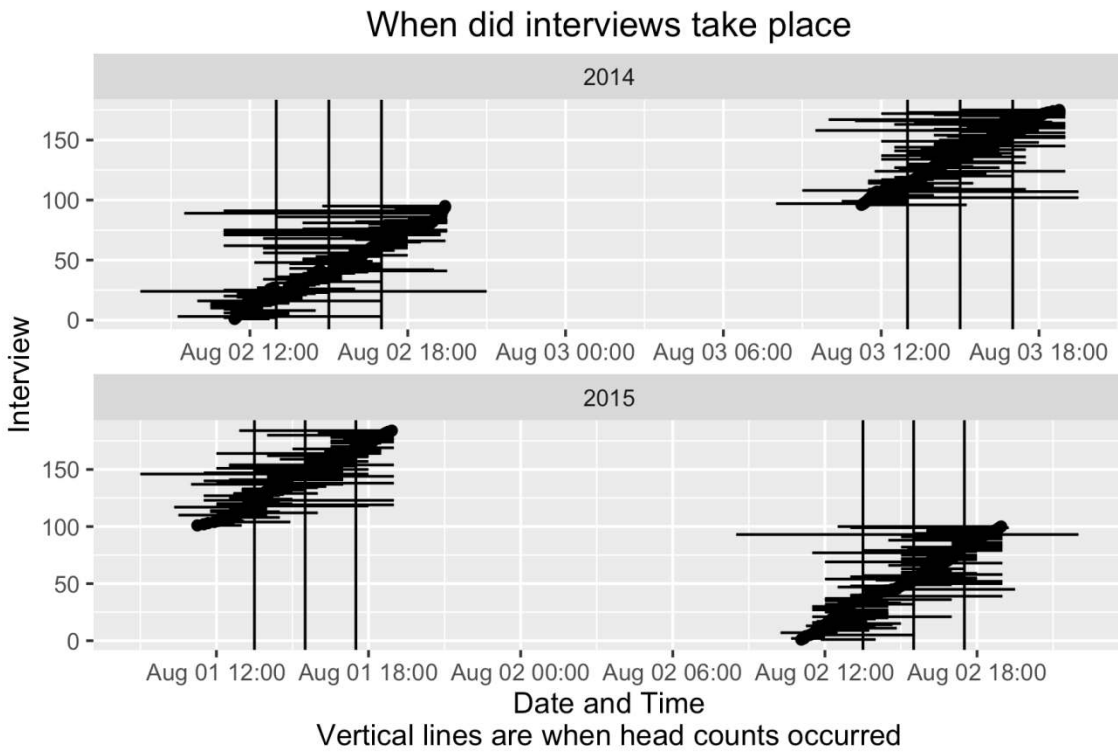
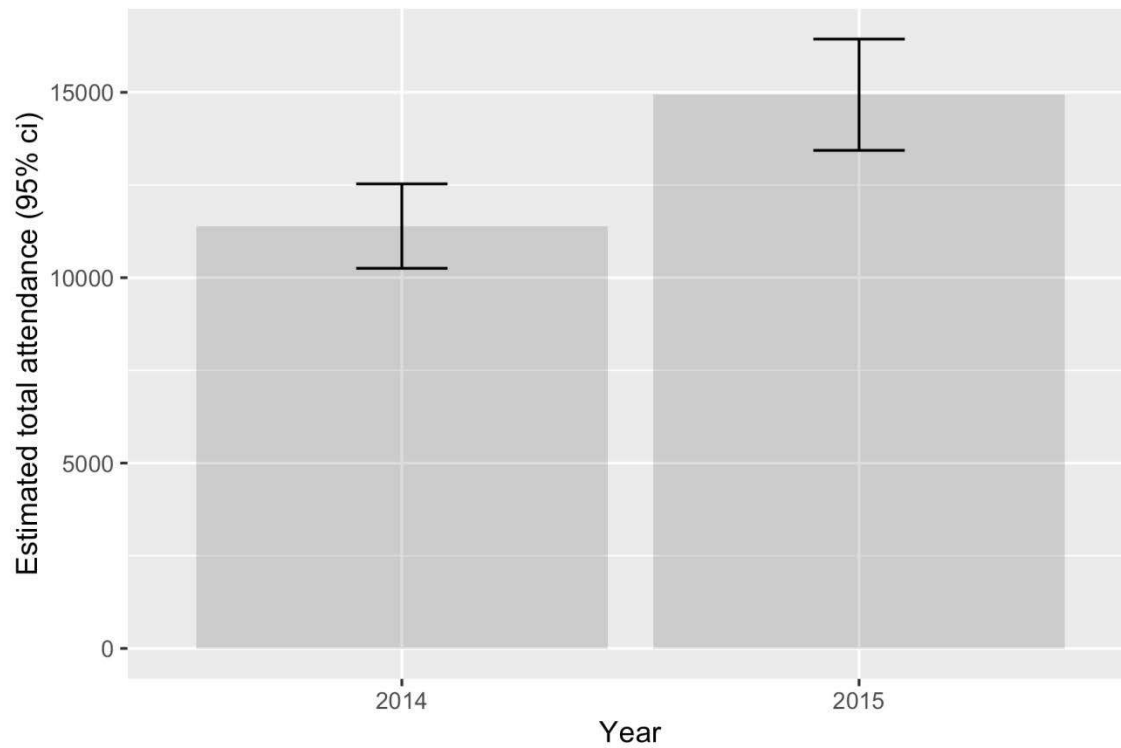Figure 1. Plot of arrival and planned departure for each interview (solid horizontal line) and times of headcounts (vertical lines).

Figure 2. Comparison of estimated attendance across years.