Short Report 1

Parker Johnson and Carl Zhang

1/20/2022

Introuction

In 2017, there was controversy over the estimated number of people that attended Trump's inauguration speech, as Trump and his associates overestimated the number of attendees. Trump initially reported that over 1 million attendees were at his inauguration, while estimates from crowd size professor Keith Still estimated a total attendance of 300,000 to 600,000 people. In this report, we explored how to estimate the size of a pseudo crowd using samples, with the population of our crowd being represented by a visual of dots.

Methodology

We applied a stratified sampling technique. Firstly we performed a boxplot exploratory data analysis using an SRS taken from each stratum to assess the reasonableness of a stratified analysis and if normal assumptions were met. We assumed that within each of the designated stratum the crowd density is roughly homogeneous, which allows us to stratify the population into the two distinct sections. Along with this, sampling elements between stratum are heterogeneous, which justifies our decision to stratify in the first place. Thus, the crowd was split into two strata (see figure 1 for strata cutoff specifications).

We designated areas that were 80 units wide by 150 units long as our sampling units for both strata (as shown by the blue boxes in figure 1). To sample a box, we manually counted the dots inside and on the border of the box. The size of the sampling units divides the strata evenly, which makes taking a simple random sample of the boxes in each stratum easier.

Next, we needed to estimate an appropriate sample size. We decided on a reasonable percent margin of error (0.05). To calculate the sample sizes for an SRS without replacement, we used the following equation:

$$n_0 = (\frac{sz}{e})^2$$

In this equation, s is our estimate for the standard error for each stratum, e is the margin of error, z is the corresponding z-score, and N is the number of items in each sampling frame for each stratum – the same in our case. Here, s and e are unknown.

To estimate the standard deviation of each stratum, we roughly chose two units that roughly contained the maximum and minimum amount of dots for each stratum and used the equation $\frac{y_{max}-y_{min}}{4}$. We took a pilot sample of four units from each stratum to estimate the margin of error for each stratum. To do this, we applied this equation:

$$e = percentError * \bar{y}$$

From this,we used this equation and the equation for n_0 to calculate the sample population needed for a simple random sample without replacement. Since we are sampling from a small sampling frame, we used the finite population correction factor to calculate the desired sample size using this equation.

$$n = \frac{n_0}{1 + n_0/N}$$

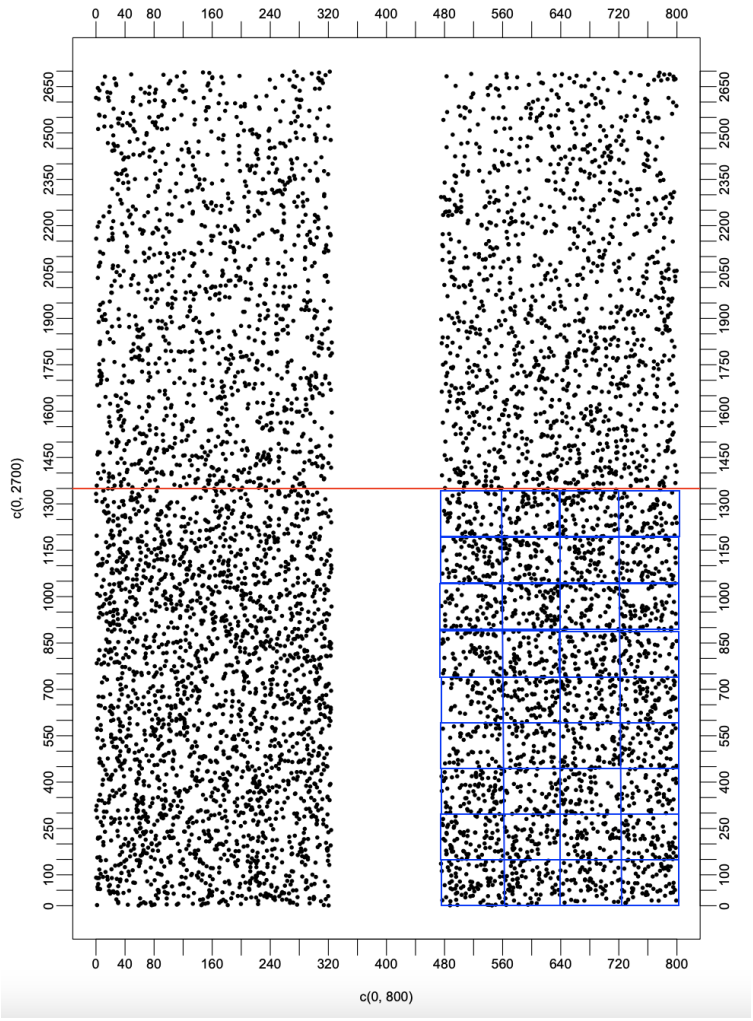


Figure 1: The provided pseudo crowd. The red line roughly represents the cutoff at y = 1350 for the two strata (the population above and below the line). The blue boxes roughly represent the sampling units (80 units wide by 150 units long).

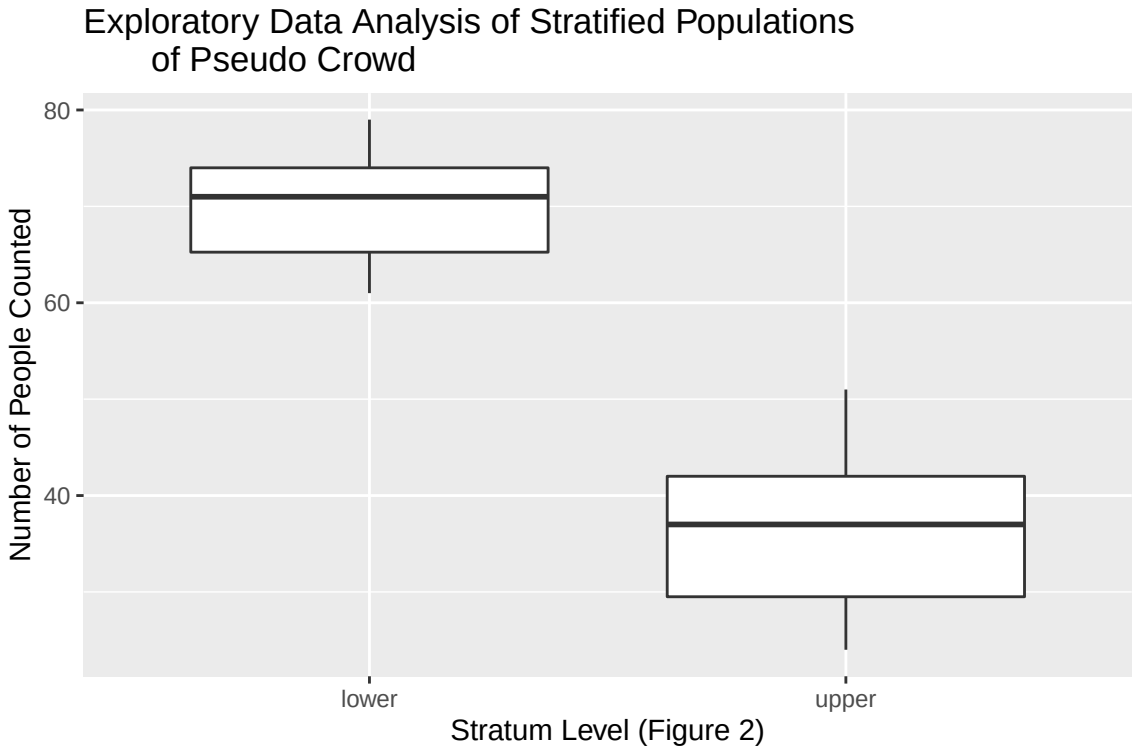
Once we calculated our desired sample size, we were able to sample each strata with their calculated corresponding sample sizes and calculate the sample mean and standard error for the number of dots in each sampling unit for the samples. To estimate the total population we combined the sample total estimates (sample mean estimates multiplied by number of units in the sampling frame for each stratum) and calculated the combined standard error for the two strata $S_{combined} = \sqrt{S_1^2 + S_2^2}$. From this we calculated our total population estimate and a 95% confidence interval.

Results

We first checked to see if a stratified sample would be appropriate to use for our data. After taking 30 samples from both the upper and lower strata, we plotted the mean and distribution of each stratum in a boxplot (see figure 2). Since the population within each stratum are noticeably different, and the distribution within each stratum is approximately normal, we can justify using a stratified sample for our data.

We first investigated what would be appropriate sample sizes to use for our estimation of the total population of the pseudo crowd. For the upper stratum, we calculated that n = 30.86, which indicates that we need to take 31 samples from this stratum to achieve our margin of error, and for the lower stratum, we got that n = 6.03, which indicates that we need to take 7 samples from this stratum.

Our next task was calculating the standard error of the population, and calculating the estimated total population of the crowd. After calculating the mean and standard errors from the samples within each stratum, we estimated that the total population of our pseudo crowd was 7871 dots, with a standard error of 200.3 dots. We are 95% that the actual population of the pseudo crowd is between 7477 and 8263 dots.



Discussion

As mentioned previously in the results section, we estimated that the total population of the pseudo crowd was 7871 dots, with a standard error of 200.3 dots. Relating this to the example of Trump's inauguration speech, a similar analysis could possibly be conducted using the aerial image of the event to estimate the total number of attendees within the crowd. The fairly large standard error indicates that precise estimates

of crowd sizes are difficult to achieve given the large number of samples that would need to be collected within each stratum.

One of the errors that could have taken place when conducting our design was overcoverage. When taking samples, dots that stood on the border of a sampling unit were counted. However, if we were to also randomly sample an adjacent sampling unit, we could have double counted dots that bordered both sampling units. Undercoverage was also possible given that some dots could have been directly on top of each other in the figure and were missed. This would be similar to a child not being counted for a real crowd if the child was hidden beneath their parents.

Within each sample, there was a large enough number of dots where the assumption of a large enough sample size for a confidence interval was met. However, since the minimum number of dots for the upper stratum was found to be 24, which is less than the recommended sample size of 30 for a confidence interval, it's possible that non-normality was introduced into our design through too small of sample sizes for the upper stratum. However, since the vast majority of our sample sizes were larger than 30, the large sample size assumptions were met for the most part within our study.

The data collection when implementing our design was the biggest difficulty and source of error. It is quite possible that we miscounted the number of dots within each sample. This could prove to be especially difficult when estimating the population of an actual crowd, given the large possibility of error in miscounting individuals.

Appendix

The following is the R code used to reach our conclusions.

First, we calculated the standard error for each strata. This was done by estimating the boxes within each strata that contained the maximum and minimum number of people, and calculating this range and dividing by 4 to get a conservative estimate of the error. We then calculated the margin of error for both strata using a pilot sample, choosing a value of .05 for how close proportionally we wanted our sample to be from the true population. From these margin of errors and standard errors, we were able to compute n_0 for each stratum, which is the sample sizes we would need to acheive our margin of error without the finite population correction.

```
#Picked box dimension of 80 by 150
upper_stata_max <- 51
upper_strata_min <- 24

lower_strata_max <- 79
lower_stata_min <- 61

s_upper_strata <- (51-24)/4  #6.75 estimated range of values in
                             #sampling units for upper strata/4
s_lower_strata <- (79-61)/4  #4.5  estimated range of values in
                             #sampling units for lower strata/4

#Calculate margin of error for each stratum
# We want sample to be .05 away from actual population. This is using a pilot sample
e_estimateUP <- (41+31+37+35)/4 * .05
e_estimateLOW <- (64+71+67+73)/4 * .05

n_0_upper <- ((s_upper_strata * 1.96)/e_estimateUP)^2
n_0_lower <- ((s_lower_strata * 1.96)/e_estimateLOW)^2
```

We then calculated the necessary sample sizes, with the finite population correction. The size of N in this case refers to the total number of sample sizes we can take within each stratum.

```
big_N <- (640/80)*(2700/150) / 2 #Per strata
n_upper <- (n_0_upper / (1 + n_0_upper/big_N))
n_upper
```

```
## [1] 30.86449
```

```
n_lower <- (n_0_lower / (1 + n_0_lower/big_N))
n_lower
```

```
## [1] 6.031889
```

Next, we took the number of necessary samples from each stratum, which was 19 samples from the upper stratum, and 6 samples from the lower stratum, and calculated their means and standard deviations. From this, we were able to calculate the total standard errors for the population of each stratum, as well as the total standard error for the entire population.

```
mean_upper <- (48+28+34+46+27+32+46+38+48+41+49+50+37+30+50+33+41+28+30+33+25+48+
                 42+34+48+45+50+43+39+36+44)/31
mean_upper
```

```
## [1] 39.45161
```

```
s_upper <- sd(c(48,28,34,46,27,32,46,38,48,41,49,50,37,30,50,33,41,28,30,33,25,48,
                42,34,48,45,50,43,39,36,44))
mean_lower <- (73+78+64+64+69+78+63)/7
mean_lower
```

```
## [1] 69.85714
```

```
s_lower <- sd(c(73,78,64,64,69,78,63))


SE_upper <- sqrt(big_N^2 * (1 - n_upper/big_N)*(s_upper^2 / n_upper))
SE_upper
```

```
## [1] 77.99252
```

```
SE_lower <- sqrt(big_N^2 * (1 - n_lower/big_N)*(s_lower^2 / n_lower))
SE_lower
```

```
## [1] 184.3152
```

```
SE_total <- sqrt(SE_upper^2 + SE_lower^2)
SE_total
```

```
## [1] 200.1372
```

Next, we were able to calculate our estimate for the total crowd size, by adding together the expected number of people within each stratum.

```
total_est <- mean_upper * big_N + mean_lower * big_N
total_est
```

```
## [1] 7870.23
```

Finally, we calculated the 95% confidence interval for our total crowd size by using our total crowd size estimation and the total standard error found previously.

```
CI_upper <- total_est + 1.96*SE_total
CI_upper
```

```
## [1] 8262.499
```

```
CI_lower <- total_est - 1.96*SE_total
CI_lower
```

```
## [1] 7477.961
```