BIOS 664 Project 'Estimation of number of rows at campus auditoriums' results

Cally Pfeiffer, Nathaniel Putnam, Preston Burns, Vasyl Zhabotynsky, Patrick Pasquariello

As detailed in our first paper, we split the auditoriums into three stratum of roughly equal predicted variance, sorted by seating capacity. We used Neyman allocation to determine that we needed to sample 3 auditoriums from the stratum with the smallest auditoriums, 2 auditoriums from the stratum with the largest auditoriums, and 2 auditoriums from the stratum in the middle. This meant that our sample was not EPSEM, but calculating the necessary sample weights was not complicated, since within-stratum sampling was done by SRS. Selection probabilities for respective strata are 3/18,2/12,2/6.

Our sampled units were very easy to measure; We simply walked into the room when it was unoccupied and counted the number of rows. In the process of measuring, we had to clarify two points: first, separation by aisles did not affect the number of rows; a row split in three by two aisles still only counted as one row. Second, if a partial row (say, a small cluster of seats on either side of a media closet in the far back of the room) was labelled as a separate row, then it was counted as a separate row, regardless of size.

Given the number of different institutions on campus that we contacted and the lengths we had to go to procure a response, it is a distinct possibility that our list of auditoriums is incomplete. However, we believe that the unknown auditoriums would be predominantly small auditoriums, since people are less likely to omit the large auditoriums that are traditionally named as such. Thus if our estimate is biased, it is likely to have upward bias. In order to evaluate the potential effect of this bias we run a sensitivity analysis assuming that the true size of the stratum with smaller auditoriums had not 18, but 20 auditoriums (with missing two being otherwise typical size for this strata)

Once the data was collected, the population mean row count was estimated using our two datasets: the average size of three theatrical auditoriums with online floor plans, $\bar{y}_3$, and the average sample estimate of the other 36 auditoriums, $\bar{y}_{36}$. As we demonstrated in the Sampling Consideration section of the previous paper the resulting estimate $\bar{y}_{39} = \frac{1}{13}\bar{y}_3 + \frac{12}{13}\widehat{\bar{y}}_{36}$ is unbiased and has variance $\widehat{V}(\widehat{\bar{y}}_{39}) = \frac{144}{169}\ \widehat{V}(\widehat{\bar{y}}_{36})$.

We also regressed the number of rows onto the square root of total auditorium seating capacity. In testing stage we consider two different linear models with, or without intercept:

$$y = a + b\,x + e, \quad (1)\ where\ e - iid\ from\ N(0, \sigma^2) \quad \text{and}$$
$$y = b\,x + e \qquad (2),\ where\ e - iid\ from\ N(0, \sigma^2)$$

Both of these models have a standard solution provided by R/lm or SAS/reg function.

The uses of this model were twofold. First, it allowed us to check our assumption that the square root of seating capacity was directly proportional to the row count. Initially, this assumption was based on the idea that most auditorium seating areas are square, but we discovered that a few of the auditoriums were of the very different 'thrust' design.

Figure 1. Paul Green Theatre seating chart, a “Thrust” design auditorium

Within our sample we observed that unless the auditorium is a ‘thrust’ design (such as a Paul Green Theatre - see Figure 1), there is a very precise linear correlation between the root of seating capacity and the number of rows, just as we had hoped.

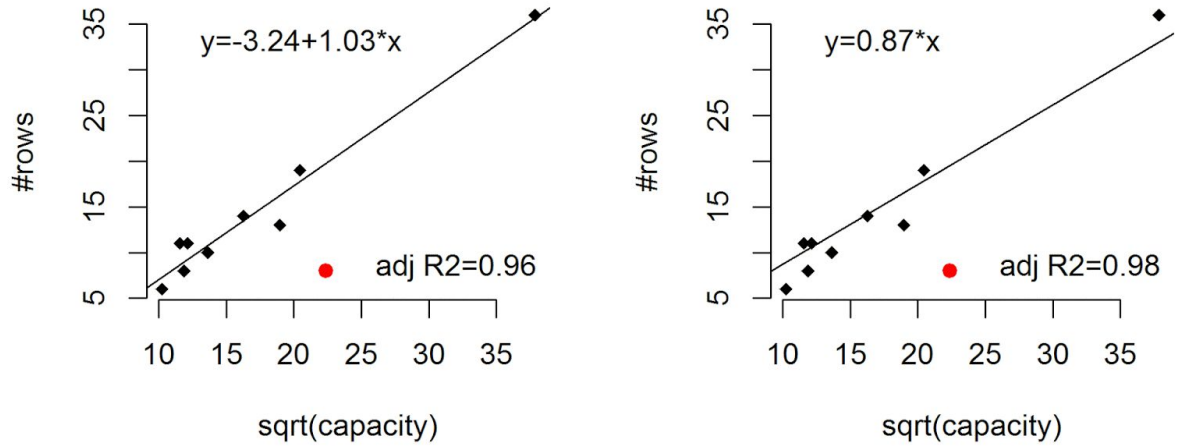

Figure 2. Linear fit with variable intercept on the left, and with 0 intercept on the right. The Red circle is Paul Green Theatre

The second use of this model was to give us an alternative estimate of the mean row count: the model’s projected row count for an auditorium with the average number of seats across all auditoriums. This mean is listed in our table of results in the row labelled lm(y~a+b x).

With our assumptions validated, the only consideration left was the worry that our estimate would be biased because of an incomplete sample frame. As mentioned above, it is possible that there are several small auditoriums on campus that we were not made aware of. To test the

ramifications of this possibility, we performed a sensitivity analysis to see what difference in mean we would find if there were two more auditoriums in the stratum with the smallest auditoriums. You can see the results in the table below, in the row labeled sensitivity analysis. All three estimates of mean row count yield very similar results, suggesting that our sampling method was quite robust. We estimate that the mean number of rows in all auditoriums on campus is between 9 and 13 (~11±2).

Table 1: Estimates from three models and sensitivity analysis.

model	Mean estimate	S.E. estimate	95%CI
lm(y~a+b x)	10.96	0.67	(9.38,12.53)
Stratified sampling	11.03	0.73	(9.23,12.82)
Sensitivity analysis	10.89	0.76	(9.04,12.75)

Appendices:

Appendix Section 1: Formulas used in the analysis

The exact formulas used to do stratified sampling can be found in the lecture slides 3 (pages 13 and 17) and are replicated below:

$$\widehat{\bar{y}} = \sum_{h=1}^{4} W_h \bar{y}_h, \ where\ \bar{y}_h = \sum_{i \in S_h} y_i / n_h \ and\ W_h = \frac{N_h}{N},$$

$$\widehat{V}(\widehat{\bar{y}}) = \sum_{h=1}^{4} W_h^2 [\frac{1-f_h}{n_h} s_h^2], \ where\ f_h = \frac{n_h}{N_h},\ s_h^2 = \sum_{i \in S_h} [y_{hi} - \bar{y}_h]^2 / (n_h - 1)\ .$$

Note that the fourth stratum here is a pseudo-strata of three buildings with online floor plans, all of which were sampled. When we apply finite size correction, our SAS and SUDAAN code adjusts the variance of this stratum to zero and calculates a properly weighted estimate and it's error (also, this way we don't need to recalculate confidence interval since the program will use a t distribution with proper degrees of freedom)

For linear regression the solution of $\beta's$ is

$$\widehat{\beta} = (X'X)^{-1} X'y \ with\ \widehat{\sigma}^2 = (Y - X'\widehat{\beta})'(Y - X'\widehat{\beta})/(n-p)\ where\ p = 2\ for\ model\ 1\ and\ p = 1\ for\ model\ 2$$

From these estimates we can easily compute the predicted y and its standard error at the mean value of x denoted X_{new}. The fitted value and its standard error were calculated using the following formulas:

$$\widehat{y}_{new} = X_{new}'\widehat{\beta}\ \ and\ V(\widehat{y}_{new}) = X_{new}'(X'X)^{-1} X_{new} \widehat{\sigma}^2$$

(computations were done using R/lm function and applying R/predict function on the result)

Appendix section 2:

SAS/SUDAAN Code used for Analysis:

```
proc import out=work.allaud
datafile = "C:\courses\2016S\BIOS664\auditorium\auditorium.csv"
      DBMS = csv replace;
      getnames = yes;
      datarow=2;
run;
proc sort data=allaud;
    by strata;
    run;
/*Create data set 'strat_popsize' with stratum population counts*/
proc freq data=allaud noprint;
    tables stratum/nocum nopercent out=strat_popsize;
run;
data strat_popsize;
      set strat_popsize;
       _total_=COUNT;
      drop COUNT PERCENT;
 RUN;
proc import out=work.aud3
      datafile = "C:\courses\2016S\BIOS664\auditorium\sampled.csv"
      DBMS = csv replace;
      getnames = yes;
      datarow=2;
run;
proc surveymeans data=work.aud3 plots=none N=strat_popsize total=strat_popsize
plots=none;;
      strata stratum;
      var numrow;
      weight weight; /*sampling weight variable*/
run;
proc descript data=work.aud3 notsorted design=wor;
      nest stratum; /*indicates that we stratify*/
      totcnt fpc;
```

```
      weight weight;
      var numrow;
print nsum="Sample Size" total wsum="Est Pop Size" mean semean lowmean upmean;
      run;
proc import out=work.audu
      datafile = "C:\courses\2016S\BIOS664\auditorium\under_sampled.csv"
      DBMS = csv replace;
      getnames = yes;
      datarow=2;
Run;
proc import out=work.strat_popsize
      datafile = "C:\courses\2016S\BIOS664\auditorium\strat_popsize_sens.csv"
      DBMS = csv replace;
      getnames = yes;
      datarow=2;
      run;
proc surveymeans data=work.audu plots=none N=strat_popsize total=strat_popsize
plots=none;;
      strata stratum;
      var numrow; /* weight*/
      weight weight; /*sampling weight variable*/
run;
proc descript data=work.audu notsorted design=wor;
      nest stratum; /*indicates that we stratify*/
      totcnt fpc;
      weight weight;
      var numrow;
print nsum="Sample Size" total wsum="Est Pop Size" mean semean lowmean upmean;
run;
```

Appendix Section 3: R code for linear regression and plots:

```
nh = c(3,2,2); n = sum(nh); fh = nh/Nh; Nh = c(18,12,6); N = sum(Nh); Wh = Nh/N
all.aud = read.csv("C:/courses/2016S/BIOS664/auditorium/auditorium.csv",as.is=T)
meanaud = mean(sqrt(all.aud$Capacity))
meanaud = mean(sqrt(all.aud$Capacity[-38]))
rows = c(6,      8,      11,     10,     11,     13,     19,     36,     14)
```

```
capr =
c(10.24695,11.87434,11.57584,13.63818,12.16553,18.97367,20.46949,37.86819,16.27882)
lm1 = lm(rows~capr)
lm2 = lm(rows~-1+capr)

png("C:/courses/2016S/BIOS664/auditorium/auditorium.png",height=3.5,width=7,units="in",res=
300)
par(mfrow=c(1,2))
plot(capr,rows,pch=18,bty="n",xlab="sqrt(capacity)",ylab="#rows",main="")
abline(a=lm1$coef[1],b=lm1$coef[2])
points(22.36068,8,pch=19,col="red")
legend("topleft",legend="y=-3.24+1.03*x",bty="n")
leg = sprintf("adj R2=%s",round(summary(lm1)$adj.r.squared,2))
legend("bottomright",legend=leg,bty="n")

plot(capr,rows,pch=18,bty="n",xlab="sqrt(capacity)",ylab="#rows",main="")
abline(a=0,b=lm2$coef[1])
points(22.36068,8,pch=19,col="red")
legend("topleft",legend="y=0.87*x",bty="n")
leg = sprintf("adj R2=%s",round(summary(lm2)$adj.r.squared,2))
legend("bottomright",legend=leg,bty="n")
dev.off()

pr.lm1 = predict(lm1, newdata = data.frame(capr=mean(meanaud)), se.fit = TRUE)
pr.lm2 = predict(lm2, newdata = data.frame(capr=mean(meanaud)), se.fit = TRUE)
```

Appendix Section 4:

Team 4 Project Plan Score Sheet

Total Points: 39

1. Comments on "Statement of the Problem" section
   - None
2. Comments on "Proposed Sample Design" section
   - Need to describe how exactly selection probabilities should be computed for each sample member using your sampling design (-1 pt.)
3. Comments on "Data Collection and Estimation" section
   - None