Attendance Rate Survey for the UCD School of Mathematical Sciences

Conducted 22nd - 27th Oct, 2007
by the students of module
STAT30020/40220 Survey Sampling

Supervisor:
Dr. Gabrielle Kelly

Report Prepared By:
Laura Cribbin
Barry Croke
Andrew Ryan
Laura White

Conducted By:
Karl Conlan
Laura Cribbin
Barry Croke
Katherine Howard
Gerard Kilroy
Manu Kumar
Samuel Lekwadi
Jade O Connor Mc Goona
Justine O Neill
Andrew Ryan
Conall Scollard
Yvonne Sylvestre-Garcia
Simon Twist
Owen Wardell
Laura White

Back Row (left to right): Dr. Gabrielle Kelly, Andrew Ryan, Manu Kumar, Conall Scollard, Gerard Kilroy, Laura Cribbin, Katherine Howard, Yvonne Sylvestre Garcia, Laura White, Justine O' Neill, Simon Twist, Prof Sean Dineen (Head of School)

Front Row (left to right): Dr. Marcus Greferath (Teaching & Learning), Owen Wardell, Barry Croke, Karl Conlan, Samuel Lekwadi

Photographer: Fionnuala Cuffe, Systems Demonstrator

Summary

For the first time the Survey Sampling class STAT30020/40220 conducted a survey to estimate attendance rates at classes in the UCD School of Mathematical Sciences. Based on a sample of 52 classes the overall attendance rate was 58% (+/- 4%). The estimates for different subgroups varied between 28% (level 0-1) and 78% (level 3-4). The estimate obtained were adjusted for the level of the class (levels 0-4), the time of the day and for attendance being more variable in larger classes. As the survey was successfully carried out in a tight time frame it shows that it is feasible to repeat it in the future and possibly include other Schools. The survey results are of value to lecturers and Head of School. Moreover, it gave us the Sampling Survey class practical hands-on experience of conducting a useful survey, was very interesting and helped greatly in understanding the lecture material.

Section 1 - Introduction

1.1 Background

We began this survey with two real objectives in mind. As a class studying Survey Sampling it was a good way to provide a platform for students to understand the real life problems with conducting surveys. It was also intended to provide an insight into the class attendance rates. Furthermore, we also hoped our survey will serve as a template for future Survey Sampling classes who can use our results and expand on our findings. Our survey is interested in the attendance rates of the students in the School of Mathematical Sciences in University College Dublin.

1.2 Data Assembly

How The Sample Frame Was Obtained?

We discussed the sample frame in class and we felt there were a number of important factors which would help us select our sample.

What Classes to Include:

Trying to include all classes in the college was obviously out of the question because our class only has 15 students. We also didn't want to make our sample frame too small by only including STATs class say. So the first decision we made was to include all classes in the School of Mathematical Sciences.

Labs & Tutorials:

We quickly realised that we would have to decide if we were going to include labs and tutorials in our survey. We thought that because these classes are often compulsory and/or marks may be given for them that it wasn't a good idea to include them.

Two Hour Lectures:

We decided to include two hour lectures even thought we felt that their attendance might be slightly above average, because it could be the only lecture of that subject all week and therefore a student would be more reluctant to miss it. This decision had a hidden implication in terms of selection of the sample which we will look at later.

Strata:

Splitting the classes up by levels seemed to make sense. This way you can see if there is a big difference between attendance from Level 1 courses and Level 4 courses and see if attendance rates increased as the level did. We discussed this in class and felt that there should be three distinct course levels in our survey; Level 0+1, Level 2 and Level 3+4. We felt that Levels 3 and Levels 4 should be put into the same group because there is a lot of overlap between courses that have two codes (e.g. our own class Survey Sampling is called STAT30020 and STAT40220) and that the attendance in Level 3 and Level 4 courses should be similar.

The next factor which we felt was an important factor in class attendance was the time of the lecture. We discussed it in class and felt that the main reason people wouldn't attend lectures is because it is on early in the morning. We decided to split lectures into two time intervals, early lectures 9am till 11am and late lectures; 11am till the end of the day.

Now we had 6 strata.

Stratum	Level	Time
1	Levels 0,1	9am-11am
2	Level 2	9am-11am
3	Levels 3,4	9am-11am
4	Levels 0,1	11am-8pm
5	Level 2	11am-8pm
6	Levels 3,4	11am-8pm

The decision to split the classes by time led to a problem with including 2 hour lectures. If a two hour lecture began at 10am and ended at 12pm which strata would it be included in? We looked at all possible two hour lectures to see if any would fall into both time frames and thankfully none of them did.

Organization of the Survey:
The organization of the Survey was done by the MSc students in the class with the help of Dr. Gabrielle Kelly. All students took part in the collection of data and in the discussions about our survey. Students were assigned 3/4 classes of which they would have to count the attendance. These classes were assigned to students based on their own timetables.

To inform the lecturers that we were going to be taking a survey of the class we prepared a letter to give to each lecturer that was selected in our survey. Here is our proposed letter.

The Letter:

Contact Details:
Dr. Gabrielle Kelly
Room L529
Library Building

Dear Doctor/Professor <Insert name here>

My name is <Insert name here> and I am a member of the Survey Sampling statistics class taught by Dr. Gabrielle Kelly. As part of our mandatory coursework we are conducting a survey of student attendance of lectures run by the School of Mathematical Sciences.

This survey will be conducted during the week commencing Monday 22nd of October. Your lecture <Insert lecture here> taught at <Insert time here> in <Insert room here> has been randomly selected to be surveyed as part of our sample. This would involve one member of our class (two if the class is large) entering the lecture ten minutes before it ends and silently counting the number of students in attendance.

Not only is this survey part of our course, but it will be presented to the Head of the School of Mathematical Sciences and can provide valuable insight into attendance habits. As such we would appreciate your permission and co-operation with this survey.

Kind Regards,

_____
<Insert name here>

In the end we couldn't send out individual letters for reasons which will be explained later.

Section 2 - Methods

2.1 Description of the Survey Design

Our survey used a design of stratified cluster sampling. Stratification is achieved by separating the population into non-overlapping groups, called strata. We then performed cluster sampling on each stratum. In our survey, each class was a cluster and each element was a student registration. Cluster sampling is a probability sample in which each sampling unit is a collection (or cluster) of elements.

2.2 Selection of the Sample Size

Using the timetable for the School of Mathematical Sciences, it was concluded that there were 203 lectures every week. A sample size had to be selected from these classes.

To be 95% sure that an estimate is in error by at most *B*, the sample size *n* is given by the formula

$$n = \frac{N\sigma^2}{(N-1)\frac{B^2}{4} + \sigma^2}$$

where N = population size and σ^2 = variation in population.

For our experiment σ^2 had to be approximated since there was no prior knowledge of the variation. Assuming approximate normality, $\sigma \approx \frac{Range}{4}$ and since the range was [0,1], σ =1/4.

Three different sample sizes were computed for different levels of *B*, using the above formula with N=203 and σ^2 = 1/16.

B=0.05 (5% bound on the error of estimation)	n=68
B=0.06 (6% bound on the error of estimation)	n=52
B=0.07 (7% bound on the error of estimation)	n=48

It was decided to use a sample size of 52 classes with a 6% bound on the error of estimation. This was selected since we wanted our error of estimation to be as low as possible, but we also had to consider our own limitations. As a small class of 15 people, it was impractical for us to consider sampling more than 3 or 4 classes each.

2.3 Allocation to the Strata

The sample size of 52 classes then had to be divided into 6 strata. It was decided to use the method of proportional allocation which assumes equal costs and variances for each stratum.

Proportional Allocation: $$n_i = n\left(\frac{N_i}{N}\right)$$

Where:
n = 52
N = 203
N_1 = 21 $\Rightarrow$ n_1 = 5
N_2 = 13 $\Rightarrow$ n_2 = 3
N_3 = 32 $\Rightarrow$ n_3 = 8
N_4 = 25 $\Rightarrow$ n_4 = 7
N_5 = 27 $\Rightarrow$ n_5 = 7
N_6 = 85 $\Rightarrow$ n_6 = 22

2.4 Estimation Methods Used

Within the cluster sampling it was decided to test three methods for three different situations of the variation. In this section y_j refers to the attendance of class j and x_j refers to its enrolment.

2.4.1 Method A:

The estimator for the mean proportion of attendance for stratum *i* is given by

$$\bar{y}_i = \frac{\sum_j^{n_i} y_j}{\sum_j^{n_i} x_j}$$

assuming that $V(y_j|x_j) = \sigma^2 x_j$ (i.e. the variance is proportional to the x_j)

2.4.2 Method B:

The estimator for the mean proportion of attendance for stratum *i* is given by

$$\bar{y}_i = \frac{\sum_{j}^{n_i} x_j y_j}{\sum_{j}^{n_i} x_j{}^2}$$

assuming that $V(y_j | x_j) = \sigma^2$ (i.e. the variance is constant.)

2.4.3 Method C:

The estimator for the mean proportion of attendance for stratum *i* is given by

$$\bar{y}_i = \frac{1}{n_i} \sum_{j}^{n_i} \frac{y_j}{x_j}$$

assuming that $V(y_j | x_j) = \sigma^2 x_j^2$ (i.e. the variance is proportional to the x_j^2)

For all three methods the variance of the estimate $\bar{y}_i$ can be estimated by

$$\hat{V}(\bar{y}_i) = \frac{1}{\sum_{j=1}^{n_i} w_j x_j{}^2} \left(\frac{\sum_{j=1}^{n_i} w_j (y_j - \bar{y}_i x_j)^2}{n_i - 1} \right)$$

Where the weights (w_j) are given by:

Method A: $w_j = \frac{1}{x_j}$

Method B: $w_j = 1$

Method C: $w_j = \frac{1}{x_j{}^2}$

2.5 Combining the Strata Results

In order to compute an average rate for the entire School of Mathematical Sciences we needed to combine the results of the six strata. This was achieved using the following standard formulae:

2.5.1 Estimator of the population mean μ :

$$\bar{y}_{st} = \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + ... + N_L\bar{y}_L)$$

$$= \frac{1}{N}\sum_{i=1}^{L} N_i\bar{y}_i$$

2.5.2 Estimated Variance of $\bar{y}_{st}$:

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\hat{V}(\bar{y}_i)$$

Where

- L = Number of Strata
- N_i = Size of Stratum i
- n_i = Size of Sample selected from Stratum i
- N = Total Size of Population = $N_1 + N_2 + \ldots + N_L$

Section 3 - Results

3.1 Data

Table 3.1 illustrates the attendances and enrolments obtained for all six strata. Note that a full stop indicates a missing attendance observation.

Table 3.1

After the attendances were collected and the data was compiled, attendance rates were calculated using three different methods.

3.2 Calculations

The attendance rate and variance was then calculated for each stratum using the following three different ratio estimators A, B and C from Methods, Section 2.

The results are as follows:

Table 3.2.1: Estimates for attendance rates, variance and standard error by stratum using estimator A

Stratum	Level	Time	Estimate	Variance	Standard Error
1	0,1	9,10	42%	0.052%	2.28%
2	2	9,10	51%	0.460%	6.78%
3	3,4	9,10	52%	0.270%	5.19%
4	0,1	All other times	28%	0.053%	2.30%
5	2	All other times	49%	0.437%	6.61%
6	3,4	All other times	78%	0.115%	3.39%

Under ratio estimator A we obtained a total population mean attendance rate of 58% with a 95% Confidence interval of (54%, 62%).

Table 3.2.2: Estimates for attendance rates, variance and standard error by stratum using estimator B.

Stratum	Level	Time	Estimate	Variance	Standard

					Error
1	0,1	9,10	41%	0.045%	2.12%
2	2	9,10	48%	0.273%	5.22%
3	3,4	9,10	46%	0.004%	0.59%
4	0,1	All other times	26%	0.025%	1.57%
5	2	All other times	45%	0.511%	7.14%
6	3,4	All other times	76%	0.184%	4.28%

Under ratio estimator B we obtained a total population mean attendance rate of 55% with a 95% Confidence interval of (51%, 59%)

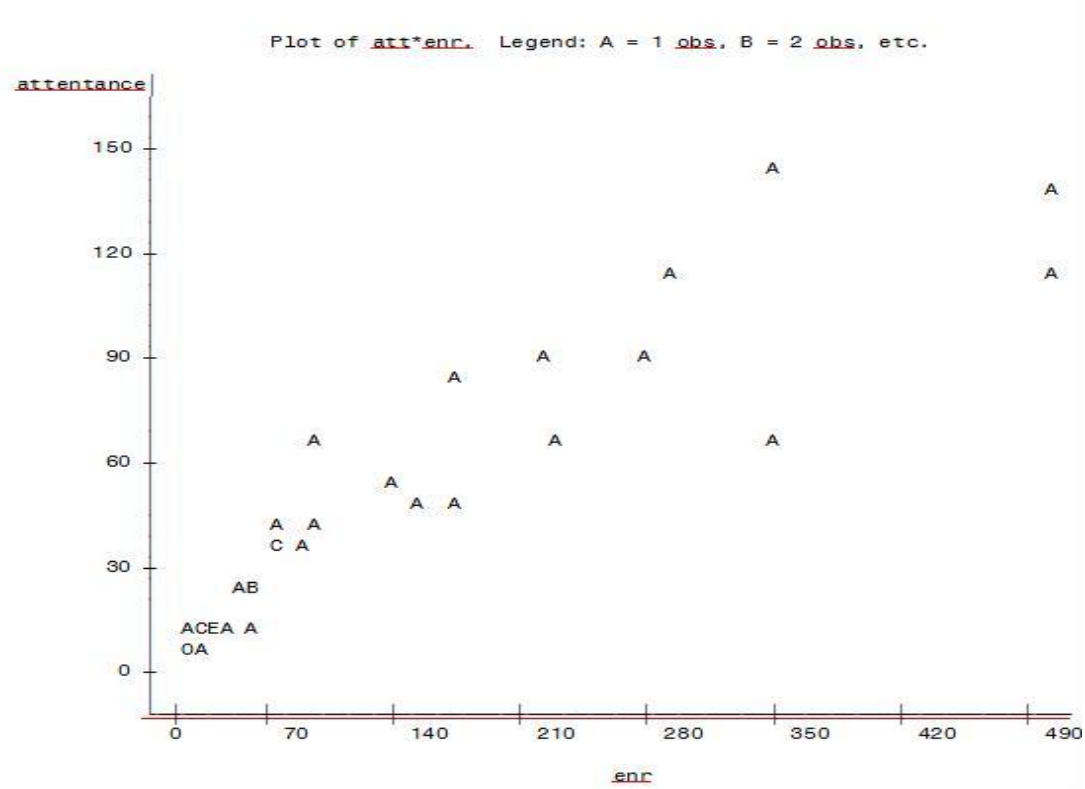
Table 3.2.3: Estimates for attendance rates, variance and standard error by stratum using estimator C.

Stratum	Level	Time	Estimate	Variance	Standard Error
1	0,1	9,10	46%	0.189%	4.34%
2	2	9,10	53%	0.417%	6.45%
3	3,4	9,10	71%	0.344%	5.86%
4	0,1	All other times	35%	0.178%	4.21%
5	2	All other times	54%	0.291%	5.39%
6	3,4	All other times	84%	0.090%	2.99%

Under ratio estimator C we obtained a total population mean attendance rate of 66% with a 95% Confidence interval of (62%, 70%).

3.4 Model Selection

Graph 3.4.1: SAS plot of attendance against enrolment.




```
NOTE: 4 obs had missing values.
```

The graph above indicates that the relationship between attendance and enrolment is approximately linear through the origin with increasing variance. The increase in variance appears to be linear so we conclude that ratio estimator A is the most suitable.

Section 4 - Discussion

4.1 Comments on the Results

Best and Worst Strata:

Attendance rates varied amongst the strata from 28% to 78%. The stratum with the highest attendance rate was stratum 6 at 78%. This was the stratum of levels 3 and 4 with lectures from 11:00 until the evening. The stratum with the lowest attendance rate of 28% was Stratum 4 which was made up of levels 0 and 1 with lectures from

11:00 until the evening. These results match our intuition; we would expect students in their final year and MSc students to have a high attendance rate because most would be quite dedicated to their course. First years, who may not take their course/lectures as seriously, would have a lower attendance rate and this is reflected in our results.

Size of Estimated Standard Errors:
Our estimated standard errors range from 2.28% to 6.78%. These are very good as they are very small and result in variances that are also very small.

Was Our Sample Size Large Enough?
We were very successful and accurate in choosing our sample size. We initially set out to attain a bound of 6% and we actually obtained a bound of 4.257%. This is quite a bit lower than what we were expecting and this is a better result.

Should We Have Chosen the Strata Differently?
There is a possibility that we should have chosen our strata differently. As mentioned above, the attendance rates amongst the levels match what our intuition tells us, so stratifying with regards to levels turned out to work quite well. However, stratifying with regards time of day may not have worked out as well as we all expected it to. This is clearly demonstrated amongst levels 0 and 1 which correspond to strata 1 and 4. The attendance rate for these students from 9:00 to 11:00 is 42%, while that from 11:00 am to the evening is 28%. This is a surprising result to obtain as it is expected that, especially among levels 0 and 1, there is a lower attendance rate in the mornings than in the afternoons/evenings. These findings could lead us to believe that perhaps stratifying by time of day may not be as interesting as expected.

4.2 Problems that Arose

Missing Values and the Reasons:
At the end of the sampling week, we had a total of six missing values. Two of these missing values resulted from the lecturer being sick. There was a call-back for these two lectures the following week and they were both successful. Another missing value resulted from the lecture being moved to a time earlier in the week. A call-back was not possible in this case. The remaining three missing values were as a result of not being able to locate the lecture, call-backs were attempted for these lectures but none were successful. Hence, our final data contains four missing values.

Why We Didn't Send Out a Letter:
After the lectures for our sample were selected, we would have liked to have sent a letter to each of the lecturers involved to let them know they had been selected. However, this was not possible because we did not have access to the full list of lecturers. We had a full list of contacts, but the contacts for some subjects were the module organiser and not specifically the lecturer. Hence, we only initially sent out a general letter to all organisers/lecturers who were potentially involved before our sample space was selected.

Problems with Enrolment Data:
We encountered some problems when it came to the enrolment data. We had a list of the number of students enrolled in each module and some lectures that were observed were made up of two module codes. To obtain the correct total enrolment figure for a module of this type, the enrolment data for both module codes were needed. In two or three instances, the observed attendance was larger than the enrolment. This was a problem we did not foresee. Some inquiries were made to the lecturers and the problem was solved. The reason for the large attendance rate was because there were PhD students sitting in on the lectures who were not actually registered for the class module. So the enrolment data for these classes were found by getting the total figure from the lecturer. The lecturer knew how many PhD students were attending the lectures and so this figure was added to the number enrolled in the class to get the accurate enrolment figure.

4.3 Other Relevant Information

There are a few discrepancies that need to be noted about this survey.

- The attendance rate of any lecture observed on the Wednesday of the sampling week would have been influenced by 'Arts Day'. There were 11 lectures sampled on Wednesday and the attendance rates could have been lower than usual on this day.
- Secondly, one of the call-backs fell on Halloween and there was a 'Mystery Tour' organised for all UCD students that day so the attendance rate observed for this lecture could be lower then usual.
- Thirdly, a total of four observations included in our survey were not taken during sampling week. Ideally, we would like all of our observations to be as consistent as possible, i.e. all observations taken on the same week. However, we believe the benefit of including these four observations resulting in a larger *n* outweighed the factor entailed due to inconsistency.

4.4 Improvements

This survey was carried out very thoroughly and professionally, but there is room for improvement.

- Several of the missing values were as a result of not being able to find the correct lecture venue or else the lecture time was moved to another time/day during the week. To avoid this happening, the observers could carry out a trial the week before the sampling week to ensure the lecture times and venues were correct. If these were found to be incorrect, then this could be rectified for the following week. This would result in less missing values; however it would be expensive, time wise.
- The data for the unregistered PhD students was needed in some instances where the attendance was found to be higher than the enrolled, but it is possible that this data was needed for more lectures. It would be hard to know which classes had PhD students sitting in if the attendance was less than enrolment, and so these cases wouldn't be brought to our attention. Hence, it is possible some of our enrolment data is slightly lower than it should be.
- Two lectures that were observed in Stratum 4 had enrolment=481 but these lectures were held in a lecture hall that had a maximum capacity of 395 students. This is a difference of 86 and it occurs twice. The enrolment data is

correct; however a large number of the students enrolled in this class are repeat students. Most repeat students do not attend lectures and this explains the large difference between attendance and enrolment.

4.5 Suggestions for the Future

This survey is an excellent starting point for future students taking Survey Sampling. A few suggestions for this survey in the future would include:

- Students having more knowledge of the college and locations of lecture halls so to prevent missing values.
- Sampling a larger school, e.g. the School of Engineering, as the results obtained from a school of this size might better reflect the tendencies of attendance rates throughout the whole of UCD.

4.6 Conclusion

The survey we carried out was a huge success. Given the time constraints and the fact that this was the very first survey we had undertaken, we feel the exercise was an excellent achievement. It was carried out in a very professional manner and we received no complaints from lecturers. Every student in the class participated and we all agree it was a very useful exercise. Seeing an application of this subject helped us understand the syllabus in greater depth. Not only did we learn how to conduct a survey, but we also witnessed the detailed preparation that was required, such as: initial compiling of the population data, carrying out the observations, data compiling and most importantly, data analysis. As well as this, we learned the advantages of stratifying data and this helped greatly in understanding the material of our course; STAT30020/40220 Survey Sampling.

APPENDIX 1 – SAS CODE

Six new datasets were created; one for each stratum. These datasets were then analysed using proc surveyreg in SAS as below.
Method A: Weight = (1/x)
Method B: Weight =1
Method C: Weight = $(1/x^2)$

```
data survey.str1;
       set survey.totallist;
       wt=(1/x);
       wt2=(1/(x*x));
       if stratum=1;
       keep stratum x y wt wt2;
run;
data survey.str2;
       set survey.totallist;
       wt=(1/x);
       wt2=(1/(x*x));
       if stratum=2;
       keep stratum x y wt wt2;
run;
data survey.str3;
       set survey.totallist;
       wt=(1/x);
       wt2=(1/(x*x));
       if stratum=3;
       keep stratum x y wt wt2;
run;
data survey.str4;
       set survey.totallist;
       wt=(1/x);
       wt2=(1/(x*x));
       if stratum=4;
       keep stratum x y wt wt2;
run;
data survey.str5;
       set survey.totallist;
       wt=(1/x);
       wt2=(1/(x*x));
       if stratum=5;
       keep stratum x y wt wt2;
run;
data survey.str6;
       set survey.totallist;
       wt=(1/x);
       wt2=(1/(x*x));
       if stratum=6;
       keep stratum x y wt wt2;
run;



/* Method A: Weight = (1/x) */

proc surveyreg data=survey.str1 total=21;
model y=x/noint;
weight wt;
run;
proc surveyreg data=survey.str2 total=13;
model y=x/noint;
weight wt;
run;
proc surveyreg data=survey.str3 total=32;
```

```
model y=x/noint;
weight wt;
run;
proc surveyreg data=survey.str4 total=25;
model y=x/noint;
weight wt;
run;
proc surveyreg data=survey.str5 total=27;
model y=x/noint;
weight wt;
run;
proc surveyreg data=survey.str6 total=85;
model y=x/noint;
weight wt;
run;

/* Method B: Weight = 1 */

proc surveyreg data=survey.str1 total=21;
model y=x/noint;
run;
proc surveyreg data=survey.str2 total=13;
model y=x/noint;
run;
proc surveyreg data=survey.str3 total=32;
model y=x/noint;
run;
proc surveyreg data=survey.str4 total=25;
model y=x/noint;
run;
proc surveyreg data=survey.str5 total=27;
model y=x/noint;
run;
proc surveyreg data=survey.str6 total=85;
model y=x/noint;
run;

/* Method C: Weight = (1/(x^2)) */

proc surveyreg data=survey.str1 total=21;
model y=x/noint;
weight wt2;
run;
proc surveyreg data=survey.str2 total=13;
model y=x/noint;
weight wt2;
run;
proc surveyreg data=survey.str3 total=32;
model y=x/noint;
weight wt2;
run;
proc surveyreg data=survey.str4 total=25;
model y=x/noint;
weight wt2;
run;
proc surveyreg data=survey.str5 total=27;
model y=x/noint;
weight wt2;
run;
proc surveyreg data=survey.str6 total=85;
model y=x/noint;
weight wt2;
run;
```