MSDS 6370 - Project

Estimating the Mean Length of Films

Damon Resnick, Rajni Goyal
11/22/2017

Abstract

This report details the analysis of estimating the mean length of films obtained from a list of films provided by James Eagan.[1] In the analysis below, we explore two of the most widely used techniques of statistical sampling, simple random sampling, and stratified sampling. We compare each technique and discuss the statistics of the dataset and conclude which technique performed better and why.

Introduction

Sampling is incredibly useful because it gives us the opportunity to have statistical confidence in the data, as well as an understanding of what might happen if the assumptions are wrong. The sample is selected in a careful way to increase the likelihood that the sample is truly representative of the population. Careful sampling is an extremely important part of the statistical estimation process, and reduces the risk of the sample being biased and non-representative of the data. Using the proper sampling design reduces the sampling error and is the major goal of any selection technique.

One of the most important parts of sampling is selecting the proper sample size. The sample should be big enough to answer the research question, but not so big that the process of sampling becomes uneconomical. In general, a larger sample size is needed for estimating values that come from a widely varying population, while a smaller sample size is needed to estimate values that come from a uniform population. Overall the larger the sample, the smaller the sampling error and the better job we can do. Therefore, large sample sizes are better for estimating values from data, but a smaller sample of the data is generally only needed depending on the confidence or margin of error one desires.

Data Description

The film industry has grown and evolved rapidly since its inception in the early 1900's. Over the years, it has gained the attention of every culture in every part of the world. These days movies are instrumental in shaping the social culture by transforming the viewers' opinions or swaying them one way or another. Several of these changes in popular English-language filmmaking practices are reflected in patterns of film style as distributed over the length and the subject of movies. In this report we make use of a movie dataset which has about 1600 movies with properties such as length, main actor and actress, director, and popularity. This movie dataset was taken and cleaned up to make it better suited to have statistical analysis performed on it.

Data Cleanup

After looking at the dataset, it was found that the data was a bit messy, so it was cleaned. It was found several rows had missing values of Popularity, Actor, Actress, and Director. We therefore removed rows that had missing data as that data was not of similar value as the other rows. Rows with Popularity of less than 10 were also removed. Several rows had the subject as "Westerns" instead of "Western", so we changed the rows from "Westerns" to "Western". Also removed were rows that only had one Subject, such as Crime, Fantasy, Romance, and Short. The Image column was also removed as it provided no use for this analysis.

The dataset obtained from all this cleaning includes 1327 rows with the columns Year, Length, Title, Subject, Popularity, and Awards. Year spans from 1920 to 1997, Length from 5 to 360, all rows have different Titles, nine different Subjects which are used as Strata and are detailed below, Popularity from 10 to 88, and Awards either Yes or No.

Task 1

The first problem is to estimate the mean length of the 1327 films in our final data set by using a stratified method and comparing that to another sampling method. Based on margin of error calculations a sample size of 120 for a simple random sample is appropriate for a MOE of 5 minutes.[2] This was determined using equation 1 and equation 2 below. However, in order to show the benefits of a stratified sample better we decided to use a margin of error of 9 minutes so that the total sample size would be roughly 1/3 of 120, or 40.

(Eq. 1)
$$n_{0,srs} = \frac{(Z_{\alpha/2}S)^2}{(I_{(1-\alpha)\%})^2}$$

(Eq. 2)
$$n_{srs} = \frac{n_{0,srs}}{1+\frac{n_{0,srs}}{N}}$$

There are 9 different Subjects of films in our data set: Action, Comedy, Drama, Horror, Musicals, Mystery, Science Fiction, War, and Western. These Subjects are taken to be different the 9 different strata. Below is Table 1 detailing the size of each strata and the mean, standard deviation, and the five-number summary for the Length of films of the entire population. These strata sizes and standard deviations are used in Equation 3 below to calculate the number of samples needed for each strata using a Neyman allocation stratified design. The calculated sample sizes are as follows: 4, 6, 19, 1, 2, 2, 1, 1, 4 for each Strata/Subject alternately. However, since the sample sizes of 1 are too small to calculate a reliable standard error the three 1 sized samples were enlarged to 2 and the 19-sized sample was reduced to 16 so that the total sample size would remain 40. The resulting samples sizes become: 4, 6, 16, 2, 2, 2, 2, 2, 4 for the 9 respective strata in the order give in the table.

(Eq. 3)
$$n_h = n \frac{\sigma_h N_h}{\sum_{h=1}^{H} N_h \sigma_h}$$

Table 1 below shows that the standard deviation of Length for the different strata varies widely. While the largest strata, Drama, has a larger than average variance it does not have the largest variance. This diversity of variances makes using a Neyman stratified method an ideal choice.

Table 1

Strata	Subject	N	Mean	Std Dev	Minimum	Maximum	Lower Quartile	Median	Upper Quartile
1	Action	151	103.3	22.49	53	226	90	99	111
2	Comedy	315	96.93	17.36	26	160	90	98	107
3	Drama	538	112.45	32	28	360	94	106	122
4	Horror	50	94.86	16.02	64	144	87	92.5	103
5	Musicals	38	90.97	41.69	5	172	60	92	118
6	Mystery	81	103.02	26.94	30	265	90	102	118
7	Sci-Fi	30	105.5	19.55	51	139	90	103	117
8	War	24	113.08	38.71	15	175	98	115	137
9	Western	100	95.07	40.08	52	298	59.5	94	110

Using a sample size of **40** a **Neyman** and **SRS** method was used to sample the 1327 population. These samples were done in SAS, with a seed value of 10, using PROC SURVEYSELECT. Then PROC SURVEYMEANS was used to perform the estimate of the mean and the standard error. (See Appendix for SAS code.)

Results of estimating the mean of the Length variable with a **Neyman** Allocation Stratified sample using a seed of **10** in SAS:

Statistics								
Variable	Mean	Std Error of Mean	95% CL for Mean		Sum	Std Error of Sum	95% CL for Sum	
Length	107.040222	7.024755	92.7131392	121.367305	142042	9321.850420	123030.336	161054.414

Results of estimating the mean of the Length variable with a **Simple Random Sample** method using a seed of **10** in SAS:

Statistics								
Variable	Mean	Std Error of Mean	95% CL for Mean		Sum	Std Error of Sum	95% CL for Sum	
Length	107.550000	6.118835	95.1734881	119.926512	4302.000000	244.753398	3806.93952	4797.06048

Comparing these two methods we see that for, this seed at least, the stratified method estimates a mean slight closer to the actual mean, **104.42**, with a higher standard error.

Task 2

Using the methods from Task 1 we now compare the results of 5 different seeds for each method. Seeds of 10, 20, 30, 40, and 50 were used with the SAS code from Task 1. The Seed, Means, Standard Error, and 95% Confidence Interval for each sample are shown in the summary tables below. The numbers in bold at the bottom of the tables are the averages of all 5 results.

Neyman Allocation Summary Table

Seed	Mean	Std Error	95% CL for Mean	
10	107.04	7.025	92.713	121.367
20	104.84	5.391	93.847	115.837
30	107.33	6.185	94.713	119.943
40	102.24	2.455	97.232	107.245
50	105.67	3.926	97.666	113.681
	105.42	**4.9964**	**95.2342**	**115.6146**

SRS Summary Table

Seed	Mean	Std Error	95% CL for Mean	
10	107.55	6.119	95.173	119.927
20	111.08	7.862	95.173	126.977
30	99.7	3.234	93.159	106.241
40	101.15	4.613	91.82	110.48
50	105	4.308	96.285	113.715
	104.9	**5.2272**	**94.322**	**115.468**

As you can see from the two summary tables the stratified method gives a better standard error and smaller 95% confidence interval on average for the five-different analyses. This is expected because of the variation in standard deviation across strata. However, this is not assured because only 5 different samples were taken and compared. There is a significant probability that the specific 5 seeds chosen to compute the 5 different samples could have provided a result that was different than this one. Many more samples would probably be needed to be more certain that this stratified design, rather than this SRS, is more appropriate to estimate the mean for this population.

The actual value of the mean Length is **104.42**. For the seed values used in the case the SRS got closer to the actual mean on average, however the standard error is smaller for the stratified method. Again this shows the stratified method performs better even though it did not get closer to the actual mean. The Neyman method did provide better certainty in the estimate. This is clear as the average standard error is smaller and the average 95% confidence interval is more narrow.

All of the seeds chosen for both the stratified and SRS designs had the actual value of the mean fall within the 95% confidence interval. It should be mentioned that the seed 40 for the Neyman allocation came the closest to falling outside of the confidence interval. Seed 40 also has the smallest error associated with it for the Neyman allocation.

It should be noted once again that since only 5 different samples were used in this comparison we cannot be certain that one method was better than the other. It is difficult to estimate exactly how many different samples we would need to show a significant difference.

Conclusion

With **these** seed values we found that a Neyman Allocation technique gives better results than a Simple Random Sample technique with a lower standard error and narrow 95% confidence intervals. This is expected since the standard deviation of the different strata are non-uniform. The use of a stratified method that takes into account the variation in standard deviation insures a highly accurate estimation from a sample that is well representative of the population. However, since the difference in the means was only 0.5%, with a standard error of roughly 5%, we **can't** really say there is a significant difference between the two techniques used because there were only 5 different samples. Many more samples are needed in order to show a significant difference between the two methods used for this particular population.

References

[1] https://perso.telecom-paristech.fr/eagan/class/igr204/datasets

[2] Excel spreadsheet of calculations of total sample size and strata sample size.

Appendix

SAS Code

```
proc import datafile='C:\Users\hp\Desktop\SMU\Statistical
Sampling\Project\filmdata_cleaned_trimmed_3c.csv' dbms=csv out=filmdata
replace;
delimiter = ",";
getnames=yes;
guessingrows=1000;
run;

data filmdata2;
      set filmdata;
      if Subject = "Action" then Strata = 1;
      if Subject = "Comedy" then Strata = 2;
      if Subject = "Drama" then Strata = 3;
      if Subject = "Horror" then Strata = 4;
      if Subject = "Music" then Strata = 5;
      if Subject = "Myster" then Strata = 6;
      if Subject = "Scienc" then Strata = 7;
```

```
      if Subject = "War" then Strata = 8;
      if Subject = "Wester" then Strata = 9;
      run;


proc means data = filmdata2 maxdec=2 N mean std min max q1 median q3;
class Strata Subject;
var Length;
      run;

data strsizes;
input Strata _total_;
datalines;
1 151
2 315
3 538
4 50
5 38
6 81
7 30
8 24
9 100
;
run;


/* Neyman */

proc surveyselect data=filmdata2 method = srs out = neysample sampsize =
(4,6,16,2,2,2,2,2,4) seed=50;
strata Strata;
title "Neyman allocation for Length of Film";
run;


proc surveymeans data = neysample sum clsum total = strsizes mean sum CLSUM
clm;
var Length;
weight SamplingWeight;
strata Strata;
title "Neyman allocation for Length of Film";
run;


/* SRS */

proc surveyselect data=filmdata2 method = srs out = srssample sampsize = 40
seed=50;
title "Simple Random Sample allocation for Length of Film";
run;

proc surveymeans data = srssample total = 1327 sum clsum mean sum CLSUM clm;
var Length;
title "Simple Random Sample allocation for Length of Film";
run;
```