STAT 344

Group Project

Exploring the online news popularity data in 2013 and 2014

Group Leader:
Xinyu Li

Role and contributions:
Wanqing Hu 65768004 (SRS proportion analysis, conclusion, final review)
Luna Li 85581379 (Stratification mean analysis, conclusion, report compilation)
Xinyu Li 70079264 (Stratification proportion analysis, intro, code integration)
Xiaolei Lin 55949507 (SRS mean analysis, intro, Part II)

2023-11-11

1 INTRODUCTION

Nowadays, online platforms are becoming the key component of digital news sharing.

Nowadays, online platforms are becoming the key component of digital news sharing. Fortunately, we found a dataset on online news popularity that can, to some extent, support the exploration of patterns related to online news, such as the number of shares of articles and the positivity of articles.

1.1 DATA SOURCE

The dataset OnlineNewsPopularity.csv is from the UC Irvine Machine Learning Repository. It summarizes a set of features about articles published by Mashable (www.mashable.com) in a period of two years, 2013 and 2014. The entire dataset contains 39,644 observations, with 18,199 articles in 2013 and 21,445 articles in 2014. Each article can be categorized into one of the following channels: "Business", "Entertainment", "Lifestyle", "Social Media", "Tech", "World" or "Other" (for articles that do not align with the predefined channels). The number of articles in each channel varies from approximately 1,000 to 5,700.

1.2 TARGETED POPULATION AND PARAMETER OF INTEREST

Given the large number of observations and the fact that they are all articles published by Mashable in a period of two years, we may consider the whole dataset as the population. Under this assumption, we can obtain information such as population size, stratum size and stratum variances.

In this report, we conduct a comprehensive analysis of online news articles spanning the years 2013 to 2014, with a specific emphasis on two parameters of interest: the change in the population mean number of shares through websites and the change in the population proportion of positive articles. For the continuous variable, we utilize the 'shares' column that records the number of shares of each article from the original data set, while the binary variable assesses the positivity of words through the 'rate_positive_words' and 'rate_negative_words' columns. The binary classification involves assigning '1' to articles with a higher rate of positive words and '0' to those dominated by negative words.

1.3 SAMPLING STRATEGY

We obtain the SRS by first grouping the population based on the years (2013 and 2014) and randomly selecting 1000 samples from 2013 news and 2014 news respectively.

Intuitively, the number of shares and proportion of positive words in certain articles varies based on the types of channels. Therefore, we consider stratifying articles based on their origin channels.

To ensure that we get the same sample every time, we set a seed of 10 beforehand.

2 INVESTIGATING AVERAGE NUMBER OF SHARES IN 2013 AND 2014

In this section, our parameter of interest is the average number of shares of articles in 2013 and 2014, and their difference.

2.1 USING SIMPLE RANDOM SAMPLING (SRS) METHODOLOGY

To estimate whether the average number of shares changed from 2013 to 2014, independent random samples were drawn from each year, each with a sample size of $\mathrm{n} = 1000$. After obtaining two simple random samples, we calculate the sample mean of each sample and take the difference between 2013 and 2014 and our vanilla estimator $\bar{y_s}$:

$$\bar{y}_{\mathrm{s}} = \overline{y_{2014}} - \overline{y_{2013}} = 4174.28 - 4153.21 = 21.07$$

We also calculate the standard error of sample mean,

$$\mathrm{SE}\overline{y_{2013}} = \frac{s_{2013}}{\sqrt{n}} = 497.44 \ , \quad \mathrm{SE}\overline{y_{2014}} = \frac{s_{2014}}{\sqrt{n}} = 727.97$$

in which s_{2013}, s_{2014} denote the sample standard deviation for each year respectively.

We then calculate the estimated standard error of our estimator with the formula:

$$\mathrm{SE}(\bar{y}_{\mathrm{s}}) = \sqrt{\left(1 - \frac{\mathrm{n}}{\mathrm{N}_{2013}}\right) \cdot \frac{\mathrm{s}_{2013}^2}{\mathrm{n}} + \left(1 - \frac{\mathrm{n}}{\mathrm{N}_{2014}}\right) \cdot \frac{\mathrm{s}_{2014}^2}{\mathrm{n}}}$$

Here, N_{2013} represents the population size in 2013, and N_{2014} represents the population size in 2014. The Finite Population Correction factor $\left(1 - \frac{n}{N}\right)$ is included in the calculation of the estimated standard error, as we assumed the original data represents the finite population. The calculation yields $\mathrm{SE}(\overline{y_s}) = 859.69$.

Finally, we construct the confidence interval at 95% confidence level using $\overline{y_s} \pm 1.96 \times S\mathrm{E}$, resulting in a confidence interval of $(-1663.93,\ 1706.06)$. Based on this interval, we cannot conclude that the average number of shares changed from 2013 to 2014 at 95% confidence level, as the confidence interval includes 0.

2.2 USING STRATIFIED SAMPLING METHODOLOGY

The summary of the data shows that each article can be categorized into various channels, including "Business", "Entertainment", and more. Therefore, stratified sampling might be a plausible method to derive an estimate of the average number of shares for news articles in 2013.

However, before proceeding, verifying whether 'channel' serves as an effective stratum is crucial. To assess this, we analyzed the between-stratum variance and found it to be greater than 1,000,000. Consequently, we may conclude that variations exist in the number of shares between channels, suggesting that stratification could be a favorable approach.

To determine the sample sizes for each stratum, we utilize optimal allocation and take into account two crucial factors: the stratum population size N_h and the stratum standard deviation S_h. The population stratum size N_h represents the total number of articles in the population within each channel, and the stratum standard deviation $\mathrm{S_h}$ reflects the variability in the number of shares within each channel. We do not consider the cost of sampling in this case. The sample size for each stratum is calculated as follows:

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} n \quad (1)$$

With total sample size $n = 1000$, the stratum sample sizes n_h calculated after rounding are:

"Business"	"Entertainment"	"Lifestyle"	"Social Media"	"Technology"	"World"	"Other"
291	118	52	40	91	67	340

We obtain the stratified sample by randomly selecting n_h samples from each channel and calculate the stratum average number of shares in our sample, $\overline{y_{s,h}}$.

After implementing stratified sampling, the stratified estimate for the average number of shares for news articles in 2013 is as below:

$$\overline{y_{str,2013}} = \sum_{h=1}^{H} \frac{N_h}{N} \overline{y_{s,h}} \quad (2)$$

$$= 3525.355$$

Furthermore, we are interested in the accuracy of this estimate by examining its standard error. The standard error for the average number of shares in each stratum can be computed as:

$$SE_{\overline{y_{s,h}}} = \sqrt{\left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}}$$

We can proceed to compute the standard error for our stratified estimate as:

$$SE_{\overline{y_{str,2013}}} = \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 SE_{\overline{y_{s,h}}}^2} \quad (3)$$

$$= 265.165$$

In comparison to our earlier estimate derived from simple random sampling (SRS), the stratified estimate has decreased from 4153.21 to 3525.36. More notably, we have successfully reduced the standard error of the estimation from 497.44 to 265.17, resulting in a 46.7% improvement in accuracy! Thus, we are confident to say that stratification is a more appropriate approach to estimate the average number of shares in this population.

For consistency, we apply the same approach to the data for the year 2014. We can compute the sample sizes within each channel using the same formula (1) as previously.

```
"Business" "Entertainment" "Lifestyle" "Social Media" "Technology" "World"   "Other"
   136         160               41           21              221           197       223
```

We calculated the stratified estimate for 2014 to be 2983.25 using formula (2), with a corresponding standard error of 176.57 computed by formula (3). The estimation is also lower than the estimate derived from SRS which is 4147.28. The improvement in accuracy is particularly significant, decreasing from 727.97 to 176.57 (75.7% better).

One possible explanation for the better performance of $SE_{\overline{y_{str,2014}}}$ comparing to $SE_{\overline{y_{str,2013}}}$ is that the between-stratum variance of 2014 data is not only absolutely larger than the between-stratum variance of 2013 data, but also relatively larger. The between-stratum variance of 2014 accounts for more than 1.5% of the total variation, while the between-stratum variance of 2013 accounts for far less than 1% of the total variation. This aligns with the intuition that stratification is more effective than SRS particularly when between-stratum variance takes a larger proportion of total variation.

Intuitively, the difference of average number of shares between 2013 and 2014 shows that there might be a decrease in the number of shares.

Stratification	Estimated average number of shares	Standard Error
2013	3525.36	265.17
2014	2983.25	176.57

We can construct a confidence interval to attest our hypothesis that the difference of average number of shares between 2013 and 2014 is not significant. The difference can be calculated as follows:

$$\overline{y_{str,2014}} - \overline{y_{str,2013}} = -542.1083$$

The corresponding SE will be:

$$SE_{\overline{y_{str,2014}} - \overline{y_{str,2013}}} = \sqrt{SE^2_{\overline{y_{str,2013}}} + SE^2_{\overline{y_{str,2014}}}} = 318.58$$

Since our sample size is large enough, we can reasonably assume a normal distribution of the sample mean by the Central Limit Theorem. Therefore, a 95% confidence interval is:

$$CI = \left(\overline{y_{str,2014}} - \overline{y_{str,2013}}\right) \pm 1.96 * SE_{\overline{y_{str,2014}} - \overline{y_{str,2013}}} = (-1166.516, 82.299)$$

The confidence interval, which includes 0, suggests that we cannot reject the hypothesis of the difference being non-significant. There is not enough evidence that the average number of shares between 2013 and 2014 decreased.

3 INVESTIGATING THE POPULATION PROPORTION OF POSITIVE ARTICLES FROM 2013 TO 2014

In this section, the binary parameter of interest is the difference between population proportion of positive articles from 2013 to 2014, with "positive articles" defined as $rate\_positive\_words - rate\_negative\_words > 0$.

"positive articles"	1
"negative articles"	0

3.1 USING SIMPLE RANDOM SAMPLING (SRS) METHODOLOGY

Having large sample size is more likely to provide us with reliable inference for the population parameter, so we aim to have a relatively big sample size. We do not need to worry much about the SRS in terms of their sample size as long as the sample size n meets $n\hat{p} \geq 10, n(1-\hat{p}) \geq 10$ to satisfy the assumption of CLT, so that we could make inference for the population proportion. A sample size of 1000 satisfies the above condition.

After selecting the samples, we compute the proportion of positive articles for 2013 news and that for 2014 news, which equal 0.824 and 0.906. Then, the estimate of change in population proportion could be approached by:

$$\widehat{\triangle} = \widehat{p_{2014}} - \widehat{p_{2013}} = 0.824 - 0.906 = -0.082$$

Here, we use the vanilla estimate, by treating sample proportion of positive articles in 2013 and 2014 as estimates of the population proportion for those years. The result of this estimate is interpreted as: The population proportion of positive articles by Mashable is estimated to decrease 0.082 from 2013 to 2014.

With N_{2014} and N_{2013} representing the population size of 2014 news and 2013 news, the standard error of our estimate is given by

$$\mathrm{SE}(\widehat{\triangle}) = \sqrt{\left(\mathrm{Var}(\widehat{p_{2014}} - \widehat{p_{2013}})\right)} = \sqrt{\left(\mathrm{Var}(\widehat{p_{2014}}) + \mathrm{Var}(\widehat{p_{2013}})\right)}$$

where

$$\mathrm{Var}(\widehat{p_{2014}}) = \frac{\widehat{p_{2014}}(1-\widehat{p_{2014}})}{1000} \cdot \left(1 - \frac{1000}{N_{2014}}\right)$$

and

$$\mathrm{Var}(\widehat{p_{2013}}) = \frac{\widehat{p_{2013}}(1-\widehat{p_{2013}})}{1000} \cdot \left(1 - \frac{1000}{N_{2013}}\right)$$

which turns out to approximately equal to 0.01479. It means that the standard error of our estimate for the change in population proportion of positive articles from 2013 to 2014 is about 0.01479.

The 95% confidence interval is

$$\left(\widehat{\triangle} \pm 1.96 \times \mathrm{SE}\left(\widehat{\triangle}\right)\right) = (-0.082 \pm 1.96 \times 0.01479) = (-0.11099, -0.05301)$$

We are 95% confident that the true change in proportion of positive articles from 2013 to 2014 falls within the interval $(-0.11099, -0.05301)$. Since the 95% confidence interval excludes 0, we conclude that there may be a change, specifically a decrease, of the proportion of positive articles from 2013 to 2014.

3.2 USING STRATIFIED SAMPLING METHODOLOGY

Using channels as stratification criteria, stratum sample size can be calculated from within-strata variance. To assess if the stratification is effective, between-strata variance is also constructed. Below has shown that within-strata variance is relatively larger than between-strata within-strata variance, indicating that strata might not be effective as it should.

$$\sum_{n=1}^{H} \frac{\mathrm{N_h}}{\mathrm{N}} \mathrm{S}^2_{\mathrm{P,h}} = 0.01 \qquad \sum_{n=1}^{H} \frac{\mathrm{N_h}}{\mathrm{N}} \left(\overline{\mathrm{p_{P,h}}} - \overline{\mathrm{p_P}}\right)^2 = 0.0002$$

Proportion of positive article inside each channel are derived separately for both articles in 2013 and 2014. With $\mathrm{p_{s,h}}$ represents the sample proportion inside each stratum, estimated proportion difference is then constructed as follow:

$$\overline{\mathrm{p_{str,2013}}} = \sum_{n=1}^{h} \frac{\mathrm{N_h}}{\mathrm{N}} \mathrm{p_{s,h}} = 0.91 \qquad \overline{\mathrm{p_{str,2014}}} = \sum_{n=1}^{h} \frac{\mathrm{N_h}}{\mathrm{N}} \mathrm{p_{s,h}} = 0.842$$

$$\overline{\mathrm{p_{str,2014}}} - \overline{\mathrm{p_{str,2013}}} = -0.071$$

Intuitively, it is obvious that proportion of the positive article had been decreased from 2013 to 2014. To test the significance of this difference, SE for both proportions are also constructed for further investigation. As the total population size for 2013 and 2014 are $N_{2013} = 18199$ & $N_{2014} = 21445$, and total sample size $n = 1000$, $\frac{n}{N_{2013}} = 0.0549 > 0.05$. FPC is used to adjust for small sample size. First, SE for the estimated proportion is calculated within each stratum:

$$SE\overline{p_{s,h}} = \sqrt{\sum_{n=1}^{h} \frac{N_h}{N}^2 \frac{s_h^2}{n_h}}$$

Second, SE for whole sample in 2013 and 2014 are calculated separately:

$$SE\overline{p_{str,2013}} = \sqrt{\sum_{n=1}^{h} \left(\frac{N_h}{N}\right)^2 \cdot SE^2_{\overline{p_{s,h}}}} = 0.010 \qquad SE\overline{p_{str,2014}} = \sqrt{\sum_{n=1}^{h} \left(\frac{N_h}{N}\right)^2 \cdot SE^2_{\overline{p_{s,h}}}} = 0.020$$

Finally, the corresponding SE for the estimated proportion change is as follow:

$$SE\overline{p_{str,2014} - p_{str,2013}} = \sqrt{\left(SE^2_{\overline{p_{str,2014}}} + SE^2_{\overline{p_{str,2013}}}\right)} = 0.023$$

Comparing to the standard error calculated using SRS sampling, which is 0.01479, stratified sampling gives a larger SE. It suggests stratifying data by channels might be inappropriate. This verify the thoughts that stratify sampling has its limitation when strata is not chosen wisely, estimation of the whole population's proportion change become less efficient.

Again, by Central Limit Theorem, a 95% confidence interval of the estimated proportion change can then be constructed:

$$\text{CI: } \left(\overline{p_{str,2014}} - \overline{p_{str,2013}}\right) \pm 1.96 * SE\overline{p_{str,2014} - p_{str,2013}} = (-0.131, -0.039)$$

As the confidence interval excludes 0, we come to the same conclusion as using SRS sampling that the decrease of the proportion of positive articles from 2013 to 2014 is significant.

4 CONCLUSION

Based on the sampling result above, the change in the average number of shares changed from 2013 to 2014 is not significant at 95% confidence level, whether using simple random sample or stratified sample. On the contrary, both results from SRS and stratified sampling indicate that there may be a change, specifically a decrease, of the proportion of positive articles from 2013 to 2014.

4.1 DISCUSSION

The advantage of simple random sampling is that SRS ensures that every member of the population has an equal chance of being selected, so it is highly representative of the population and minimizes the risk of selection bias. However, SRS may not represent the subgroups or reveal rare characteristics in the population adequately, which is a drawback compared to stratified sampling that takes samples from each subgroup. Stratified sampling performs particularly well in terms of accuracy when the between-stratum variation is large, whether relatively or absolutely.

However, one limitation of stratified sampling by channels, as we notice when comparing the within-stratum variance and between-stratum variance, is that the between-stratum variance can be relatively small comparing to the within-stratum variance. The within-stratum variation can account for over 98% of the total variation, while the between-stratum variation is less than 2% of the total variation. Hence, exploring alternative strata that can offer more pronounced variations between stratums may be a meaningful next step to enhance the effectiveness of stratified sampling.

Although both methods reach the same conclusion about the decrease in the proportion of positive articles from 2013 to 2014, it cannot be generalized to bigger population with “bigger population” defined as news articles from other sources or news articles over a longer time frame. This is because news reports from different media platforms often have distinct styles, leading to variations in the choice of wording. Besides, time is an important factor in determining the proportion of positive articles, since the positivity of an article is mostly based on the inherent quality of the events being described. Therefore, if we were to alter the time frame, we are taking more events into consideration so the proportion of positive articles is likely to differ. The time frame as a factor also influences the number of shares taken place, hence the conclusion about the mean shares cannot be generalized as well.

Reference

Fernandes,Kelwin, Vinagre,Pedro, Cortez,Paulo, and Sernadela,Pedro. (2015). Online News Popularity. UCI Machine Learning Repository. https://doi.org/10.24432/C5NS3V.

Part II

In this paper, the authors address the criticism that the likelihood ratio test (LRT) has faced in the past two centuries, particularly in multiparameter hypothesis testing problems. Critics believe that the LRT may be biased in certain cases. They claim that LRT is "inferior" by suggesting alternative tests that are less biased or even unbiased. However, the authors argue that these suggested alternative testing methods are flawed and inappropriate. They present several examples that were purported to perform poorly with the LRT, demonstrating that how LRT can, in fact, be applied successfully to obtain desirable results. In the end, the authors emphasize the role of intuition in statistical science.

Appendix

Group members

Nov 12th, 2023

```
library(tidyverse)
# Parameter: change in avg number of shares from 2013 to 2014
# SRS
# Note: data from 2013 and 2014 are independent
news <- read.csv("OnlineNewsPopularityOriginal.csv", header = T) %>%
  mutate(year_of_publish = as.integer(str_extract(url, "\\d{4}")),
         Positive_article =
           ifelse(rate_positive_words -  rate_negative_words > 0, 1, 0),
         channel = case_when(
           data_channel_is_lifestyle == 1 ~ "lifestyle",
           data_channel_is_entertainment == 1 ~ "entertainment",
           data_channel_is_bus == 1 ~ "business",
           data_channel_is_socmed == 1 ~ "socmed",
           data_channel_is_tech == 1 ~ "tech",
           data_channel_is_world == 1 ~ "world",
           TRUE ~ "other"
         )) %>%
  select(year_of_publish, channel, Positive_article, shares)

# set seed, sample from 2013 and 2014
set.seed(10)
n <- 1000
sample_2013 <- news |>
  filter(year_of_publish == 2013) |>
  sample_n(size = n, replace = FALSE)
sample_2014 <- news |>
  filter(year_of_publish == 2014) |>
  sample_n(size = n, replace = FALSE)

# calculate average of shares
mean_2013 <- mean(sample_2013$shares)
mean_2014 <- mean(sample_2014$shares)

# vanilla estimator
mean_diff <- mean_2014 - mean_2013
mean_diff
```

```
## [1] 21.067
```

```
# calculate se of shares
var_2013 <- var(sample_2013$shares)
```

```r
var_2014 <- var(sample_2014$shares)
N_2013 <- news |>
  filter(year_of_publish == 2013) |>
  nrow()
N_2014 <- news |>
  filter(year_of_publish == 2014) |>
  nrow()
se_mean_diff <- sqrt((1 - n / N_2013) * var_2013 / n +
                       (1 - n / N_2014) * var_2014 / n)
se_mean_diff
```

```
## [1] 859.6924
```

```r
# construct 95% confidence interval
CI <- data.frame(
  lower_ci = mean_diff - 1.96 * se_mean_diff,
  upper_ci = mean_diff + 1.96 * se_mean_diff
)
CI
```

```
##   lower_ci upper_ci
## 1 -1663.93 1706.064
```

```r
# Since the confidence interval contains 0, we cannot conclude if
# there is a change of average shares between 2013 and 2014.




# Parameter: change in population proportion of positive articles from 2013 to
# 2014
# SRS

#2013 subset
news_2013 <- news %>%
  filter(year_of_publish == 2013)

#2014 subset
news_2014 <- news %>%
  filter(year_of_publish == 2014)

#SRS: set sample size = 1000 for each year
set.seed(10)
combined_srs_prop <- as.data.frame(
  cbind(sample_2013$Positive_article, sample_2014$Positive_article))
colnames(combined_srs_prop) <- c("positive_2013", "positive_2014")

srs_2013_prop <- combined_srs_prop %>%
  summarize(positive_prop_2013 = mean(positive_2013 == 1)) %>%
  pull()

srs_2014_prop <- combined_srs_prop %>%
  summarize(positive_prop_2014 = mean(positive_2014 == 1)) %>%
```

```r
  pull()

# Estimate of change in population proportion (vanilla)
srs_prop_estimate <- srs_2014_prop - srs_2013_prop
srs_prop_estimate
```

```
## [1] -0.082
```

```r
# Standard error of our estimate
sample_size_each_yr <- 1000
srs_prop_estimate_se <- sqrt(
  srs_2013_prop * (1 - srs_2013_prop) / sample_size_each_yr *
    (1 - sample_size_each_yr  / nrow(news_2013)) + # FPC
    srs_2014_prop * (1 - srs_2014_prop) / sample_size_each_yr *
    (1 - sample_size_each_yr  / nrow(news_2014))) # FPC
srs_prop_estimate_se
```

```
## [1] 0.01479006
```

```r
# 95% confidence interval for change of population proportion
srs_prop_ci <- data.frame(
  lower_ci = srs_prop_estimate - 1.96 * srs_prop_estimate_se,
  upper_ci = srs_prop_estimate + 1.96 * srs_prop_estimate_se
)
srs_prop_ci
```

```
##     lower_ci    upper_ci
## 1 -0.1109885 -0.05301149
```

```r
# Since 95% confidence interval excludes 0, we conclude that there may be a
# change, specifically a decrease, of the proportion of positive articles from
# 2013 to 2014.


# Parameter: change in avg number of shares from 2013 to 2014
# Stratified sampling

news_2013$channel <- as.factor(news_2013$channel)
news_2014$channel <- as.factor(news_2014$channel)


#between stratum variance 2013
attach(news_2013)
N_2013 <- length(shares)
N_h2013 <- tapply(shares, channel, length)
N_h2013
```

```
##      business entertainment     lifestyle         other        socmed          tech
##          3194          2862          1191          3007          1369          3942
##         world
##          2634
```

```r
avg_all <- mean(news_2013$shares)
avg_b <- mean(news_2013$shares[news_2013$channel == "business"])
avg_e <- mean(news_2013$shares[news_2013$channel == "entertainment"])
avg_l <- mean(news_2013$shares[news_2013$channel == "lifestyle"])
avg_s <- mean(news_2013$shares[news_2013$channel == "socmed"])
avg_t <- mean(news_2013$shares[news_2013$channel == "tech"])
avg_w <- mean(news_2013$shares[news_2013$channel == "world"])
avg_o <- mean(news_2013$shares[news_2013$channel == "other"])

avg_stratum <- c(avg_b, avg_e, avg_l, avg_o, avg_s, avg_t, avg_w)
var_between <- sum((avg_stratum-avg_all)^2*(N_h2013/N_2013))
var_between
```

```
## [1] 1259600
```

```r
#stratum size
s_b <- sd(news_2013$shares[news_2013$channel == "business"])
s_e <- sd(news_2013$shares[news_2013$channel == "entertainment"])
s_l <- sd(news_2013$shares[news_2013$channel == "lifestyle"])
s_s <- sd(news_2013$shares[news_2013$channel == "socmed"])
s_t <- sd(news_2013$shares[news_2013$channel == "tech"])
s_w <- sd(news_2013$shares[news_2013$channel == "world"])
s_o <- sd(news_2013$shares[news_2013$channel == "other"])

s_all <- c(s_b, s_e, s_l,  s_o, s_s, s_t, s_w)
var_within <- sum((s_all^2)*(N_h2013/N_2013))
var_within
```

```
## [1] 193088642
```

```r
n <- 1000
n_h2013 <- (N_h2013*s_all/sum(N_h2013*s_all))*n
n_h2013
```

```
##      business entertainment     lifestyle         other        socmed          tech
##     290.92520     118.11732      51.59988     340.29705      40.29318      91.37970
##         world
##      67.38767
```

```r
n_h2013 <- round(n_h2013)
n_h2013
```

```
##      business entertainment     lifestyle         other        socmed          tech
##           291           118            52           340            40            91
##         world
##            67
```

```r
detach(news_2013)

#stratum mean
```

```r
attach(news_2013)
channels <- unique(channel)

STR.sample.prop2013 <- NULL
set.seed(10)
for (i in 1:length(channels)){
  row.indices <- which(channel == channels[i])
  sample.indices <- sample(row.indices, n_h2013[i], replace=FALSE)
  STR.sample.prop2013 <- rbind(STR.sample.prop2013, news_2013[sample.indices,])
}
ybar_h2013 <- tapply(STR.sample.prop2013$shares, STR.sample.prop2013$channel, mean)
var_h2013 <- tapply(STR.sample.prop2013$shares, STR.sample.prop2013$channel, var)
se_ybar_h2013 <- sqrt((1-n_h2013/N_h2013)*var_h2013/n_h2013)
rbind(ybar_h2013, se_ybar_h2013)
```

```
##                business entertainment lifestyle     other   socmed      tech     world
## ybar_h2013     2383.093     2933.4674 2966.8647 5696.7802 4361.060 3730.7885 2585.4000
## se_ybar_h2013   149.332      569.1975  699.1982  979.7712 1111.632  692.7222  426.2263
```

```r
detach(news_2013)

#stratum mean
ybar_str2013 <- sum(ybar_h2013*(N_h2013/N_2013))
SE_ybar_str2013 <- sqrt(sum(se_ybar_h2013^2*(N_h2013/N_2013)^2))
str.pop2013 <- c(ybar_str2013, SE_ybar_str2013)
cat("stratified mean",":", ybar_str2013,"\n")
```

```
## stratified mean : 3525.355
```

```r
cat("stratified SE",":", SE_ybar_str2013,"\n")
```

```
## stratified SE : 265.1654
```

```r
#For 2014
attach(news_2014)

N_2014 <- length(shares)
N_2014
```

```
## [1] 21445
```

```r
N_h2014 <- tapply(shares, channel, length)
N_h2014
```

```
##      business entertainment     lifestyle         other        socmed          tech
##          3064          4195           908          3127           954          3404
##         world
##          5793
```

```r
avg_all2014 <- mean(news_2014$shares)
avg_b2014 <- mean(news_2014$shares[news_2014$channel == "business"])
avg_e2014 <- mean(news_2014$shares[news_2014$channel == "entertainment"])
avg_l2014 <- mean(news_2014$shares[news_2014$channel == "lifestyle"])
avg_s2014 <- mean(news_2014$shares[news_2014$channel == "socmed"])
avg_t2014 <- mean(news_2014$shares[news_2014$channel == "tech"])
avg_w2014 <- mean(news_2014$shares[news_2014$channel == "world"])
avg_o2014 <- mean(news_2014$shares[news_2014$channel == "other"])

avg_stratum2014 <- c(avg_b2014, avg_e2014, avg_l2014, avg_o2014, avg_s2014, avg_t2014, avg_w2014)

var_between2014 <- sum((avg_stratum2014-avg_all2014)^2*(N_h2014/N_2014))
var_between2014
```

```
## [1] 1405712
```

```r
s_b2014 <- sd(news_2014$shares[news_2014$channel == "business"])
s_e2014 <- sd(news_2014$shares[news_2014$channel == "entertainment"])
s_l2014 <- sd(news_2014$shares[news_2014$channel == "lifestyle"])
s_s2014 <- sd(news_2014$shares[news_2014$channel == "socmed"])
s_t2014 <- sd(news_2014$shares[news_2014$channel == "tech"])
s_w2014 <- sd(news_2014$shares[news_2014$channel == "world"])
s_o2014 <- sd(news_2014$shares[news_2014$channel == "other"])

s_all2014 <- c(s_b2014, s_e2014, s_l2014,  s_o2014, s_s2014, s_t2014, s_w2014)
var_within2014 <- sum((s_all2014^2)*(N_h2014/N_2014))


n <- 1000
n_h2014 <- (N_h2014*s_all2014/sum(N_h2014*s_all2014))*n
n_h2014
```

```
##      business entertainment     lifestyle         other        socmed          tech
##     135.88299     160.42930      40.82148     222.79994      21.42599     221.40565
##         world
##     197.23466
```

```r
sum(n_h2014)
```

```
## [1] 1000
```

```r
n_h2014 <- round(n_h2014)
n_h2014
```

```
##      business entertainment     lifestyle         other        socmed          tech
##           136           160            41           223            21           221
##         world
##           197
```

```r
detach(news_2014)

#stratum mean
attach(news_2014)
channels <- unique(channel)

STR.sample.prop2014 <- NULL
set.seed(10)
for (i in 1:length(channels)){
  row.indices <- which(channel == channels[i])
  sample.indices <- sample(row.indices, n_h2014[i], replace=FALSE)
  STR.sample.prop2014 <- rbind(STR.sample.prop2014, news_2014[sample.indices,])
}
ybar_h2014 <- tapply(STR.sample.prop2014$shares, STR.sample.prop2014$channel, mean)
var_h2014 <- tapply(STR.sample.prop2014$shares, STR.sample.prop2014$channel, var)
se_ybar_h2014 <- sqrt((1-n_h2014/N_h2014)*var_h2014/n_h2014)
rbind(ybar_h2014, se_ybar_h2014)
```

```
##                business entertainment lifestyle     other    socmed      tech      world
## ybar_h2014    3145.6502     2839.2562  3699.471 5602.4634 3248.2640 2438.8015 1751.80952
## se_ybar_h2014  557.0672      496.3236   743.486  762.2798  777.0406  139.3923   72.74968
```

```r
detach(news_2014)

#stratum mean
ybar_str2014 <- sum(ybar_h2014*(N_h2014/N_2014))
SE_ybar_str2014 <- sqrt(sum(se_ybar_h2014^2*(N_h2014/N_2014)^2))
str.pop2014 <- c(ybar_str2014, SE_ybar_str2014)

rbind(c("stratified mean","stratified SE"), round(str.pop2013,digits = 2)
      , round(str.pop2014, digits = 2))
```

```
##      [,1]              [,2]
## [1,] "stratified mean" "stratified SE"
## [2,] "3525.36"         "265.17"
## [3,] "2983.25"         "176.57"
```

```r
#CI
MeanDiff <- ybar_str2014 - ybar_str2013
SE_MeanDiff <- sqrt(SE_ybar_str2013^2 + SE_ybar_str2014^2)
CI <- c(MeanDiff - 1.96*SE_MeanDiff, MeanDiff + 1.96*SE_MeanDiff)
CI
```

```
## [1] -1166.51594    82.29943
```

```r
#The CI contains 0, and the difference is very small as we can tell.

# Parameter:  change in population proportion of positive articles from 2013 to
# 2014
# Stratified sampling
news_2013$Positive_article <- as.factor(news_2013$Positive_article)
```

```r
news_2014$Positive_article <- as.factor(news_2014$Positive_article)

#stratum size 2013 optimal allocation
attach(news_2013)

positive_2013 <- news_2013[news_2013$Positive_article == "1",]
negative_2013 <- news_2013[news_2013$Positive_article == "0",]
N_Pos_2013<- tapply(positive_2013$Positive_article, positive_2013$channel, length)
N_Neg_2013<- N_h2013 - N_Pos_2013
prop_2013 <- N_Pos_2013/N_h2013
avg_prop_2013 <- sum(N_Pos_2013)/sum(N_h2013)

prop_sd_within <- sqrt(prop_2013 * (1-prop_2013) / N_h2013)

var_h2013_between <- (prop_2013 - avg_prop_2013)^2
sum(var_h2013_between)
```

```
## [1] 0.01616685
```

```r
sum(prop_sd_within^2)
```

```
## [1] 0.0002112503
```

```r
n <- 1000
n_h2013_prop <- (N_h2013*prop_sd_within/sum(N_h2013*prop_sd_within))*n
n_h2013_prop
```

```
##      business entertainment     lifestyle         other        socmed          tech
##     145.40457     169.06544      74.54125     220.69923      74.06791     137.80518
##         world
##     178.41641
```

```r
n_h2013_prop <- round(n_h2013_prop)
n_h2013_prop
```

```
##      business entertainment     lifestyle         other        socmed          tech
##           145           169            75           221            74           138
##         world
##           178
```

```r
n <- sum(n_h2013_prop)  #adjust sample size due to rounding

detach(news_2013)

#strata size 2014 optimal allocation
attach(news_2014)

positive_2014 <- news_2014[news_2014$Positive_article == "1",]
negative_2014 <- news_2014[news_2014$Positive_article == "0",]
N_Pos_2014<-tapply(positive_2014$Positive_article, positive_2014$channel, length)
```

```r
N_Neg_2014<-N_h2014 - N_Pos_2014
prop_2014 <- N_Pos_2014/N_h2014
avg_prop_2014 <- sum(N_Pos_2014)/sum(N_h2014)

prop_sd_within2014 <- sqrt(prop_2014 * (1-prop_2014) / N_h2014)

var_h2014_between <- (prop_2014 - avg_prop_2014)^2
sum(var_h2014_between)
```

```
## [1] 0.06830889
```

```r
sum(prop_sd_within2014^2)
```

```
## [1] 0.0003311377
```

```r
n <- 1000
n_h2014_prop <- (N_h2014*prop_sd_within/sum(N_h2014*prop_sd_within))*n
n_h2014_prop
```

```
##      business entertainment     lifestyle         other        socmed          tech
##     112.79487     200.38934      45.95452     185.58919      41.73808      96.22677
##         world
##     317.30724
```

```r
n_h2014_prop <- round(n_h2014_prop)
n_h2014_prop
```

```
##      business entertainment     lifestyle         other        socmed          tech
##           113           200            46           186            42            96
##         world
##           317
```

```r
n <- sum(n_h2014_prop)  #adjust sample size due to rounding

detach(news_2014)

#positive article proportion in 2013 using stratified sampling
attach(news_2013)
set.seed(10)
##partitioning data by chanel
channels <- unique(channel)
STR.sample.prop2013 <- NULL
for (i in 1:length(channels)){
  row.indices <- which(channel == channels[i])
  sample.indices <- sample(row.indices, n_h2013_prop[i], replace=FALSE)
  STR.sample.prop2013 <- rbind(STR.sample.prop2013, news_2013[sample.indices,])
}
##get proportion of positive article inside each chanel
###get size for positive and negative artical under each chanel
STR_2013_positive <- STR.sample.prop2013[STR.sample.prop2013$Positive_article == "1",]
```

```r
STR_2013_negative <- STR.sample.prop2013[STR.sample.prop2013$Positive_article == "0",]
STR_n_Pos_2013<-tapply(STR_2013_positive$shares, STR_2013_positive$channel, length)
STR_n_Neg_2013<-tapply(STR_2013_negative$shares, STR_2013_negative$channel, length)
t_STR_n_2013<-STR_n_Pos_2013+STR_n_Neg_2013

###estimated proportion of positive article in 2013
str_2013_prop <- STR_n_Pos_2013/(t_STR_n_2013)
prop_bar_2013 <- sum(str_2013_prop*(N_h2013/N_2013))
prop_bar_2013
```

```
## [1] 0.9131263
```

```r
### calculate SE for the estimate
str_2013_prop_bar_var <- (1-n_h2013_prop/N_h2013)*
  str_2013_prop*(1-str_2013_prop)/(t_STR_n_2013)
prop_bar_2013_se <- sqrt(sum(str_2013_prop_bar_var*
                              (N_h2013/N_2013)^2))
prop_bar_2013_se
```

```
## [1] 0.01037432
```

```r
detach(news_2013)


#positive article proportion in 2014 using stratified sampling
attach(news_2014)
set.seed(10)
##partitioning data by chanel
channels <- unique(channel)
STR.sample.prop2014 <- NULL
for (i in 1:length(channels)){
  row.indices <- which(channel == channels[i])
  sample.indices <- sample(row.indices, n_h2014_prop[i], replace=FALSE)
  STR.sample.prop2014 <- rbind(STR.sample.prop2014, news_2014[sample.indices,])
}
##get proportion of positive article inside each channel
###get size for positive and negative article under each channel
STR_2014_positive <- STR.sample.prop2014[STR.sample.prop2014$Positive_article == "1",]
STR_2014_negative <- STR.sample.prop2014[STR.sample.prop2014$Positive_article == "0",]
STR_n_Pos_2014<-tapply(STR_2014_positive$shares, STR_2014_positive$channel, length)
STR_n_Neg_2014<-tapply(STR_2014_negative$shares, STR_2014_negative$channel, length)
t_STR_n_2014<-STR_n_Pos_2014+STR_n_Neg_2014

###estimated proportion of positive article in 2014
str_2014_prop <- STR_n_Pos_2014/(t_STR_n_2014)
prop_bar_2014 <- sum(str_2014_prop*(N_h2014/N_2014))
prop_bar_2014
```

```
## [1] 0.8284099
```

```r
### calculate SE for the estimate
str_2014_prop_bar_var <- (1-n_h2014_prop/N_h2014)*
  str_2014_prop*(1-str_2014_prop)/(t_STR_n_2014)
prop_bar_2014_se <- sqrt(sum(str_2014_prop_bar_var*
                              (N_h2014/N_2014)^2))
prop_bar_2014_se
```

```
## [1] 0.02093435
```

```r
detach(news_2014)

#Estimate of change in population proportion
delta_STR_prop <- prop_bar_2014 - prop_bar_2013
delta_STR_prop
```

```
## [1] -0.08471643
```

```r
#SE of our estimate
delta_STR_prop_SE <- sqrt((prop_bar_2013_se)^2 + (prop_bar_2014_se)^2)
delta_STR_prop_SE
```

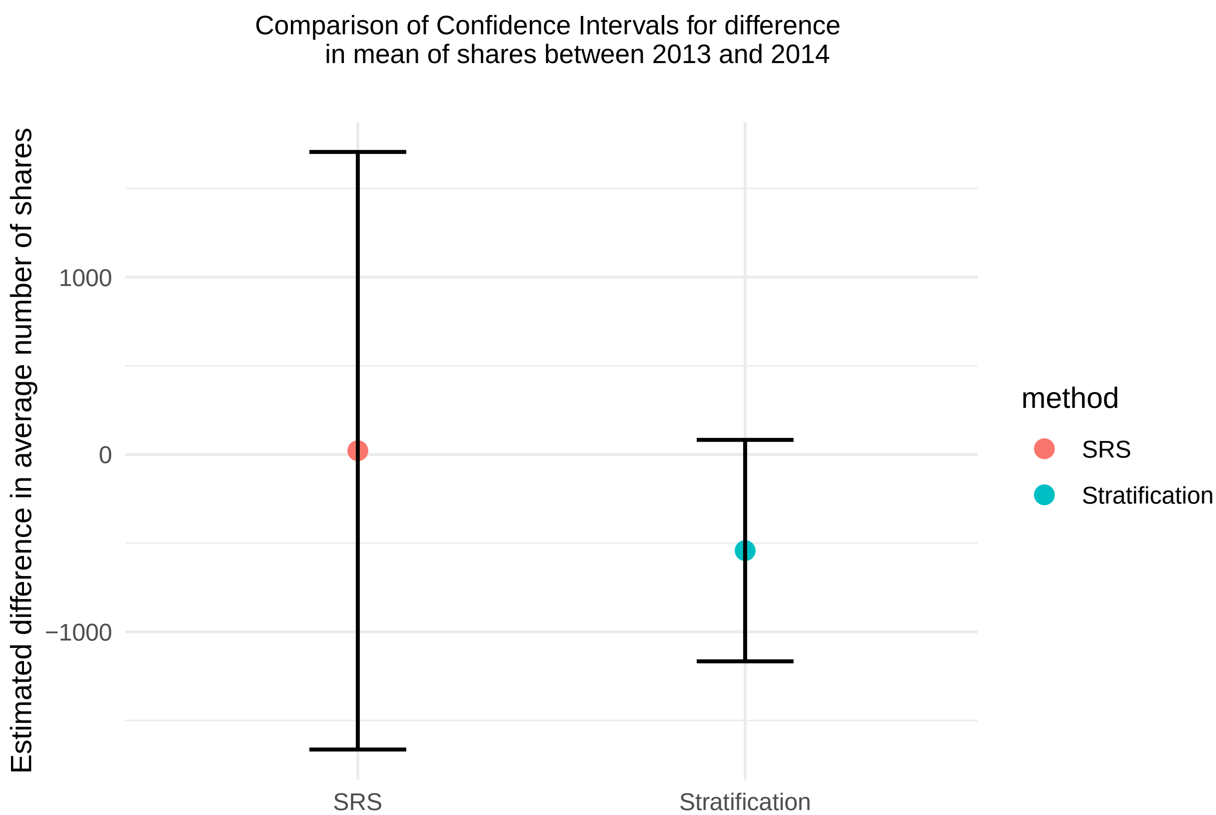
```
## [1] 0.02336394
```

```r
# 95% confidence interval for change of population proportion
str_prop_ci <- data.frame(
  lower_ci = delta_STR_prop - 1.96 * delta_STR_prop_SE,
  upper_ci = delta_STR_prop + 1.96 * delta_STR_prop_SE
)
str_prop_ci
```

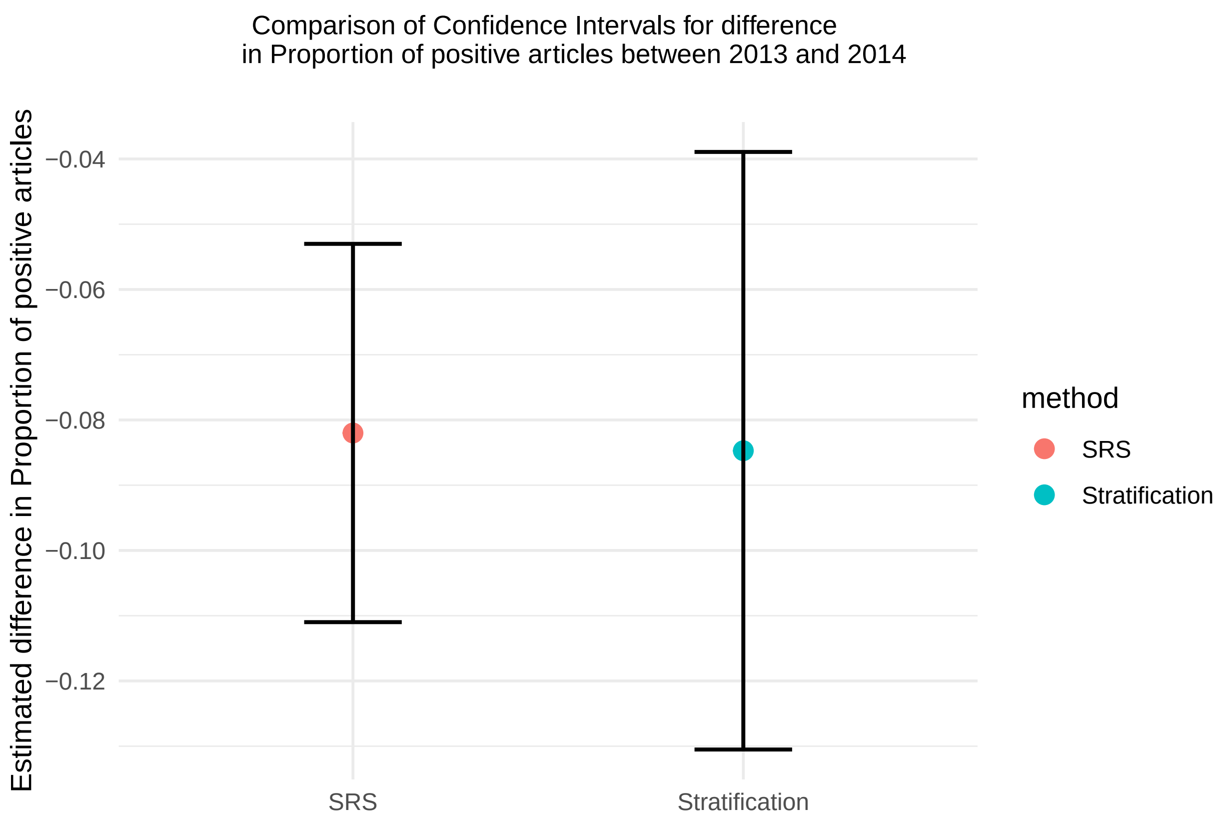
```
##     lower_ci    upper_ci
## 1 -0.1305098 -0.03892311
```

```r
# Since 95% confidence interval excludes 0, we come to the same conclusion that
# there may be a decrease of the proportion of positive articles from
# 2013 to 2014.

library(ggplot2)

lower_srs_mean <- -1663.93
upper_srs_mean <- 1706.064

lower_strat_mean <- -1166.51594
upper_strat_mean <- 82.29943


lower_srs_prop <- -0.1109885
upper_srs_prop <- -0.0530114

lower_strat_prop <- -0.1305098
```

```r
upper_strat_prop <- -0.03892311

# Creating a data frame
diff_mean <- data.frame(
  method = c("SRS", "Stratification"),
  point = c((lower_srs_mean + upper_srs_mean) / 2, (lower_strat_mean + upper_strat_mean) / 2),
  lower = c(lower_srs_mean, lower_strat_mean),
  upper = c(upper_srs_mean, upper_strat_mean)
)

diff_prop <- data.frame(
  method = c("SRS", "Stratification"),
  point = c((lower_srs_prop + upper_srs_prop) / 2, (lower_strat_prop + upper_strat_prop) / 2),
  lower = c(lower_srs_prop, lower_strat_prop),
  upper = c(upper_srs_prop, upper_strat_prop)
)

ggplot(diff_mean, aes(x = method, y =  point, ymin = lower, ymax = upper, color = method)) +
  geom_point(position = position_dodge(width = 0.4), size = 3) +
  geom_errorbar(width = 0.25, position = position_dodge(0.4), aes(ymin = lower, ymax = upper),
                color = "black", size = 0.7) +
  labs(title = "Comparison of Confidence Intervals for difference
       in mean of shares between 2013 and 2014",
       y = "Estimated difference in average number of shares",
       x = "") +
  theme_minimal() +
theme(plot.title = element_text(size = 10, hjust = 0.5, margin = margin(b = 20)))
```




```
ggplot(diff_prop, aes(x = method, y =  point, ymin = lower, ymax = upper, color = method)) +
  geom_point(position = position_dodge(width = 0.4), size = 3) +
  geom_errorbar(width = 0.25, position = position_dodge(0.4), aes(ymin = lower, ymax = upper),
              color = "black", size = 0.7) +
  labs(title = "Comparison of Confidence Intervals for difference
       in Proportion of positive articles between 2013 and 2014",
       y = "Estimated difference in Proportion of positive articles",
       x = "") +
  theme_minimal()+
  theme(plot.title = element_text(size = 10, hjust = 0.5, margin = margin(b = 20)))
```


Comparison of Confidence Intervals for difference
in Proportion of positive articles between 2013 and 2014
Estimated difference in Proportion of positive articles
−0.04
−0.06
−0.08
−0.10
−0.12
SRS
Stratification
method
SRS
Stratification