BIO310 Assignment

Madheshvaran S

29/01/2021

Estimation of Journal Impact Factor for Journal of Cell Biology in the year 2013

Introduction:

Journal Impact Factor (JIF) is a metric devised by Eugene Garfield, the founder of the Institute for Scientific Information in Philadelphia, USA. It is used as one of the metrics which quantifies the impact of research articles published in a journal. According to Wikipedia, it is defined as "average number of citations that articles published in the last two years in a given journal received" for that specific year.

In this assignment, the journal impact factor is calculated for the Journal of Cell Biology (JCB) in the year 2013. This is done through generating a random sample dataset from the population of articles published in 2011 and 2012 and analyzing the data to obtain the journal impact factor for JCB in 2013. From the obtained data, we plot the distribution of the citations received in 2013 and calculate the confidence intervals for the JIF obtained. The assignment also examines the difference between research articles and reviews in terms of the number of citations they received in 2013.

The raw data, R code and the results of this analysis are provided in the following GitHub repository. The results are reproducible and this can be done using the raw data and R code provided. The link to the GitHub repository is:
https://github.com/madheshvaran/BIO310

1. Generation of the Sample:

Question: What is the sampling method that you will use to choose the 26 articles?

(a) Describe the method

Before generating the sample dataset used for the analysis, I performed a set of preliminary analysis to ensure that the sample selected is random and the sample represents the population (i.e., the entire set of articles published in 2011 and 2012).

To start the preliminary analysis, I manually created a csv file containing the details of the issues (Volume Number, Issue Number, Date, Number of Research Articles, Number of Review Articles) published in JCB in the year 2011 and 2012. The csv file can be accessed through the GitHub repository.

To ***ensure that the sample (to be used in the analysis) represents the population of all articles published in 2011 and 2012***, I randomly selected 10 issues from the list and estimated the mean proportion of research and review articles in that sample. I repeated this process for five times and took the mean of mean proportion of research aticles estimated from the five samples. The code used for it is given below. Note that I used *set.seed()* to ensure the reproducibility of the code.

```
setwd("C:/Users/Madheshvaran/Desktop/Biostatistics")

# Library
library(knitr)
library(ggplot2)

prelim_data <- read.csv("raw data/List of Volumes and Issues for year 2011
and 2012.csv")
prop_list <- c()
for (i in 1:5) {
  set.seed(2021+i)

  sno <- sample(prelim_data$S_No, 10)
  sdata <- prelim_data[sno,]
  total_article <- sum(sdata$Research_Articles)
  total_review <- sum(sdata$Review_Articles)

  prop_article <- total_article/(total_article+total_review)
  prop_list <- c(prop_list, prop_article)
}
mean(prop_list)

## [1] 0.8393
```

From the results, I estimated that the proportion of research articles in the population is 0.8393. Now, since we are selecting a total of 13 articles in each year for the sample dataset, so the number of research articles to be chosen are **11** (rounding off the product of proportion of research articles estimated and total number of articles, i.e., 13 articles). Therefore, to make acertain that the sample represents the population, we need to choose **11** research articles and **2** review articles in each year to get a total of 26 articles.

Now, to ***ensure that the sample is chosen randomly***, the sampling method is described below. Since it is difficult to get data about the entire population (which can be done manually entering the data or by using web scrapping tools), I wrote a code that randomly select 13 issues out of 26 issues pubished in each year. From the selected issues, the code randomly picks an article (each article in an issue has a index which can be uniquely mapped to it). This code randomly selects 11 research articles and 2 review articles for each year.

```r
### Sampling Method  ###

# For 2011
sdata <- prelim_data[which(prelim_data$Year == "2011"), ]

# Selection of Research Articles
set.seed(29)
article11 <- sample(sdata$S_No, 11)
list_article_11 <- c()
for (i in article11) {
  set.seed(1000+i)
  x <- sample(1:sdata$Research_Articles[which(sdata$S_No == i)],1)
  list_article_11 <- c(list_article_11, x)
}
temp <- data.frame(S_No = article11, Article_Type = rep("Research Article"),
                   Article_Chosen = list_article_11)
final_article_11 <- merge(sdata, temp, by = "S_No")

# Selection of Reviews
set.seed(49)
review11 <- sample(sdata$S_No, 2)
list_review_11 <- c()
for (i in review11) {
  set.seed(500+i)
  x <- sample(1:sdata$Review_Articles[which(sdata$S_No == i)],1)
  list_review_11 <- c(list_review_11, x)
}
temp <- data.frame(S_No = review11, Article_Type = rep("Review Article"),
                   Article_Chosen = list_review_11)
final_review_11 <- merge(sdata, temp, by = "S_No")
final_selection_11 <- rbind(final_article_11, final_review_11)[,-c(7:8)]


# For 2012
sdata <- prelim_data[which(prelim_data$Year == "2012"), ]

# Selection of Research Articles
set.seed(31)
article12 <- sample(sdata$S_No, 11)
list_article_12 <- c()
for (i in article12) {
  set.seed(10+i)
  x <- sample(1:sdata$Research_Articles[which(sdata$S_No == i)],1)
  list_article_12 <- c(list_article_12, x)
}
temp <- data.frame(S_No = article12, Article_Type = rep("Research Article"),
                   Article_Chosen = list_article_12)
final_article_12 <- merge(sdata, temp, by = "S_No")
```

```
# Selection of Reviews
set.seed(51)
review12 <- sample(sdata$S_No, 2)
list_review_12 <- c()
for (i in review12) {
  set.seed(70+i)
  x <- sample(1:sdata$Review_Articles[which(sdata$S_No == i)],1)
  list_review_12 <- c(list_review_12, x)
}
temp <- data.frame(S_No = review12, Article_Type = rep("Review Article"),
                   Article_Chosen = list_review_12)
final_review_12 <- merge(sdata, temp, by = "S_No")
final_selection_12 <- rbind(final_article_12, final_review_12)[,-c(7:8)]

final_selection <- rbind(final_selection_11, final_selection_12)
```

The variable *final_selection* has the list of all issues and the corrosponding articles to be selected for the sample dataset. Using this instructions, I manually created a csv file that contains the selected articles name, type of article and number of citations received in the year of 2013. The first five entries of *final_selection* is displayed below. For the entire dataset, it is provided as a csv file in the results folder of the GitHub Repository.

```
kable(final_selection[1:5, ], format = "markdown")
```

S_No	Volume_Number	Issue_Number	Date	Month	Year	Article_Type	Article_Chosen
1	192	1	10	January	2011	Research Article	7
2	192	2	24	January	2011	Research Article	4
3	192	3	7	February	2011	Research Article	9
5	192	5	7	March	2011	Research Article	2
7	193	1	4	April	2011	Research Article	3

(b) Provide a table with details of the articles chosen (name of article, type of article and year) and how many citations they got in the year being considered

The following table shows the list of all randomly selected articles along with further information:

```
data <- read.csv("raw data/Sample Dataset.csv")
kable(data, format = "markdown")
```

Name	Type	Year	Citations
Correlated fluorescence and 3D electron microscopy with high sensitivity and spatial precision	Research Article	2011	32
IGF-II is regulated by microRNA-125b in skeletal myogenesis	Research Article	2011	31
The serine/threonine kinase Par1b regulates epithelial lumen polarity via IRSp53-mediated cell–ECM signaling	Research Article	2011	9
Defining the earliest step of cardiovascular progenitor specification during embryonic stem cell differentiation	Research Article	2011	12
System analysis shows distinct mechanisms and common principles of nuclear envelope protein dynamics	Research Article	2011	7
SLAIN2 links microtubule plus end–tracking proteins and controls microtubule growth in interphase	Research Article	2011	5
Tinman/Nkx2-5 acts via miR-1 and upstream of Cdc42 to regulate heart function across species	Research Article	2011	7
Spatial code recognition in neuronal RNA targeting: Role of RNA–hnRNP A2 interactions	Research Article	2011	8
Caspase-8 inactivation in T cells increases necroptosis and suppresses autoimmunity in Bim?/? mice	Research Article	2011	2
Cryoelectron tomography of radial spokes in cilia and flagella	Research Article	2011	14
Synergy between the ESCRT-III complex and Deltex defines a ligand-independent Notch signal	Research Article	2011	7
On emerging nuclear order	Review Article	2011	20
TOR kinase complexes and cell migration	Review Article	2011	8
Structural specializations of ?4?7, an integrin that mediates rolling adhesion	Research Article	2012	11
A neuropeptide signaling pathway regulates synaptic growth in Drosophila	Research Article	2012	4
Human telomeres replicate using chromosome-specific, rather than universal, replication programs	Research Article	2012	3
Trichoplein and Aurora A block aberrant primary cilia assembly in proliferating cells	Research Article	2012	9
Stoichiometry of Nck-dependent actin polymerization in living cells	Research Article	2012	9
The IFT-A complex regulates Shh signaling through cilia structure and membrane protein trafficking	Research Article	2012	15
Follistatin-mediated skeletal muscle hypertrophy is	Research	2012	13

regulated by Smad3 and mTOR independently of myostatin	Article		
Bnip3 and AIF cooperate to induce apoptosis and cavitation during epithelial morphogenesis	Research Article	2012	1
Clathrin promotes centrosome integrity in early mitosis through stabilization of centrosomal ch-TOG	Research Article	2012	5
Critical role for the kinesin KIF3A in the HIV life cycle in primary human macrophages	Research Article	2012	12
Mutual antagonism between IP3RII and miRNA-133a regulates calcium signals and cardiac hypertrophy	Research Article	2012	6
Breaking the ties that bind: New advances in centrosome biology	Review Article	2012	15
The cellular and molecular basis for malaria parasite invasion of the human red blood cell	Review Article	2012	17

2. Distribution of the Number of Citations:

Question: Plot histograms to show the distribution of number of citations for the 26 articles.

The following histogram shows the distribution of number of citations for the 26 articles sorted by the year of publication:

```
data$Type <- factor(data$Type)
data$Year <- factor(data$Year)
data$Type_and_Year <- paste(data$Year, " - ", data$Type, sep = "")

ggplot(data = data, aes(x = Citations, fill = Year))+
  geom_histogram(binwidth = 2, color = "black")+
  scale_x_continuous(breaks = seq(0, 34, by = 2))+
  labs(title = "Distribution of Number of Citations By Year of Publication",
       x = "Number of Citations", y = "Frequency",
       fill = "Legend Box")+
  scale_fill_manual(values = c("olivedrab1", "steelblue2"))+
  theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour =
"black"))
```

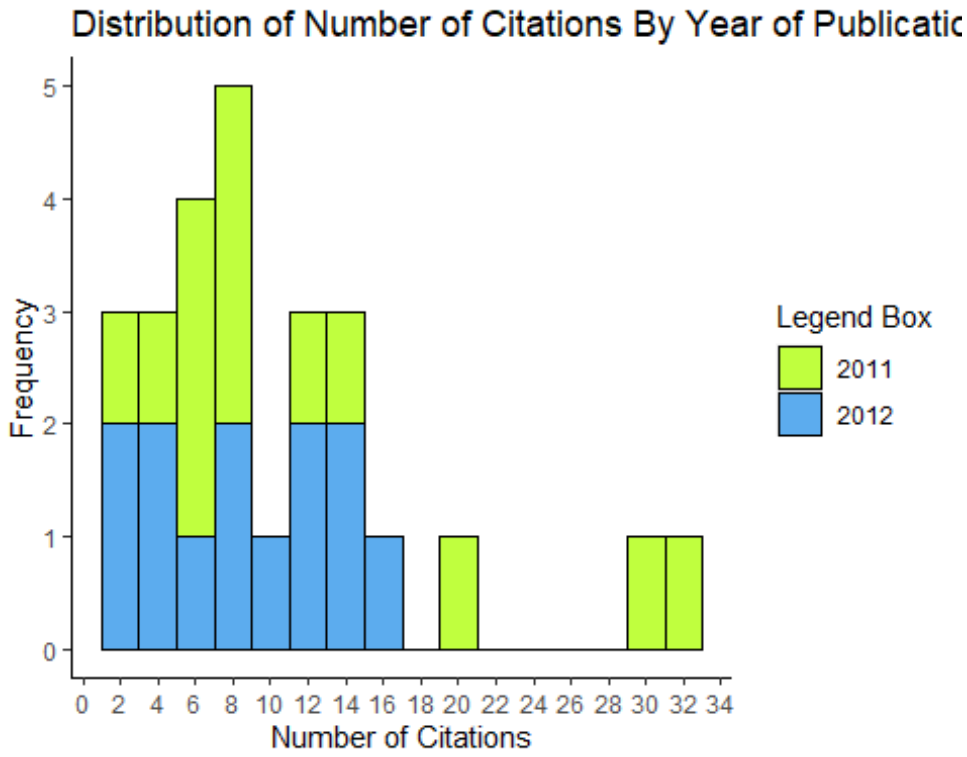

Figure 1: The distribution of number of citations for the 26 articles *by the year of publication*.

The following histogram shows the distribution of number of citations for the 26 articles sorted by the type of article:

```
ggplot(data = data, aes(x = Citations, fill = Type))+
  geom_histogram(binwidth = 2, color = "black")+
  scale_x_continuous(breaks = seq(0, 34, by = 2))+
  labs(title = "Distribution of Number of Citations By Type of Article",
       x = "Number of Citations", y = "Frequency",
       fill = "Legend Box")+
  scale_fill_manual(values = c("grey60", "grey20"))+
  theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour =
"black"))
```

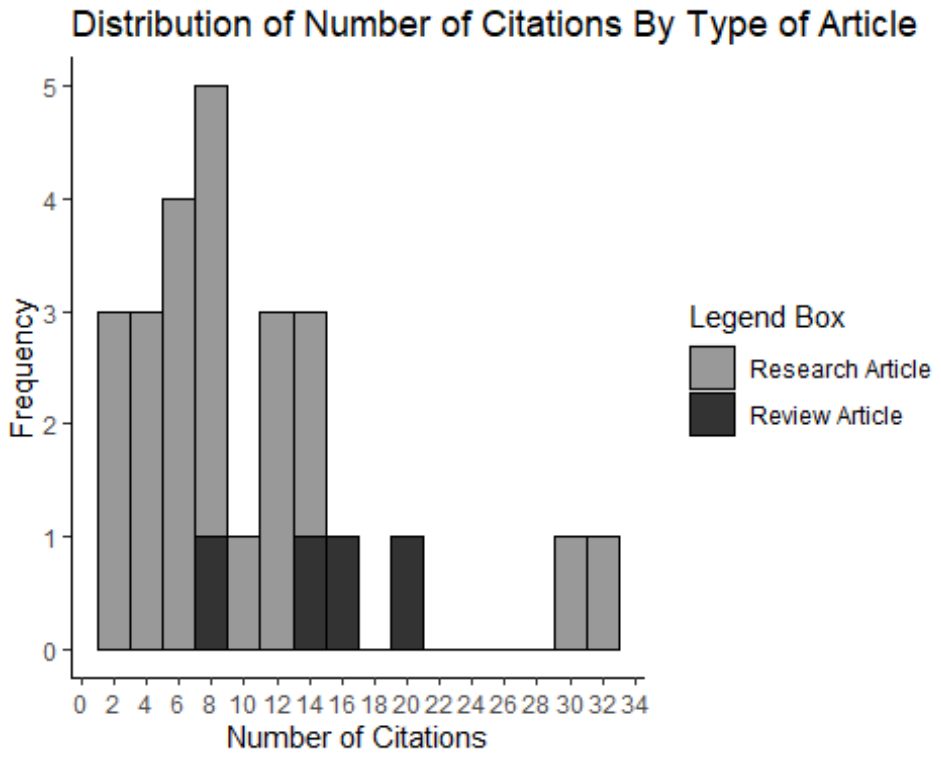

Figure 2: The distribution of number of citations for the 26 articles *by the type of article*.

The following histogram shows the distribution of number of citations for the 26 articles sorted both by the year of publication and the type of article:

```
ggplot(data = data, aes(x = Citations, fill = Type_and_Year))+
  geom_histogram(binwidth = 2, color = "black")+
  scale_x_continuous(breaks = seq(0, 34, by = 2))+
  labs(title = "Distribution of Number of Citations By Type of Article and
Year of Publication",
       x = "Number of Citations", y = "Frequency",
       fill = "Legend Box")+
  scale_fill_manual(values = c("olivedrab1", "olivedrab4", "steelblue1",
"steelblue4"))+
  theme(panel.grid.major = element_blank(), panel.grid.minor =
element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour =
"black"))
```

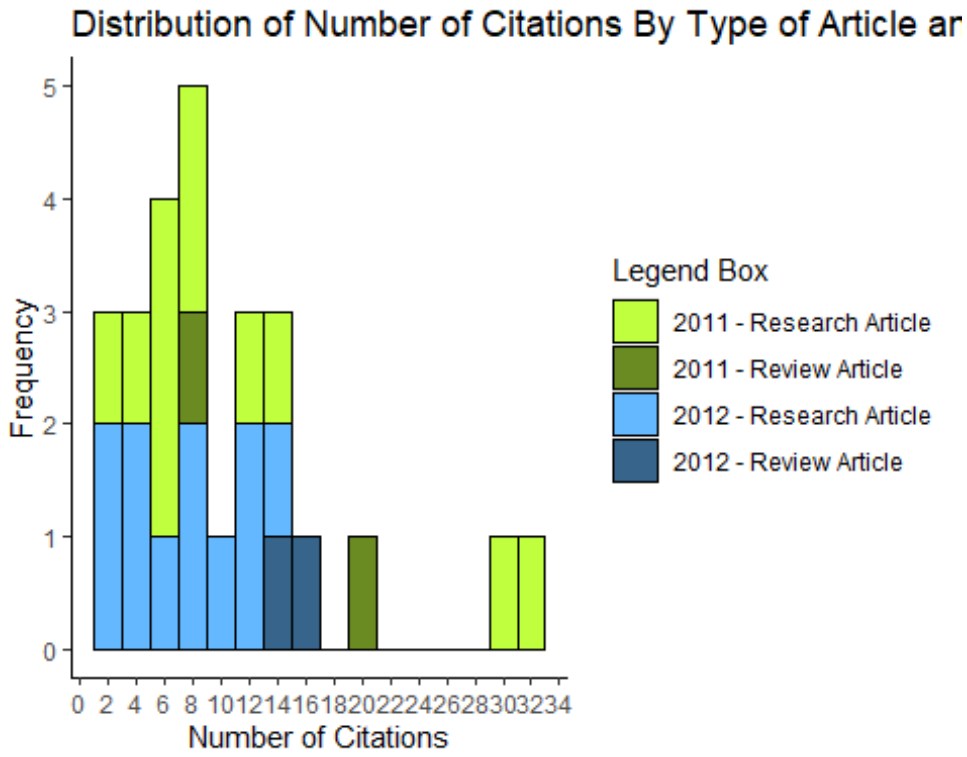

Figure 3: The distribution of number of citations for the 26 articles both *by the year of publication and the type of article*.

3. Calculation of Mean and Median:

Question: Calculate the mean and median for the number of citations. Are these two the same? If not, comment on why they are different.

The mean and median of the number of citations is calculated below:

```
jif <- mean(data$Citations)
med <- median(data$Citations)
```

From the calculation, the mean of number of citations is 10.8461538 and median of number of citations is 9. The mean and median are **not same**. This is because the mean is affected by extreme values (as some articles tend to get very high number of citations) but median is not affected by extreme values.

Assuming I have a paper in this journal in 2020, how many citations would I expect to get in 2021? Which measure should I consider from above as a measure of central tendency and why?

Since the assignment calculates the JIF for JCB in 2013, I can't comment about a paper published in 2020. But if I modify the question, i.e., assuming I have a paper in JCB in 2013, the paper is most likely to get cited **9** times in 2014. We need to use median in this case because there are some exceptional articles published in JCB which gets cited very high

(due to their relevance or popularity). So, the distribution is positively skewed and hence, mean is always over-estimated. However, the median tends to show the central tendency of a typical population. Therefore, it is recommended to use median in this case.

4. Difference between Research and Review Articles in terms of number of citations:

Question: Do reviews and research articles get significantly different number of citations? Explain using the data and an appropriate statistical test.

The hypothesis in this case are as follows:
Null Hypothesis (Ho): The mean of number of citations of research and review articles are same.
Alternative Hypothesis (Ha): The mean of number of citations of research and review articles are not same.

Since the sample size of the review articles is small (4 review articles in total of 26 articles), we can't ensure the normality of the review articles. Also due to its small size, we cannot be sure about the assumption of same variance of two groups. So, a **two-way t test** becomes less powerful to test the hypothesis. However, we can perform a non-parametric test called **Wilcox Rank Sum (or Mann-Whitney) Test** which does not assume the normality of the groups. The p-values calculated from these two tests are as follows:

```
## Two-way T test
research_article_group <- data[which(data$Type == "Research Article"),
"Citations"]
review_article_group <- data[which(data$Type == "Review Article"),
"Citations"]
t.test(research_article_group, review_article_group, alternative =
"two.sided")$p.value

## [1] 0.1591214

## Wilcox Rank Sum Test
wilcox.test(research_article_group, review_article_group, alternative =
"two.sided")$p.value

## [1] 0.06938036
```

In both cases, the p-value calculated in more than the significance value (0.05). So, we **CANNOT** reject the null hypothesis. However, we need to increase the sample size of review articles to reconfirm the statistical test.

5. Calculation of 95% Confidence Intervals:

Question: List two ways (no need to explain) in which you can calculate 95% Confidence Intervals (CIs) for the mean JIF from #3 (above). Using any one of the methods, calculate the 95% CIs for mean JIF.

We can calculate the 95% Confidence Intervals by two ways:
(i) Calculation of Confidence Intervals by using Student's T distribution (or converting it to a standard normal distribution (or Z-distribution))
(ii) Calculation of Confidence Intervals with Bootstrapping method

In this case, we can use the first method to calculate the 95% Confidence Interval. The following code calculates the 95% confidence intervals for the estimated JIF. Note that *qt(0.975, df)* calculates the t(alpha = 0.05, df) value.

```
error <- qt(0.975, df = length(data$Citations) -
1)*(sd(data$Citations)/sqrt(length(data$Citations)))
left <- mean(data$Citations) - error
right <- mean(data$Citations) + error
```

From the above calcuation, the 95% Confidence Interval are **(7.7453713, 13.9469364)**.

Now can you give me another answer for #3a (above)?

With the 95% Confidence Interval, we can say that the parameter (true value) lies within the calculated confidence interval 95 times out of 100 if we repeat this process 100 times. Then it is most likely that the paper published in 2013 can have a number of citation in 2014 within this interval but if this sample collected falls in the remaining 5 out of 100, then the parameter lies outside the interval. So, most likely the number of citations received in 2014 is within (7.7453713, 13.9469364) but one can't say with full confidence.

6. Comaparasion to the Real JIF:

Question: As mentioned at the top, real JIFs for each journal are calculated by Clarivate. Can you use one of the methods listed in Q5 to calculate 95% CIs for these JIFs?

According to the Journal Citation Reports 2013 published by Clarviate, the Journal Impact Factor of the Journal of Cell Biology is ***9.786*** for the year 2013. This value falls in the 95% Confidence Interval we calculated above.

No. We cannot use any of the above methods to calculate the 95% Confidence Intervals for the real JIF. This is because the real JIF calculated by Clarivate is done for the whole population. So, it is a **parameter**(true value of the population). The Confidence Intervals is calculated for an estimate not for a parameter. It doesn't make sense to do so.