Analysing the Methods to Optimize Online Advertising Campaign

Yiyi Tan - 1006872706

April 4, 2021

Abstract

When it comes to online advertising, Facebook becomes one of the effective marketing platforms over the last few years. As Facebook clients associate with the platform, adding comments, liking posts, and providing demographics information, Facebook constructs a profile of that client dependent on what their identity is and what they're keen on. This study aims to determine how Facebook campaign can be better targeted in order to increase conversions (conversion occurs when a visitor to the website completes the desired goal). Specifically, it investigates ways to improve conversion rates with the right target group.

Data in this report were collected from Kaggle. This data set represents an Anonymous organization's social media ad campaign. We will clean the data and summarize it within the numerical and graphical procedure. After observing the data, we want to test the hypothesis that women and 30-40 age range lead to a higher conversion rate. Data were analyzed using Bayesian Credible Interval and Hypothesis Test. Additionally, the Maximum Likelihood Estimator examined the optimal mean of conversion rate and we calculated the confidence interval for mean click-through rate. The results support that females will in general make the purchase more in the wake of tapping the ads than men and the population means of the 30-34 age range's conversion rate is 13% (higher than the mean of conversion rate 10.9%). Moreover, it showed a low mean for click-through rate with a high mean for conversion rate.

These results suggest that ideal target demographics are women and age group between 30 to 34. On this basis, these two groups should be taken into account when designing Facebook campaigns.

Introduction

Facebook is one of the most attractive online advertising platforms for different kinds of advertisers. The goal of creating an online campaign varies from one firm to another, but Facebook has three objective categories: Awareness, Consideration, and Conversion. For this project, we will assume the company is interested in the conversion of the campaign and help them to maximize the amount of revenue by improving its ads on Facebook. The data set used in this project is from an anonymous organization's Facebook ad campaign and generated from Kaggle https://www.kaggle.com/loveall/clicks-conversion-tracking. The major variables that will be used in this project are 'age', 'gender', 'Impressions', 'Clicks' and 'Approved_Conversion'. Besides, there are going to be three new variables be created include Click Through Rate (clicks/impressions), Conversion Rate (total_conversion/clicks) and Cost Per Click (spent/clicks) to assess the performance of this marketing campaign.

The main goal of this project is to help this anonymous organization (xyz company) make the best possible performance of its Facebook campaign and give them several suggestions on future campaign strategies. In driving the growth of the business, optimize the impact of Facebook ads is one of the important elements in the marketing campaign. This analysis will be separated into two aspects of improving Facebook campaign. On one hand, conversion rates are especially significant when running Facebook ads of the fact that they can gauge the accomplishment of each campaign. On the other hand, with 2.74 billion monthly active users, targeting the right group with the correct message is key to creating a successful social marketing strategy (McLachlan, 2021).

We will focus on building up the conversion rate with the right target group by analyzing the data set. As a result, the study will address two research questions:

1. How to optimize this social advertising campaign to get a higher conversion rate?
2. What are the perfect target group in this campaign?

To address these two research questions, 6 main methodologies will be utilized in this project include Goodness of Fit Test, Simple Linear Regression, Bayesian Credible Interval, Hypothesis Test, Maximum Likelihood Estimator, and Confidence Interval. The first two methodologies (Goodness of Fit Test and Simple Linear Regression) pays attention to clicks and impressions. The Goodness of Fit Test will test the distribution of clicks and Simple Linear Regression will perform whether there exists a linear relationship between clicks and impressions. The next two methodologies (Bayesian Credible Interval and Hypothesis Test) analyze the demographics group xyz company is supposed to aim at so that getting more conversions. Bayesian Credible Interval focuses on the proportion of women who made the approved conversion. In addition, the Hypothesis Test looks into whether 30 to 34 age range will make a higher mean conversion rate than the mean conversion rate. The last two methodologies (MLE and Confidence Interval) examine the mean of conversion rate and click-through rate.

Data

i. Data Collection

Data for this study are a subset drawn from an Advertisement Conversions data set generated from Kaggle https://www.kaggle.com/loveall/clicks-conversion-tracking. This data set represents an Anonymous organization's (denote as xyz company) social media ad campaign which contains 1143 observations in 11 variables from the original file 'KAG_conversion_data.csv'. Description of the important variables is in the following Variable Description section.

ii. Data Cleaning

In this report, we compiled and cleaned the exported data from Kaggle to ensure data is correct, consistent, and usable. To begin with, in order to assess the performance of this marketing campaign, there are some missing crucial variables include Click Through Rate (clicks/impressions) which describes the percentage of how many of the impressions became clicks, Conversion Rate (total_conversion/clicks) which indicate the proportion of people exposed to the purpose of these ads (ie. bought the products) who clicked on it (ie. link clicks), and Cost Per Click (spent/clicks) calculate how much did each click cost. With the variables we have in the data set, we can easily create the CTR and CVR figures using the mutate function. Besides, removing missing values and select the columns that are needed for later analysis.

iii. Variable Description

Below table are the 7 variables that will be used in the later analyzing process with its detailed description. The first two rows of this table are the categorical variables age and gender. The rest of the rows are the numerical variables.

Variable	Description
age	Age of the person to whom the ad is shown
gender	Gender of the person to whom the ad is shown
Impressions	Number of times the ad was shown
Clicks	number of clicks on for that ad.
Approved_Conversion	Total number of people who bought the product
CTR	Percentage of how many of the impressions became clicks
CVR	Proportion of people exposed to the purpose of this ads over who clicked on it
CPC	How much each click cost

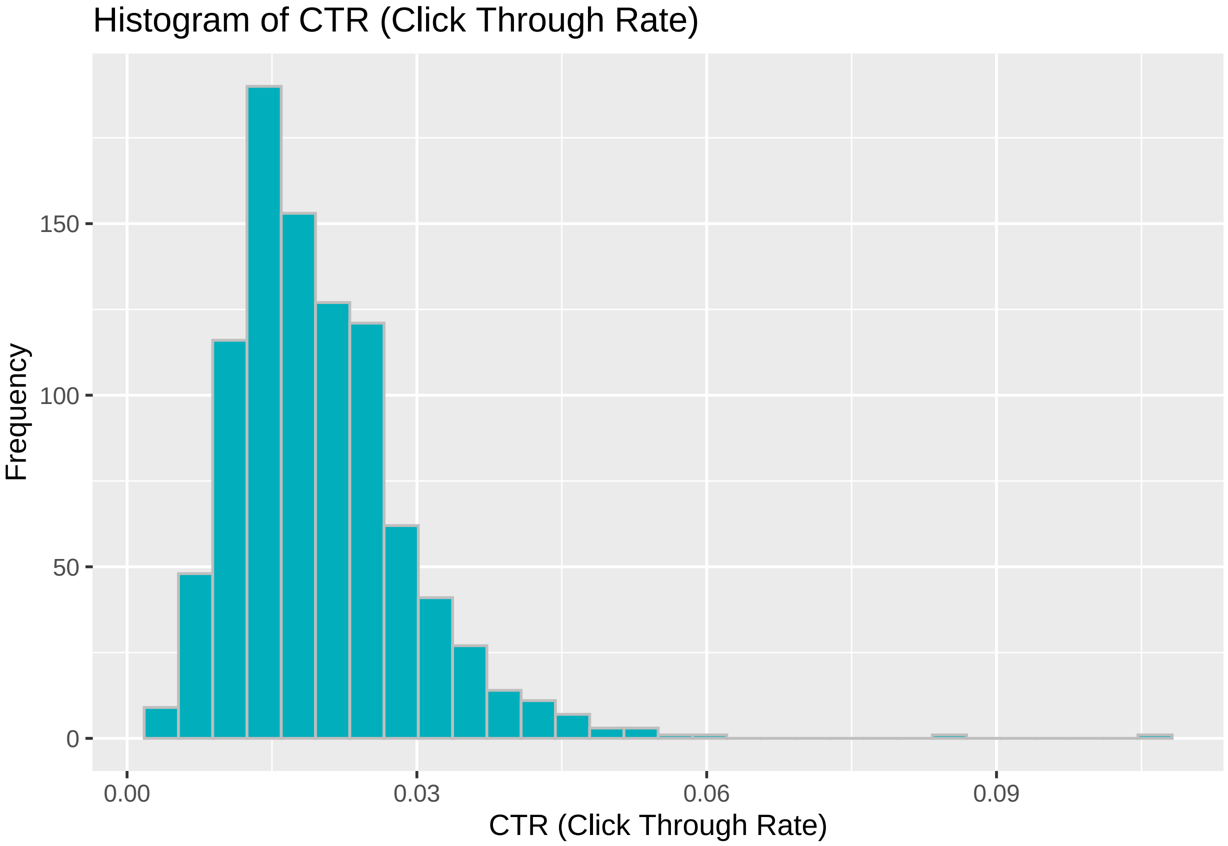
iv. Data Summaries

(a) Numerical Summaries

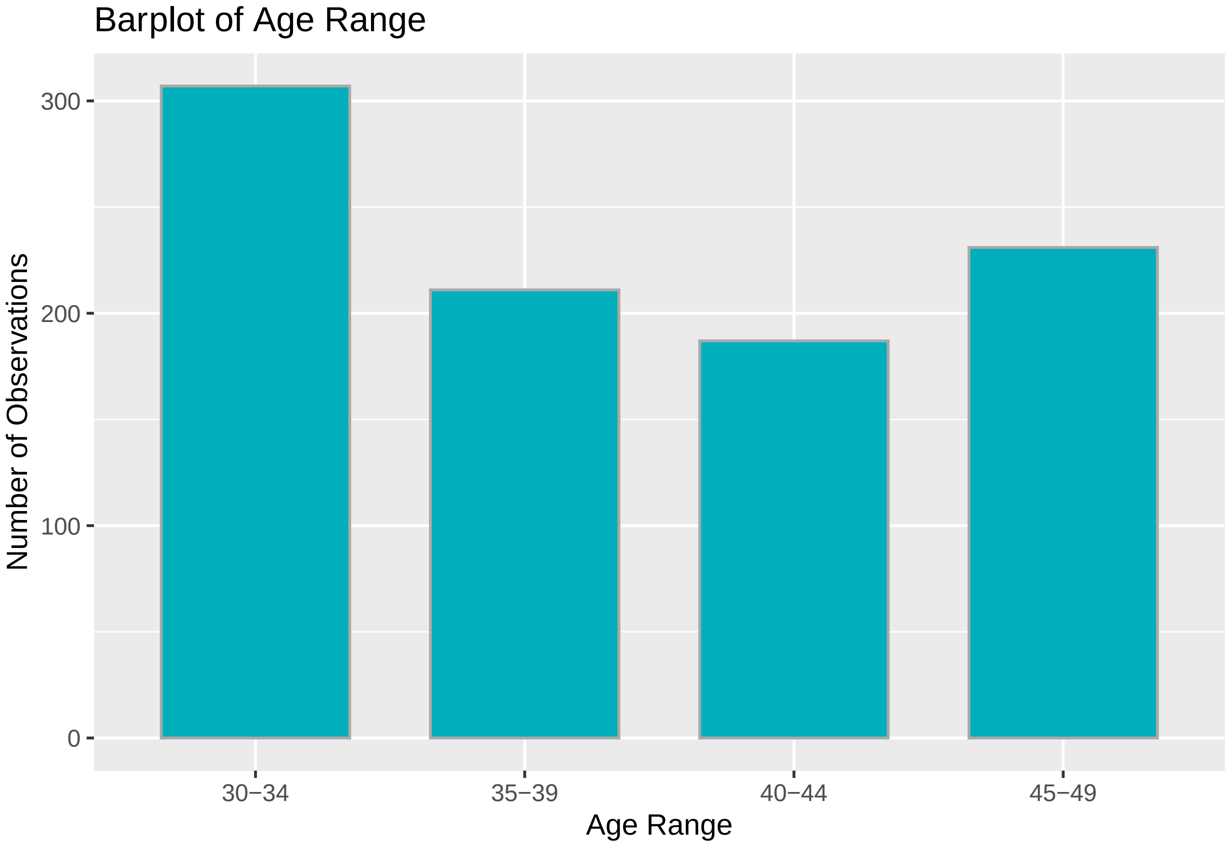
Variable	Min	Mean	Median	Max
Impressions	944	227506	98336	3052003
Clicks	1.00	40.77	16.00	421.00
Approved_conversion	0.00	1.072	1.00	21.00
CTR (percentage)	0.00307%	0.0201%	0.0182%	0.106%
CVR (percentage)	0.0%	10.903%	1.254%	200%
CPC (dollar)	0.180	1.499	1.498	2.212

In the above table, the mean click-through rate is 0.0201% which is much lower then the average CVR on Facebook Ads (0.89%) according to Irvine. Although the mean click-through rate for xyz company's campaign is not that good, the mean conversion rate performs better than the average on Facebook Ads (9.11%) across all industries. The average cost per click (CPC) on Facebook ads is $1.68 across all industries compared to $1.499 of CPC for xyz's campaign which seems fine but not enough to optimize the profit. Therefore, we will analyze the reasons for having a poor click-through rate and help xyz company to improve their CTR.

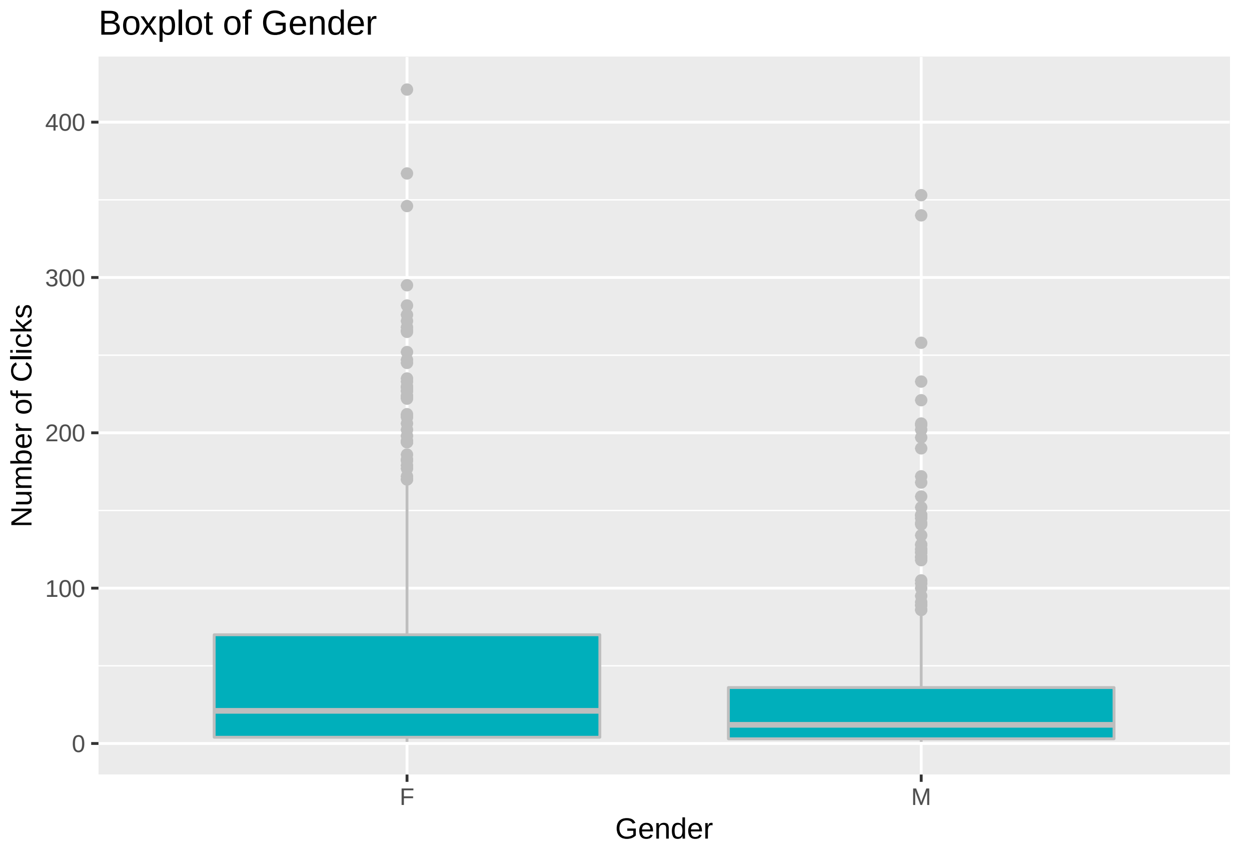
(b) Graphical Summaries



In order to make an assumption for the Maximum Likelihood Estimator we are interested in viewing the shape of the CTR (Click Through Rate)'s distribution. Variable Click Through Rate is a numerical data we created then we are eligible to plot a histogram of CTR. The X-axis is intervals that show the scale of CTR which the measurements fall under. Y-axis shows the number of times that the values occurred within the CTR. The width is the same across all bins. From this graph, we can observe that shape of the graph is right-skewed (positive skewness) and is unimodal which means there is only a single highest value. In addition, the center of the graph is about 0.017% which means most of the ads have 0.017% CTR, and the spread of CTR range between 0.00307% to 200% (from numerical summaries). After observing the histogram, CTR is the most range between 0.005% to 0.03%. Noticeably, this graph tends to be a bell curve indicates that the data (CTR) follows the normal distribution. Thus in the Method section, we will estimate the true mean of CTR by the methodology called Confidence Interval (Z approach) by assuming CTR follows Normal distribution.



This bar graph compares different categories contains 30-34, 35-39, 40-44, and 45-49 age ranges. The most common category is the age range between 30 to 34 and the least common category is the age range between 40 to 44. 1.72 times is more common 30-40 age range than 40-44 age range. At this point, we know 30-34 years have higher observations than other age range. Hypothetically, the mean of 30-34 conversion rate is higher than the mean total conversion rate. Detailed will be analyzed in Hypothesis Test.



This is a boxplot of number of clicks for two categories (F: female and M: male). From the boxplot, we can easily see that both of the categories are right skew and there are approximately 20 outliers. Therefore, there are about 20 observations whose number of clicks is higher significantly than others in both categories. The spread of the number of clicks for the female is about 60 and the spread of the second boxplot is 40. The first box plot is much higher than another which suggests a difference between these two groups. The median number of clicks seems to be different for both categories. In the first boxplot, the center is at around 24 but for the second one, the center is at 15. In conclusion, the mean of the number of clicks for the female is higher than male and is more variable than male.

Methods

1. Goodness of Fit Test

Clicks can seem like the most important step in the online advertising process because a click represents an interested visitor. In this section, we will introduce a methodology called the goodness of fit test to help xyz company have a better understanding of clicks in the campaign. Assume the clicks we count will be a random sample from the population of all clicks.

For this methodology, we will first make a null hypothesis that the number of clicks (X) follow Poisson distribution with parameter $\lambda = 41$ i.e.$H_0 : X \sim Poi(\lambda = 41)$. Then, the alternative hypothesis is that the number of clicks (X) does not follow Poisson distribution with parameter $\lambda = 41$ i.e.$H_\alpha : X$ not follow $Poi(\lambda = 41)$. Next, we will use the data to find a test statistic and p-value (the result will be shown in the Results section). Besides, using the Chi-squared test and calculate p-value by R. Let $\alpha = 6$ as a cut-off for rejecting or not rejecting H_0. If p-value $< \alpha$ then we will reject H_0. In contradiction, if p-value $> \alpha$ then we won't reject H_0.

2. Simple Linear Regression

According to the variable description in the Data section, we informed that clicks are the number of clicks on for that ad in the campaign. Approved conversion is the total number of people who complete a goal established on the website after seeing the ad. By the definition of these two variables, it seems like there is a relationship between them. Linear regression attempts allow us to model the relationship between these two variables by fitting a linear equation to observed data.

For this model, we assumed approved conversion and clicks have a linear relationship ($Y_i = \alpha + \beta x_i + U_i$), U_i has equal spread ($E(U_i) = 0$) and constant variance ($Var(U_i) = \sigma^2$).

Variable	Description
Y_i	Approved Conversion
x_i	Clicks
α	intercept of the regression line
β	slope of the regression line
U_i	error of the equation

3. Bayesian Credible Interval

Only knowing the distribution of clicks is certainly not enough for the advertisers or firms to improve the performance of their campaign. A decent pointer of how the ads are performing is computing the conversions. According to the data section, we are informed that there are more clicks made by female than male does which raise a question that is the conversions and clicks have the same pattern?

Hence, we want to check if the proportion of women who buy the product after clicking the ads will be greater than men. We will find the Bayesian Credible Interval here which is an interval that has a specified probability of containing the parameter, given the observed data. Let X be the number of conversions made by female (number of products women bought). By the fact that, the conversions are either made by female or made by male in our data set which leads to be a binary response. Then suppose our data is a random sample of Binomial random variables with probability p (proportion of women who bought the products), total number of conversions n (total products bought by all people), with the prior distribution of p is assumed to be $Beta(5, 8)$ in hopes of yielding a neutral/non-informative prior. We are interested in finding a 95% credible interval of the parameter p. Then we have posterior distribution (in Result section) by derivation. Thus, we can use the 2.5th and 97.5th percentiles of this distribution to derive a range of values which p has 95% probability of falling into. All derivations regarding the posterior distribution can be found in Section 1 of the Appendix.

4. Hypothesis Test

Based on the observations of the barplot in Data section, we noticed that the age range of 30-34 years has higher observations than other age range. Besides, we calculated the total impressions the campaign made for different age range in the table below.

Age Range	Sum of Impressions
30-34	67674431
35-39	42021238
40-44	39551158
45-49	63699120

Age range of 30-34 has the highest total impressions among all 4 age ranges. Therefore, we are interested in whether mean of 30-34 conversion rate is higher than the mean of conversion rate (10.9% in Maximum Likelihood Estimator). Assume that the population mean of conversion rate between 30-34 age range is 13%. Hypothesis test will assess the rationality of the hypothesis by using sample data.

We have formulated a null and alternative hypothesis by assuming the population mean of conversion rate between 30-34 age range is 13% i.e. $H_0 : \mu = 13$. Then, the alternative hypothesis is that the population mean of conversion rate between 30-34 age range is not 13% i.e.$H_\alpha : \mu \neq 13$. Additionally, we will use data that is random generalized to find a test statistic and p-value. Assume that the distribution of the (data and thus the) sample mean is normally distributed. Then we know test statistic $= \frac{(\bar{X}-\mu)}{s/\sqrt{n}} \sim t_{n-1}$ which implies that p-value: $p = P(|t_{n-1}| > |t|)$ ($\bar{X}$: Sample Mean, s: Sample Standard Deviation, n: Sample Size). Hence we are able to use the data to find a test statistic and p-value which will be presented in the Results section. Let $\alpha = 0.06$ as a cut-off for rejecting or not rejecting H_0. If p-value $< \alpha$ then we will reject H_0. In contradiction, if p-value $> \alpha$ then we won't reject H_0.

5. Maximum Likelihood Estimator

Conversion is a key element in the campaign strategy; after all, every firms are looking for turning potential customers into buyers (at a high rate). According to Kim, conversion rate optimization enables firms to maximize every cent of their campaign spend by finding that sweet spot that convinces the maximum percentage of their prospects to take action. "When you run a transformation optimization activity, your goal is not only to increase the number of transformations, but also to increase the average value of the transformations" indicates that having knowledge of the mean for the conversion rate could give the xyz company some suggestions to improve it (Filip, 2020). As a result, we are interested in knowing the mean of conversion rate and Maximum Likelihood Estimator can help us estimate it. MLE (Maximum Likelihood Estimator) is the method of finding the value of one or more parameters for a given statistic which makes the known likelihood distribution a maximum.

To begin with, we assume the data of conversion rate is a random sample of Exponential random variables with parameter, λ (based on the conversion rate in data section). The goal of the Maximum Likelihood Estimator is to fit a distribution to the data and find "optimal" parameter (in exponential distribution $mean = \frac{1}{\lambda}$). Hence, we will use the maximum likelihood estimator (MLE) approach to estimate the mean, $\frac{1}{\lambda}$. The result of MLE for $\frac{1}{\lambda}$ will be presented in the Result section. All derivations regarding the MLE can be found in Section 2 of the Appendix.

6. Confidence Interval

After calculating the unknown parameter using MLE for Conversion Rate, now we want to know the mean for the Click Through Rate. Estimating the true mean by the methodology called Confidence Interval (Z approach) which is a kind of estimation type calculated from the observation data. This gives us a range of values for the unknown parameter (mean μ). The decision of choosing Z approach is based on the assumption (data follows Normal distribution), population standard deviation (σ) is unknown and the sample size we generated is large enough. Using a higher confidence level results in a less accurate confidence interval. In this case, instead of 99% Confidence level we will choose a 95% so that the result of estimation will be more precise.

Assume that we have a sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma)$, and we estimate the unknown parameter μ using the estimator $\hat{\mu} = \bar{X_n}$. By central limit theorem that the distribution of the sample means will be approximately normally distributed. Then, we know that

$$\frac{\bar{X_n} - \mu}{s/\sqrt{n}} \sim \mathrm{N}(0, 1)$$

Further calculation and result will be shown in the Result section.

Results

There will be 6 results of each methodologies included in the Methods. By Goodness of Fit Test, we do not reject the hypothesis for the distribution of clicks. Result of Linear Regression seems not practical since it does not support the linear relationship between approved conversion and clicks. According to Bayesian Credible Interval and Hypothesis Test, two groups of people (women and age between 30 to 34) tend to have higher conversion rate. Maximum Likelihood Estimator estimates the mean for conversion rate and Confidence Interval estimates the mean for click through rate.

1. Goodness of Fit Test

Assume the clicks we count will be a random sample from the population of all clicks. The chi-square statistic of 0.559 that we calculated corresponds to a particular location on a chi-square distribution with one degrees of freedom. We now need a p-value, to determines the probability of obtaining a test statistic at least as extreme as 0.559 while assuming the null hypothesis is true. Using the 'pchisq()' function in R Language get the p-value 0.281 which is greater than the α, thus we won't reject H_0.

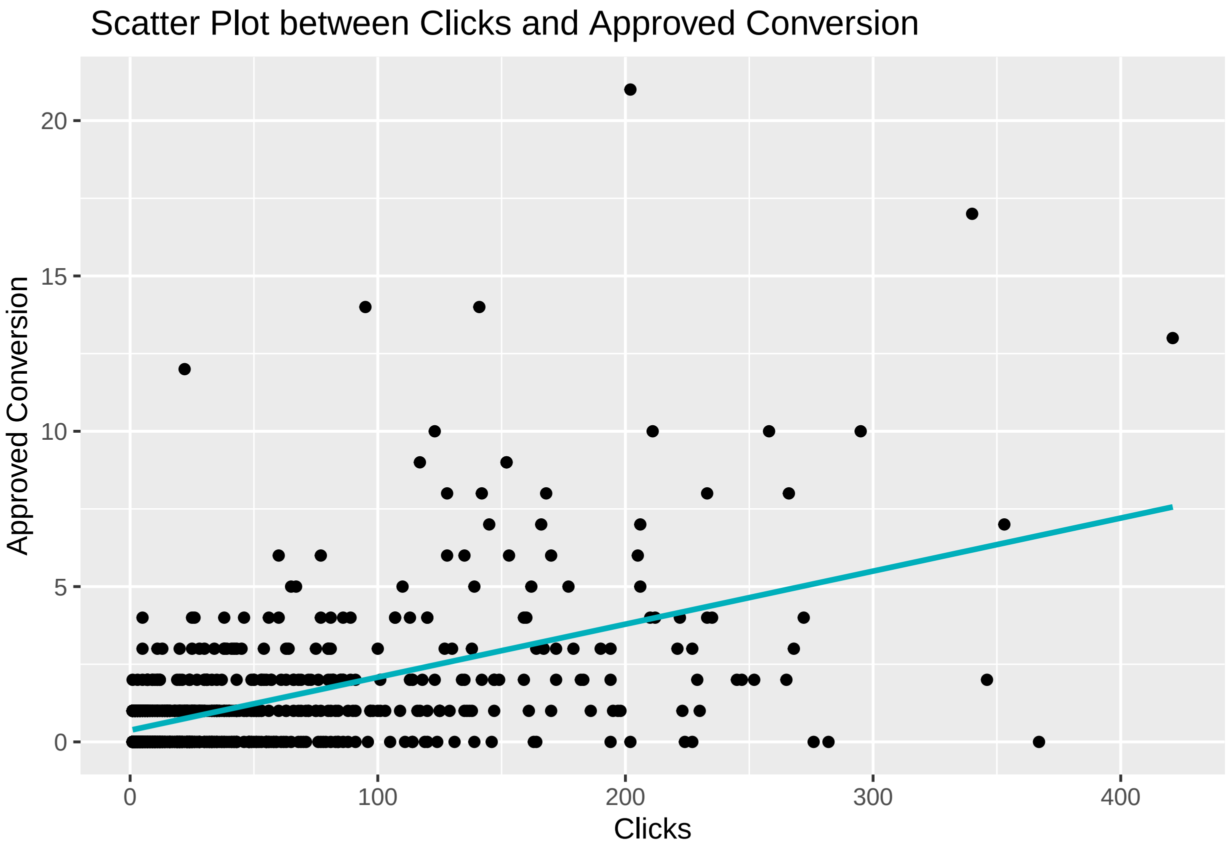
Based on the calculations, the distribution of the number of clicks is statistically significant with a p-value of 0.281, well above the defined alpha of 0.06. Last but not least, the study supports the null hypothesis that number of clicks (X) follow Poisson distribution with parameter $\lambda = 41$.

The conclusion supports that average number of clicks per day is 41. However, it's not possible to determine how many of the clicks the ad campaign is benefiting from since we don't know the conversions. Thus, we couldn't draw any other conclusions even though we have an idea of the distribution of clicks.

2. Linear Regression

First, use R code get the estimates population regression line for $\hat{\alpha}$ and $\hat{\beta}$. See below table for exact value for $\hat{\alpha}$ and $\hat{\beta}$.

$\hat{\alpha}$	0.375
$\hat{\beta}$	0.0171



In this estimated linear regression model for the bivariate data set $(\hat{x}_1, \hat{y}_1), (\hat{x}_2, \hat{y}_2), ..., (\hat{x}_n, \hat{y}_n)$ assume that $\hat{x}_1, ..., \hat{x}_n$ are non-random and that $\hat{Y}_1, ..., \hat{Y}_n$ satisfying: $\hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i + U_i$ for $\mathrm{i} = 1, 2, \ldots, \mathrm{n}$

1) Examine the regression line:

For this model, $\hat{Y}_i = 0.375 + 0.0171\hat{x}_i$ ($\hat{Y}_i$: Approved Conversion and $\hat{x}_i$: clicks). $\hat{\alpha} = 0.375$ represents the intercept of the estimated regression line above. If there are no clicks at all, then the approved conversion is 0.375. It does not make practical sense as there should be 0 number of conversion if no click, but we don't need to pay much attention when there is no click in the campaign. $\hat{\beta} = 0.0171$ represents the slope of the estimated regression line. If there is one more click made by the potential customer, + 0.0171 approved conversion. This do make practical sense, imagine that, if there are more clicks, then there will be more conversions. Similarly, if there are less clicks, then there will be less conversions made by the potential customer.

2) Examine the graph:

From the graph, we can see that all these points spread evenly on both sides of the line which implies that the assumption $\mathrm{E}(U_i) = 0$ holds. However, these points are not parallel to the line then it has a non-constant variance $\mathrm{Var}(U_i)$ since the residuals increase with the fitted values in a pattern, the errors do not have constant variance. This is a violation to the assumption of a constant variance. Therefore, in this case, the model we created is not valid and we couldn't claim Approved Conversion and clicks have linear relationship. The result is not what we expected and does not have practical value.

3. Bayesian Credible Interval

The assumption of the prior in the method section allows us to derive the posterior pdf of the parameter of interest. Thus, posterior distribution of p is $Beta(x+5, n-x+8)$. This pdf can then be used to answer probability questions regarding to the conversions made by female. After calculation using R programming, the 2.5th and 97.5th sample percentiles are estimates of the 0.025 and 0.975 quantiles of the distribution of $Beta(x+5, n-x+8)$. We have 95% credible interval of the proportion of the conversion made by female is $[0.492, 0.527]$. In conclusion, there is a 95% probability that the true proportion of conversions made by female is between 0.492 and 0.527 (rounding to the third decimal place). Hence, this credible interval gives us a sense of the proportion of conversions made by female is slightly more than male's.

Overall, women tend to make the purchase more after clicking the ad than men. Getting to know who and where the target group is, the manner by which they carry on and for what reason is an indispensable piece of running a meaningful ads.

4. Hypothesis Test

Follow the hypothesis test in Method Section, test statistic $= \frac{(\bar{X}-\mu)}{s/\sqrt{n}} \sim t_{n-1}$, p-value: $p = P(|t_{n-1}| > |t|)$, $\alpha = 0.06$. Assume that the distribution of the (data and thus the) sample mean is normally distributed and data was collected randomly. First, we calculated the sample mean (15.44) by taking the sum of the 30-34 conversion rate and divided by its observations. Similarly, we have the sample standard deviation. The test statistic of 1.413 that we calculated corresponds to a particular location on a t distribution with (sample size - 1) degrees of freedom. We now need a p-value, to determines the probability of obtaining a test statistic at least as extreme as 1.413 while assuming the null hypothesis is true. Using the 'pt()' function in R Language get the p-value 0.159 which is greater than the significance level α, thus we do not reject H_0.

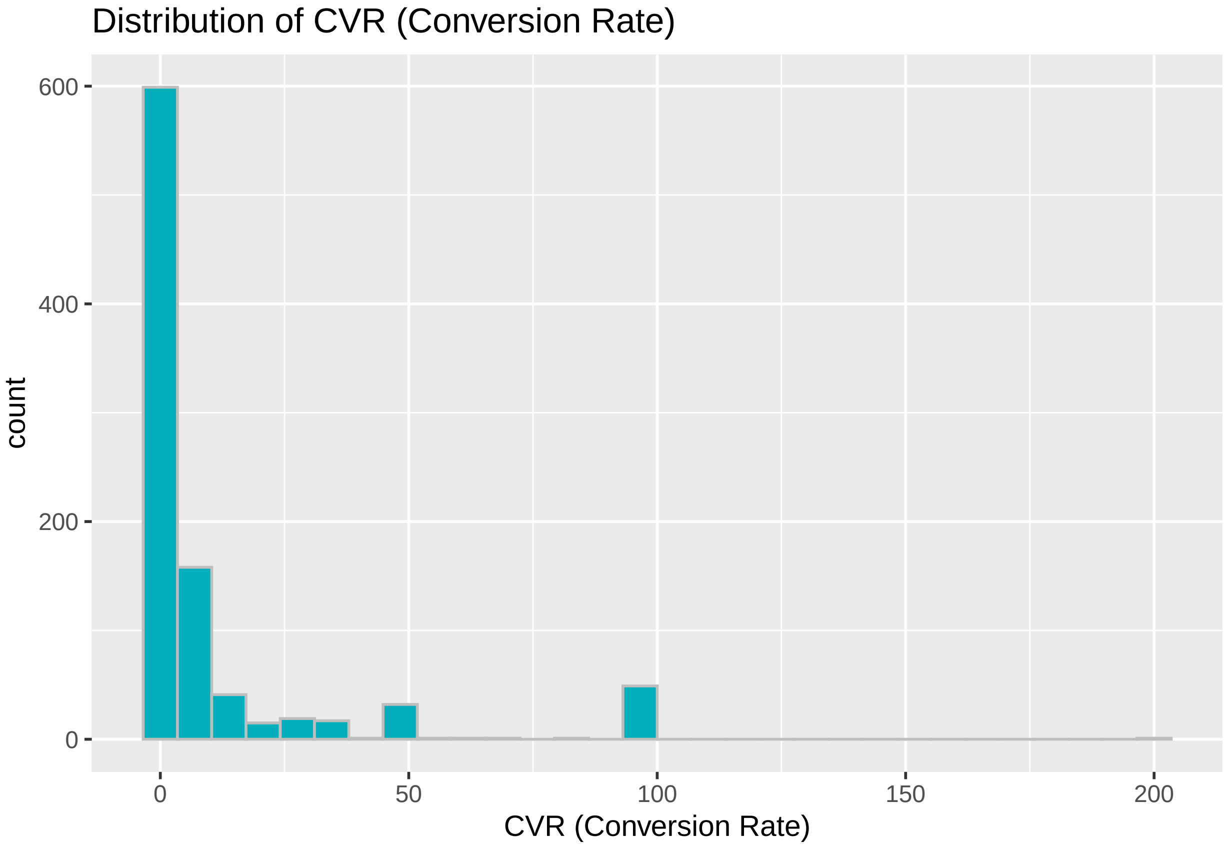
Based on the calculations, the population mean of the 30-34 age range's conversion rate is statistically significant with a p-value of 0.159, well above the defined alpha of 0.06. In the final analysis, the study supports the null hypothesis that population mean of the 30-34 age range's conversion rate is 13%.

Conversion rate does provide objective and accurate information about the performance of the products. However, higher conversion rates doesn't necessary lead to more profits. The key numbers will ultimately determine whether new transformation generates sufficient ROI and whether the campaign can be considered a "success" which could be a next step for xyz company (Filip, 2020).

5. Maximum Likelihood Estimator

We have assumed that the data of conversion rate is a random sample of Exponential random variables with parameter, λ. By applying the approach that MLE of λ is $\frac{1}{\bar{X}}$. Then the mean $\frac{1}{\lambda}$ is $\bar{X}$ and using the function in R to calculate $\hat{\lambda} = 0.0917$ i.e.$\frac{1}{\hat{\lambda}} = \bar{X} = 10.903$. Histogram below is the Distribution of CVR (Conversion Rate) with the data generated. This frequency graph shows exponential decrease is a pattern of data that shows greater decreases with larger conversion rate which creating the curve of an exponential function.

It's helpful to know where xyz company stand compared to others in the industry. WordStream, a SAAS advertising platform, analyzed some data to provide a benchmark for this situation. Overall, the average conversion rate for Facebook ads is between 9-10% across all industries. The mean conversion rate of xyz company's campaign (10.9%) seems like a "fine" percentage compare to the average conversion rate on Facebook Ads. Despite the good conversion rate, there are still opportunities for further enhancements by using A/B testing to optimize their website's funnel (used to track the steps that lead up to that conversion) in online marketing.



6. Confidence Interval

Get the sample size from the data set which is 936 and calculate the sample mean (0.02%) and standard deviation (0.0094%) of the Click Through Rate. To calculate the confidence interval for mean, we have $\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$. The 95% confidence interval for the mean CTR is $[0.0194, 0.0207]$. Hence, we are 95% confident that the true mean lies between 0.0194%

and 0.0207% (rounding to the third decimal place). However, CTR varies between different industries.

To determine what a good click-through rate would look like for xyz company, we need to start by researching industry's average click-through rates. According to Irvine(2020), the average Click Through Rate on Facebook Ads is 0.90% but xyz company's campaign obtain the mean click through rate between 0.0194% and 0.0207% which is much lower then 0.90%. Such low click through rate may indicate that this campaign may targeting the wrong audience or the ads may not greatly attractive to the potential customers.

Conclusions

The motivation behind this project is to help this xyz company to increase their conversion rate of the Facebook marketing campaign. In this section, we will wrap up by addressing two research questions (1. How to optimize this social advertising campaign to get a higher conversion rate? 2. What are the perfect target group in this campaign?) To answer these two research question, let us examine this issue from three points of view include clicks, target group and CVR & CTR. Besides, we will highlights some future analysis recommendations regarding the objective.

I. Clicks

When visitors click on an ad, the most common scenario is that they are taken to the advertiser's website, hoping to have a sale or a potential customer captured there. All the transactions occurs after the "single" click. In this study, we first analyze the clicks variable using goodness of fit test by making an assumption of its distribution. Goodness of fit test do support our assumption, nonetheless, to perceive how the xyz's campaign are faring, it is necessary to look further than simply the quantity of clicks the advertisements are getting. Analyzing the influence factor of conversion could have a deep insight of the campaign. In practical, more the number of times the ad is clicked more approved conversion is achieved. However, the linear model we created is not valid thus we can not claim there is a positive relationship between approved Conversion and clicks.

II. Target Group

Since Facebook marketing is focused on the correct target group who will purchase from xyz company. Finding the perfect target demographics from this data set include both age and sex to optimize the social ad campaigns. By the result of Bayesian Credible Interval, female will in general make the purchase more in the wake of tapping the ads than men. Despite from that, different age range perform a different conversion patter. Hypothesis test supports our assumption that population mean of the 30-34 age range's conversion rate is 13% (higher than the mean of conversion rate 10.9%). Thus age group of 30-34 should be the main aim for xyz's marketing campaign. There are 3 ads in this data set but we didn't analyze it separately, so next step could be compare the individual ad performance and find the targeted audiences independently.

III. Conversion Rate & Click Through Rate

Conversion rate are a powerful method of analyzing the presentation of numerous promoting channels. Through Maximum Likelihood Estimator, the mean conversion rate of xyz company's campaign is 10.9% which is an exciting percentage compare to all industries. For future analyses, calculating ROI (return on investment) is an available option since CVR can likewise be utilized to set ROI assumptions when scaling a campaign. Even if it is not truly fair and accurate to compare xyz's conversion rate to that of a whole industry, we don't have more information about the category industry of xyz's business. Although the mean of conversion rate for xyz's marketing campaign is pretty high among the average in all industries, the click through rate is much lower than the average by confidence interval. From one perspective, having irrelevant keywords will target the wrong audience or a low ad rank (which determined by CPC bid) both leads to below than average CTR. In another point of view, the high-intent of people clicking through who know exactly what they want and don't click through until they're ready to take action (Quick, 2020). For instance, the employment and job training industry, has the lowest CTR, but when people do click through, they convert at the third-highest conversion rate. Apparently, the drawback here is the same as the comparison of average conversion rate.

Weaknesses

MLE (Maximum Likelihood Estimator) and Bayesian Credible Interval both requires strong assumptions about the structure of the data which is the disadvantage in the study. For MLE, the likelihood equations should be explicitly turned out for a given distribution problem. In particular, $L(\lambda)$ need to be worked out for Exponential distribution in our data which is non-trivial. For Bayesian Credible Interval, a wrong assumption of prior will lead to the wrong estimation, finally it will give us a totally wrong prediction. It requires translating intuitive prior into a mathematically formulated prior. For our problem, we assumed $p \sim Beta(5, 8)$ leads to the posterior $P(p \mid \text{ data }) \sim \text{Beta}(x + 5, n - x + 8)$. Thus, having more information of prior could be considerate.

Discussion

To sum up, the performance of campaign in this data set is up to the mark compare to the average Facebook marketing. For future improvement of conversion rate, conducting an A/B test to find the right layout to drive higher percentage of website visitors to convert into transaction or customers. Besides, getting the right visitors to the ads is an important way to optimize the conversion rate . The ideal target demographics are women and age group between 30 to 34 as these group tends to resulting a higher conversion rate.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)

2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

4. Gokagglers. *Sales Conversion Optimization.* https://www.kaggle.com/loveall/clicks-conversion-tracking. (Last Accessed: September 26, 2017)

5. Irvine, M. (2020) *Facebook Ad Benchmarks for YOUR Industry [2019].* https://www.wordstream.com/blog/ws/2019/11/12/facebook-ad-benchmarks. (Last Accessed: July 17, 2020)

6. Kim, L. (2020) *What's a Good Conversion Rate?.*https://www.wordstream.com/blog/ws/2014/03/17/what-is-a-good-conversion-rate. (Last Accessed: August 5, 2020)

7. Filip, A. (2020) *The unexpected difference between conversions and ROI.* https://www.omniconvert.com/blog/the-unexpected-difference-between-conversions-and-roi/. (Last Accessed: September 14, 2020)

8. McLachlan, S(2021) *27 Facebook Demographics to Inform Your Strategy in 2021.* https://blog.hootsuite.com/facebook-demographics/.(Last Accessed: January 20, 2021)

9. Quick, T(2020) *Facebook Advertising Benchmarks for Your Industry.* https://instapage.com/blog/facebook-advertising-benchmarks. (Last Accessed: November 2, 2020)

Appendix

1. Bayesian Posterior Distribution Derivation

By assumption made in Method section, X is the number of conversions made by female, n is the total conversions and p is proportion of female who bought the products. By assumption that $X_1, X_2, ..., X_n \overset{iid}{\sim} \text{Bin}(n, p)$, then $X_i \sim \text{Bin}(n, p)$. Therefore, probability mass function of X is:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Additionally, $p \sim Beta(\alpha, \beta)$, $\alpha = 5, \beta = 8$. Probability density function of p is:

$$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \text{ where } 0 \leq p \leq 1, \alpha, \beta > 0$$

$$\begin{aligned}
\text{Prior of p} : f(p) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\
\text{Likelihood of p} : L(p) &= \prod_{i=1}^{n} \binom{n}{x_i} p^{x_i}(1-p)^{1-x_i} \\
&= \left[\prod_{i=1}^{n} \binom{n}{x_i}\right] p^{\Sigma_1^n x_i}(1-p)^{\sum_1^n 1-x_i} \\
\text{Since xi is a single trial (0 or 1)} &= \left[\prod_{i=1}^{n} \binom{n}{x_i}\right] p^{x}(1-p)^{n-x}
\end{aligned}$$

$$\begin{aligned}
\text{By Bayesian Statistics: } P(p \mid data) &= \frac{P(\text{ data } \mid p)P(p)}{P(\text{ data })} \\
&\propto P(data \mid p)P(p) \propto L(p)f(p) \\
&= \left[\prod_{i=1}^{n} \binom{n}{x_i}\right] p^{x}(1-p)^{n-x} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\
&\propto p^{x+\alpha-1}(1-p)^{n-x+\beta-1}
\end{aligned}$$

Let $\alpha^* = x + \alpha, \beta^* = n - x + \beta$. Then $P(p \mid data) \propto p^{\alpha^*-1}(1-p)^{\beta^*-1}$.

Since the probability density function of Beta distribution is:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \text{ where } 0 \leq x \leq 1, \alpha, \beta > 0$$

Therefore, we are able to observe that $P(p \mid data) \sim \text{Beta}(\alpha^*, \beta^*)$, which implies that $P(p \mid data) \sim \text{Beta}(x+5, n-x+8)$. In conclusion, posterior follows the Beta distribution with parameter $\alpha^* = x + 5, \beta^* = n - x + 8$.

2. MLE Derivation

Let Conversion Rate = X, by assumption that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$, then $X_i \sim \text{Exp}(\lambda)$ Therefore, probability density function of X is: $f(x) = \lambda e^{-\lambda x}, x \geq 0$. For a Exponential distribution the likelihood function (for an iid sample) is:

$$\begin{aligned} L(\lambda) &= f(x_1) \cdots f(x_n) \\ &= \lambda e^{-\lambda x_1} \cdots \lambda e^{-\lambda x_n} \\ &= \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \end{aligned}$$

The loglikelihood function is:

$$\begin{aligned} l(\lambda) &= ln(L(\lambda)) \\ &= ln(\lambda^n e^{-\lambda \sum_{i=1}^n x_i}) \\ &= nln(\lambda) - ln(e^{-\lambda \sum_{i=1}^n x_i}) \\ &= nln(\lambda) - \lambda \sum_{i=1}^n x_i \end{aligned}$$

Derivative of loglikelihood and set it be zero: $\frac{dl}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0$

Then, $\lambda = \frac{n}{\sum_{i=1}^n x_i}$

Second Derivative test of loglikelihood: $\frac{d^2 l}{d\lambda^2} = -\frac{n}{\lambda^2}$.

Since sample size $n > 0, \lambda^2 > 0$, then $-\frac{n}{\lambda^2} < 0$. Thus, $l(\lambda)$ has a maximum at λ. In conclusion, the maximum likelihood estimator of $\hat{\lambda}$ is

$\hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{X}}$