Data Analysis on Wine Review Dataset: Finding which wine is better Wine1 or Wine2

Himanshu Gupta

1.a. (i)Which type of wine is better rated (ii)What is the probability that the Sauvignon Blanc will be better?

Introduction: Nowadays extracting useful information from the data and applying analytical and logical skills has led to many accurate predictions. This analysis is an imperative part of every domain like medical science, financial, IT industry etc. The kaggle has wide range of datasets pertaining to most of the branch. One of the dataset is about reviews of wines in different part of world. This data is suitable for exploratory analysis as based on summary statistics and various graphs, we can detect the pattern and test hypothesis etc.

Dataset Description: The dataset used here can be found on Kaggle website https://www.kaggle.com/zynicide/wine-reviews . This dataset gives the information about varieties of wines consumed in various parts of world. We can see in the description, details of respective wines are given like colour, aroma, different types of fruits used for preparation etc. For each of these records, individual score which varies from 80 to 100, price of wine and their wineries are given. This dataset contains 14 columns and 129970 rows, and do have some missing values. **Below are the fields in dataset**:

Number of rows	Country -> where the wine made
description -> review of wine	Designation
points -> rating of wine out of 100	Price -> cost of wine
province -> province where win is made	region_1 -> region where wine is made
region_2	taster_name -> reviewer of wine
taster_twitter_handle -> twitter id	Title -> consist of winery + year
variety -> type of wine	Winery -> original place of wine

Data Handling : In order to begin with analysis, we first load the dataset present in csv into R. As per question, the important features required are country, price, points and variety, so these features are selected using 'sqldf' function. Now we have filtered out the data with price equals to '15' euros, country Chile with wine variety 'Chardonnay' and country South Africa with wine variety 'Sauvignon Blanc'. We used rbind() function to combine the filtered data. Now the dataset contains some missing values, these values doesn't have any impact on the analysis, hence we dropped the missing values records. There were also some duplicate rows in dataset which were removed using function unique().

Data Analysis: We can perform t-test to find the significant difference between the mean of two samples, related by same features. In this problem we may assume the null hypothesis as mean of Chardonnay and Sauvignon Blanc is equal. **There are few conditions which should be satisfied to carry out t-test**:

1. The sampling distribution should be normally distributed: Plotted histograms for 'Chardonnay' only, 'Sauvignon Blanc' only and total data (both wines). Below are the results:

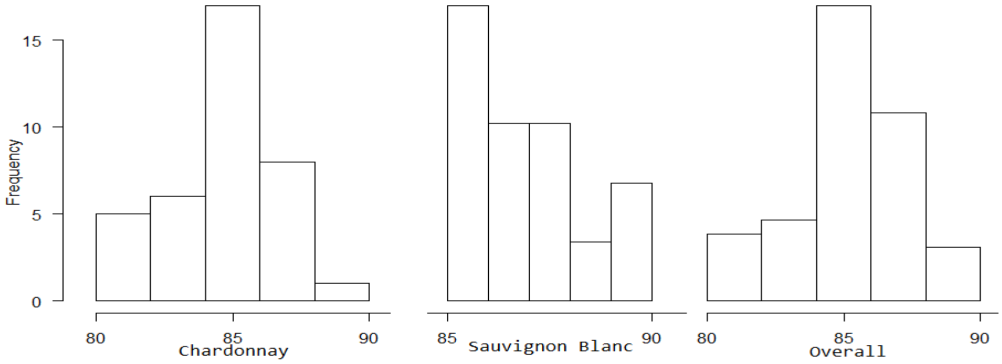


Fig – 1

Above histograms resemble to normal distribution. Hence this condition satisfies. Also we can check the normal distribution condition by plotting **Q-Q plots** (is scatterplot) where quantiles are plotted against one other. If the data points form a straight line diagonally then the distribution can be normally distributed as shown for Chardonnay and Sauvignon Blanc below:

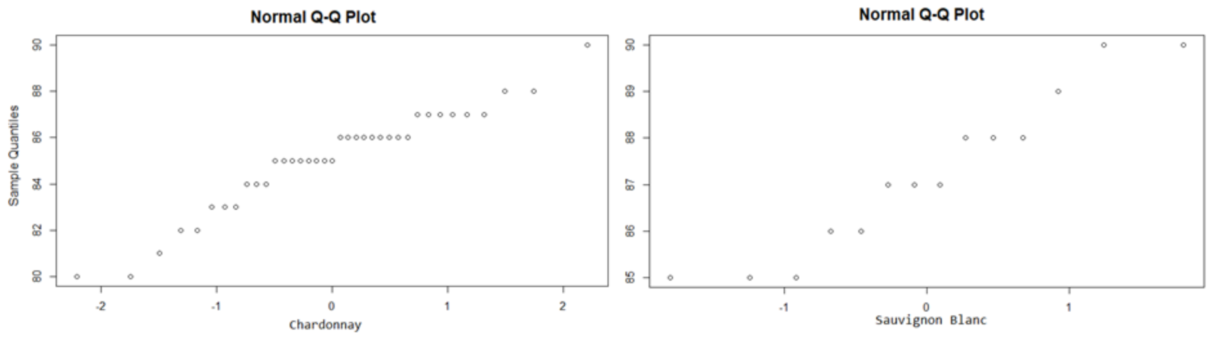
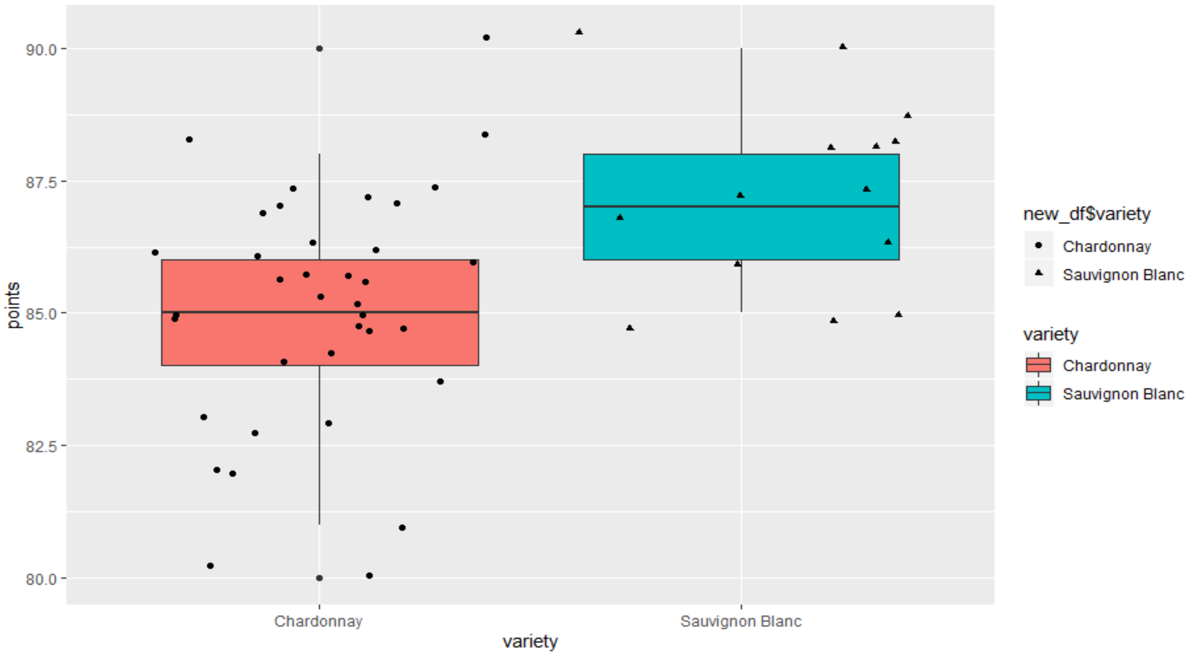


Fig - 2

2. Both samples have Homogeneity of variance, is also a condition: We can check using F-Test:
   var.test(points ~ variety, data=new_df) -> F-test code
   F = 1.6452, num df = 36, denom df = 13, p-value = 0.3372, ratio of variances: 1.645237
   As p-value is greater than 0.05, the distribution is normally distributed.

In order to **visualize** the data we have used ggplot() function because it can plot the complex data with multiple variables with ease and automatic legends, colours etc. Also to explore about the data, I have used box plot as it can handle large data easily, displays outliers and can provide a clear summary at a go.

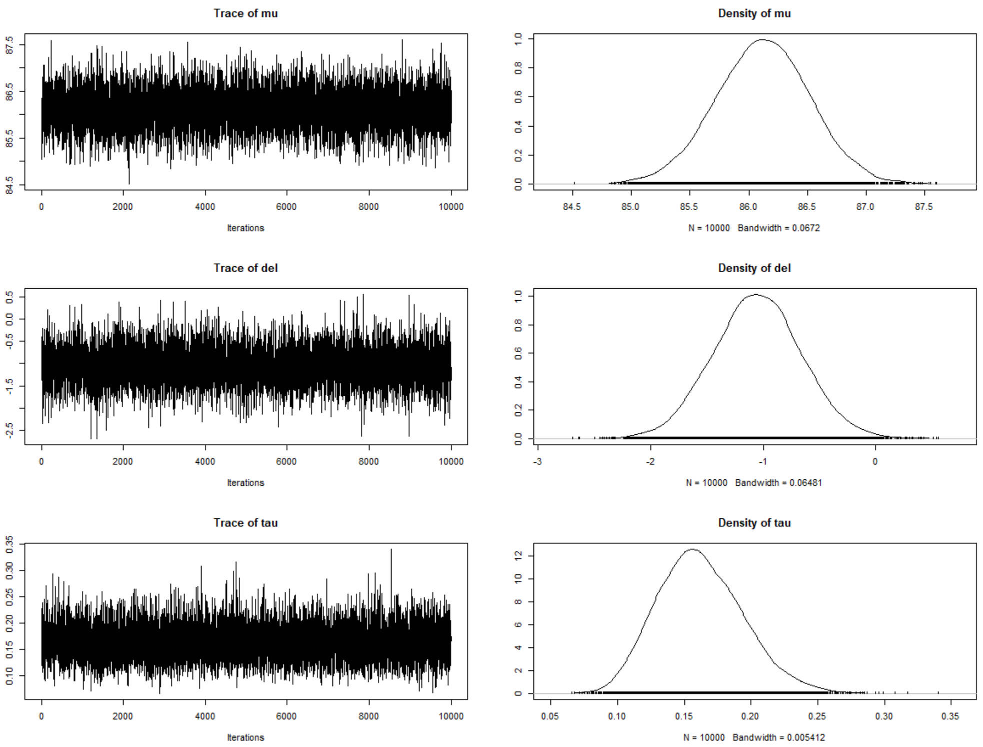
Fig - 3



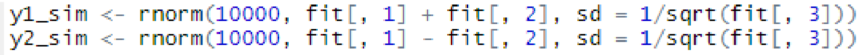
Above figure Fig - 3 shows box plots of two varieties of wines: Chardonnay from Chile is shown with colour Pink and Sauvignon Blanc from South Africa is shown with colour Blue. It can be easily observed from the plot that mid value or median of points for Chardonnay and Sauvignon Blanc are 85 and 87 respectively. The data in Chardonnay is concentrated around 84 to 96 while for Sauvignon Blanc it’s 85 to 88.

Mean Comparison in Bayesian Model

In order to model the comparison between the mean points of individual variety (wines), we will use Gibbs sampler. Hence we will create a function compare_points_gibbs which will compare the differences in means.We have passed various model parameters l function as mu0, tau0, gamma0, a0 and b0 with values 50,1/400, 1/400, 1, 50 respectively. Here we do not have much information about the prior distribution and also to cover maximum number of observations, the mean and S.D can be adjusted with large values. Basically Gibbs sampler function produces iteratively chain of Markov samples where each samples depends upon the previous ones. It provides efficient depiction of marginal posterior densities. Normally, using smaller chain of sample does not provide the desired distribution hence it is recommended to use large chain of sample to have good estimate of the true posterior. We have taken large iterations as 10000; mu, del and tau are rightly distributed. Also we can see that original distribution resembles with distribution formed using generated samples. **Fig - 4**



Below is the code for sample generation based on the Gibbs sampling output. We compared the below given difference means of two wines namely Chardonnay and Sauvignon Blanc and found Sauvignon Blanc to be better



```
y1_sim <- rnorm(10000, fit[, 1] + fit[, 2], sd = 1/sqrt(fit[, 3]))
y2_sim <- rnorm(10000, fit[, 1] - fit[, 2], sd = 1/sqrt(fit[, 3]))
```

Below Fig-5 distribution shows that sampling differences between the two wines which follow normal distribution:

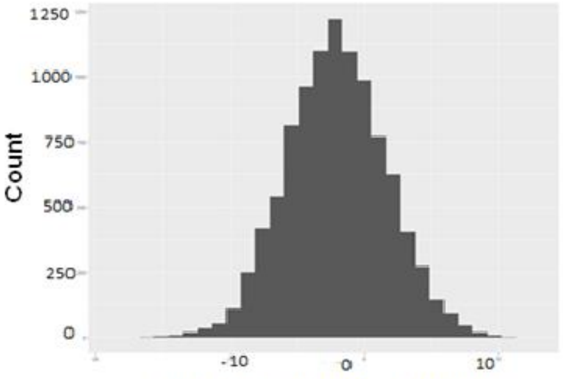


Fig - 5 Simulated Wine Difference

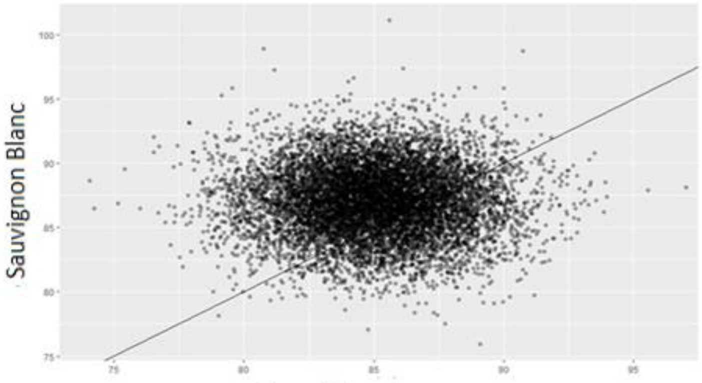


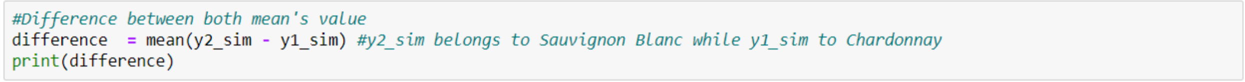
Fig - 6 Chardonnay

The above scatter plot (Fig -6) represents generated samples of Chardonnay vs. generated samples of Sauvignon Blanc. It can be observed that the line split the distribution in two unequal parts, so two samples are not same. Also the Sauvignon Blanc is more distributed above the line than Chardonnay.

Conclusion: From above graphs and discussion, we can infer that:

1.a.(i) Which type of wine is better rated? How much better

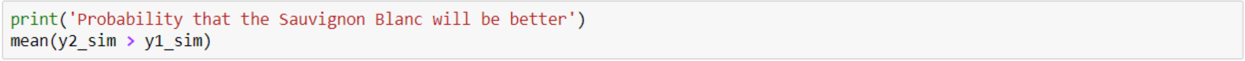
It can be clearly observed that Sauvignon Blanc from South Africa is better rated than Chardonnay from Chile and with **value 2.056 %.** Below is the code snipped for reference:



```
#Difference between both mean's value
difference  = mean(y2_sim - y1_sim) #y2_sim belongs to Sauvignon Blanc while y1_sim to Chardonnay
print(difference)
```

1.a.(ii)What is the probability that the Sauvignon Blanc will be better?

The probability that Sauvignon Blanc is better is **71.3%**. Below is the code snippet:



```
print('Probability that the Sauvignon Blanc will be better')
mean(y2_sim > y1_sim)
```

1.b. Consider the Italian wines in the dataset. Which regions produce better than average wine? Limit your analysis to wines costing less than €20 and to regions which have at least four such reviews.

Data Handling: We extracted the relevant features needed to answer above question which were country, price, points, variety and region_1 selected using 'sqldf' function. As our analysis is restricted to Italy, we filtered out the data points having country equals to 'Italy', having price of wines greater than 20 euros. Also it is mentioned in problem to limit our analysis to regions having at least 4 such reviews which can be interpreted as the regions having number of occurrence of same value more than 3. It was done using the group_by() and filter() functions. There were some missing values in the dataset which were dropped as they did not contribute much towards answer. There were also some duplicate rows in the dataset which were removed using the function unique(). It is important to note that we will use points and region_1 fields after filtering.

Data Analysis: The filtered data has 152 regions with corresponding values of points. The total average value of mean comes out to be 86.47. To visualize the data and outliers we plot boxplot using the ggplot() function. There are 152 boxplots plotted which represents distribution of wine data for Italy plotted with Points vs. Region_1 as shown in Fig-7. We can also see some of the points outside these plots which are outliers. Fig-8 shows wide range of sample sizes.

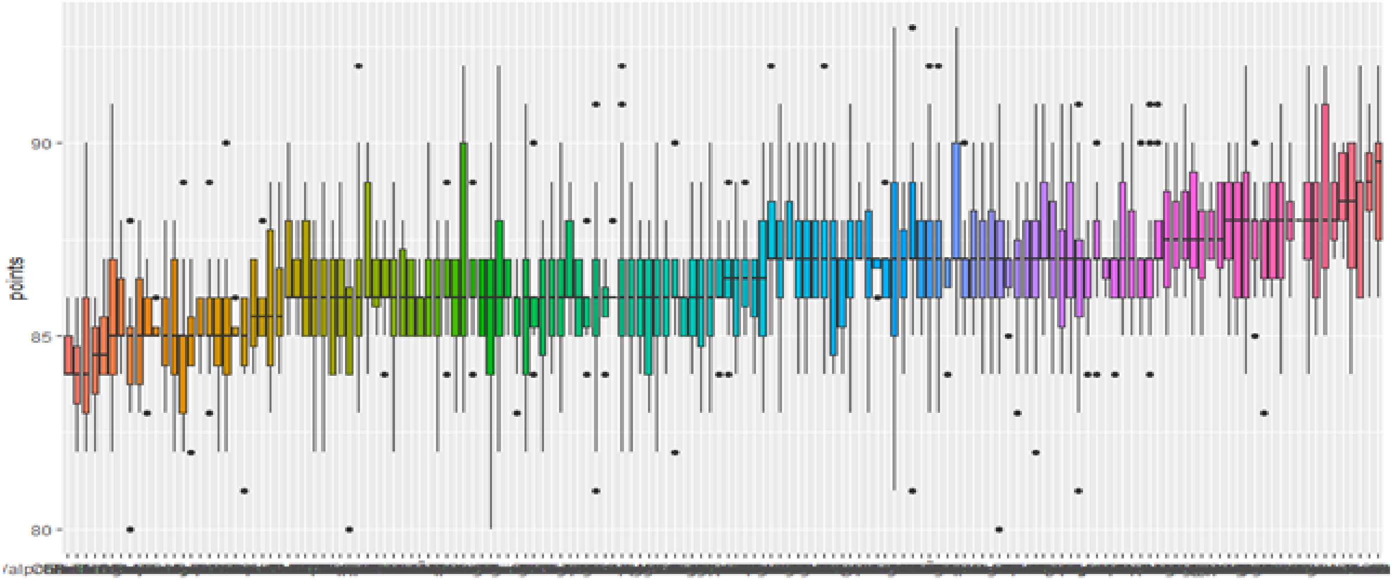


Fig - 7

The Fig-8 shows broad range of sample sizes in each region. The Fig-9 states that the wine point values in the data are very variable. Also maximum rating values lies in dataset between 85.5 – 88.

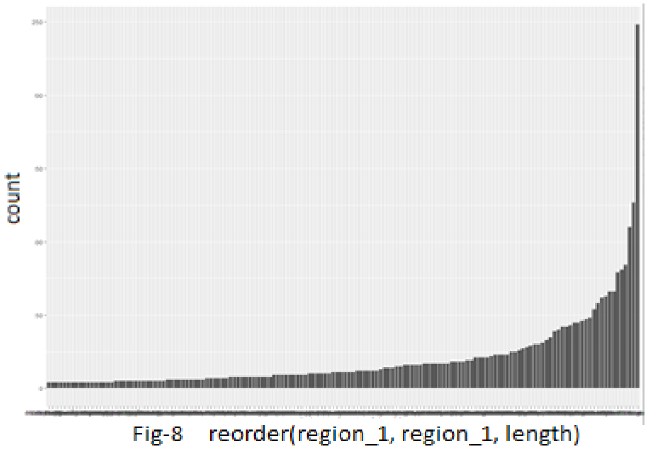


Fig-8 reorder(region_1, region_1, length)

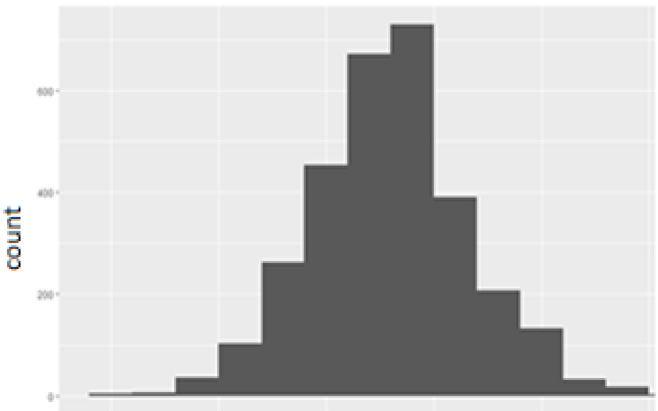


Fig - 9 Points

It can be clearly seen in Fig-10 that for less sample size, most of the mean point value exist. Mostly the mean values lies between 85.5 to 88. In order to get good results we may take some variations into account by modelling mean point values from same population. We can model this via Gibbs Sampler where differences of mean points can be compared.

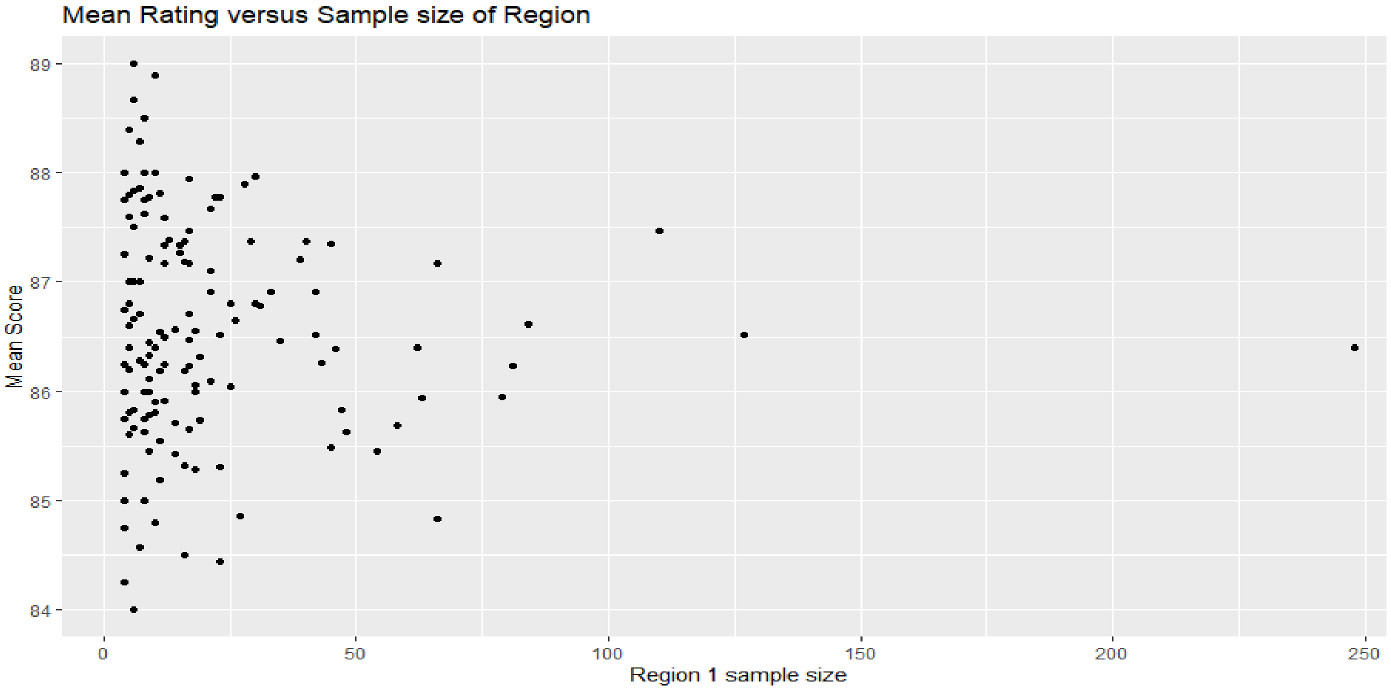


Fig-10

Here we will be use Gibbs Sampler function to produce the posterior for these 152 different samples. This function compare_m_gibbs takes parameters like mu0, tau0, gamma0, a0 and b0 with values 50,1/400, 1/400, 1, 50 respectively as explained in previous problem. Now at every iteration Gibbs sampling function will update the parameter where each samples depends upon the previous ones. Here we have taken one of the parameter as iteration which is equal to 10000, hence after all iterations the distribution produced resemble with the actual one. The Fig shows the boxplots of different generated samples using Gibbs sampling function:

Data Analysis on Wine Review Dataset: Finding which wine is better Wine1 or Wine2

Himanshu Gupta

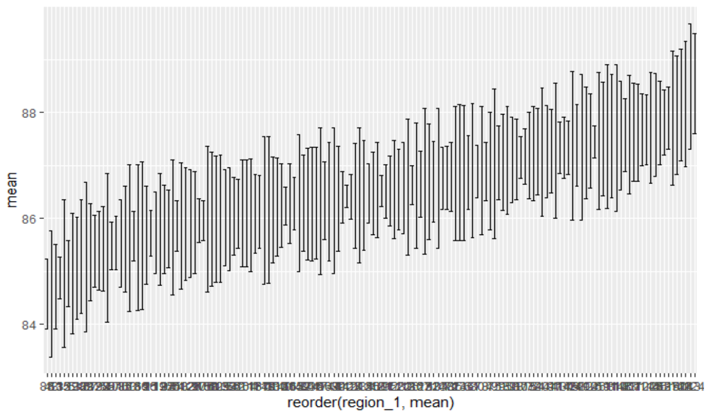


Fig-11 Upper Bound and Lower bound of theta

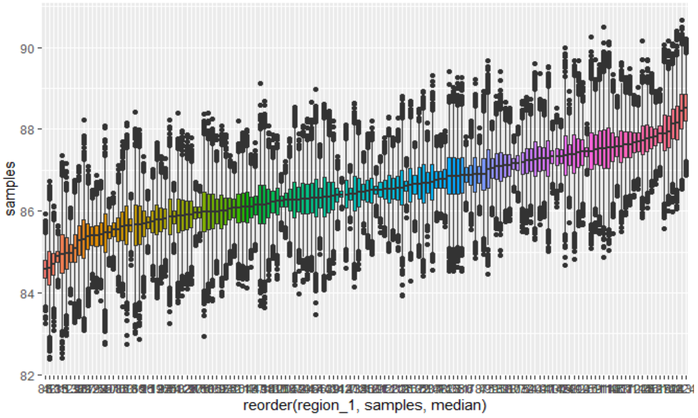


Fig-12 Boxplot for generated sample via Gibbs Function

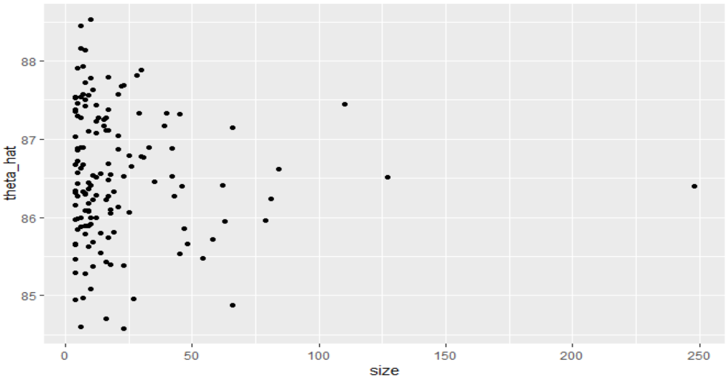


Fig-13 theta_hat vs. relative sample size

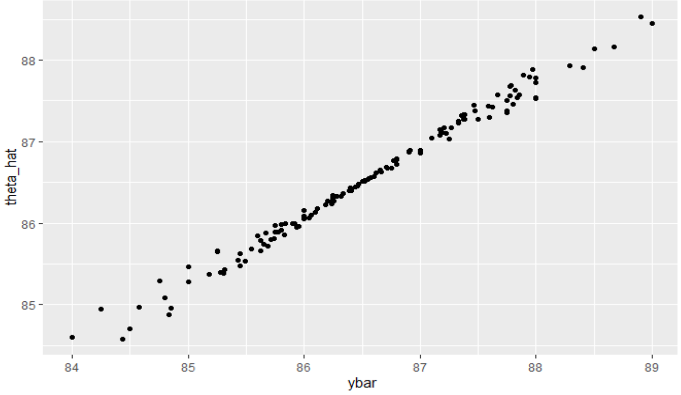


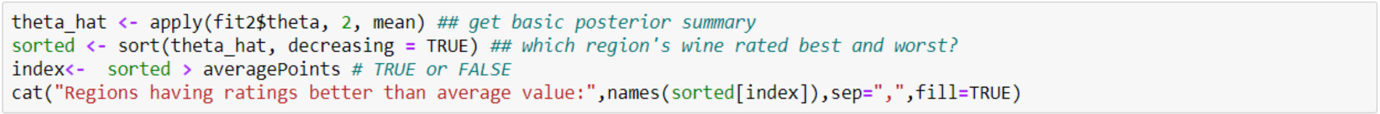
Fig-14 Compares parameter estimate vs. sample mean

Conclusion:

- **Which regions produce better than average wine?**

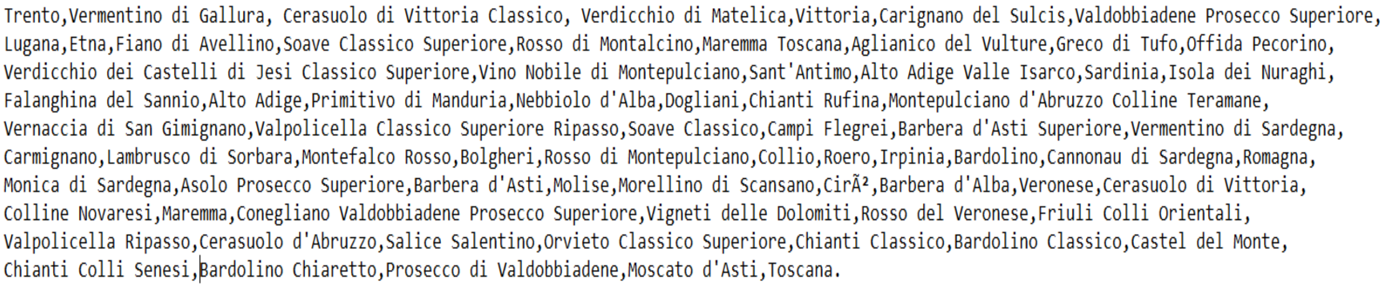
As per analysis, it was found that in Italy there are 73 regions which produced better wine than average.

The results are inferred using above graphs and below code snippet. averagePoints = 86.47



```
theta_hat <- apply(fit2$theta, 2, mean) ## get basic posterior summary
sorted <- sort(theta_hat, decreasing = TRUE) ## which region's wine rated best and worst?
index<-  sorted > averagePoints # TRUE or FALSE
cat("Regions having ratings better than average value:",names(sorted[index]),sep=",",fill=TRUE)
```

Below are the regions which produce better than average wine:



```
Trento,Vermentino di Gallura, Cerasuolo di Vittoria Classico, Verdicchio di Matelica,Vittoria,Carignano del Sulcis,Valdobbiadene Prosecco Superiore,
Lugana,Etna,Fiano di Avellino,Soave Classico Superiore,Rosso di Montalcino,Maremma Toscana,Aglianico del Vulture,Greco di Tufo,Offida Pecorino,
Verdicchio dei Castelli di Jesi Classico Superiore,Vino Nobile di Montepulciano,Sant'Antimo,Alto Adige Valle Isarco,Sardinia,Isola dei Nuraghi,
Falanghina del Sannio,Alto Adige,Primitivo di Manduria,Nebbiolo d'Alba,Dogliani,Chianti Rufina,Montepulciano d'Abruzzo Colline Teramane,
Vernaccia di San Gimignano,Valpolicella Classico Superiore Ripasso,Soave Classico,Campi Flegrei,Barbera d'Asti Superiore,Vermentino di Sardegna,
Carmignano,Lambrusco di Sorbara,Montefalco Rosso,Bolgheri,Rosso di Montepulciano,Collio,Roero,Irpinia,Bardolino,Cannonau di Sardegna,Romagna,
Monica di Sardegna,Asolo Prosecco Superiore,Barbera d'Asti,Molise,Morellino di Scansano,CirÃ²,Barbera d'Alba,Veronese,Cerasuolo di Vittoria,
Colline Novaresi,Maremma,Conegliano Valdobbiadene Prosecco Superiore,Vigneti delle Dolomiti,Rosso del Veronese,Friuli Colli Orientali,
Valpolicella Ripasso,Cerasuolo d'Abruzzo,Salice Salentino,Orvieto Classico Superiore,Chianti Classico,Bardolino Classico,Castel del Monte,
Chianti Colli Senesi,Bardolino Chiaretto,Prosecco di Valdobbiadene,Moscato d'Asti,Toscana.
```