Verifying the Central Limit Theorem with Whisky

Created	@Oct 25, 2020 4:47 PM
Tags	Personal Project

Overview

Scotch, be it with a soda, on the rocks, or neat is an icon of Scotland. The drink is known for its complexity, and wide variety, across a number of price points and levels of quality. Recently I was introduced to a simple but fundamental theorem in statistics, "The Central Limit Theorem" which states the following:

> The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement , then the distribution of the sample means will be approximately normally distributed. [1]

With that in mind I wanted to convince myself that this theorem held. Delving into the depths of Kaggle . I found a data set of roughly 2200 Scotch whisky reviews [2]. The

factor of most interest is the review scores themselves. Each whisky received a rating on a scale of 0 to 100. Of the 2247 whiskies, there are 2223 unique entries. With that in mind it is fair to say there is a huge variety present. To perform the following analysis and visualization the R programming language was used.

Initial Analysis

Firstly we begin looking at the dataset itself. It is composed of 6 features:

- 'name': The name on the label of the bottle
- 'category': The classification of whisky eg: Single Malt.
- 'review.point': A score on the scale of 0 - 100 given by the reviewer.
- 'price': Price paid for the bottle.
- 'currency': Currency used to pay for the bottle.
- 'description': Notes from the reviewer about the whisky.

With any dataset it is key to investigate for any abnormalities in the data. With the exception of some missing characters in the 'name' parameter looked ordinary.

'category' is composed of 5 values, none of which were abnormal, these could be used in future analysis to compare reviews or price across different whiskies.

The 'review.point' despite being a 0 to 100 scale has a minimum value of 63, and a max of 97. This seems to allude to Scotch's testament as a luxury product, with even the worst whisky having some redeeming qualities.

'price' has quite a large range of values and rather inconsistent formatting, and conventions. The data here is characters rather than floats or doubles. Some of the figures have commas, and in one case the phrase '$15,000 or $60,000/set'. All of this would lead to some cleaning if 'price' is to be analysed.

As for 'currency' every single entry is '$'. Therefore this column is essentially redundant at the moment as all the prices are in USD.

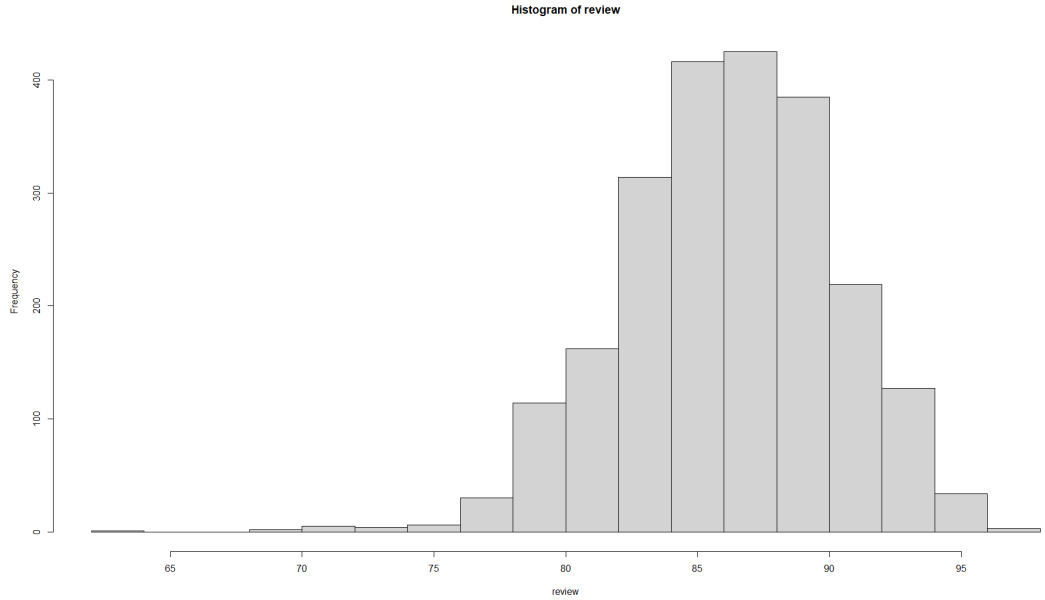
Finally 'description' is quite wordy, with some missing or error characters. Perhaps in future I will take a look at this to perform some sentiment analysis, however for my analysis this will not be necessary.

While 'review.point' is the only piece of data I am currently interested it in, it is worthwhile being familiar with the entire dataset if future testing is to be done.

Initial Visualizations

Following this initial analysis some visuals may aid in gaining some additional insight.

```
set.seed(1343)
scotch <- read.csv("scotch_review.csv")
review <- scotch$review.point
par(mar = c(4, 4, .1, .1))
rePlot <- hist(review)
```

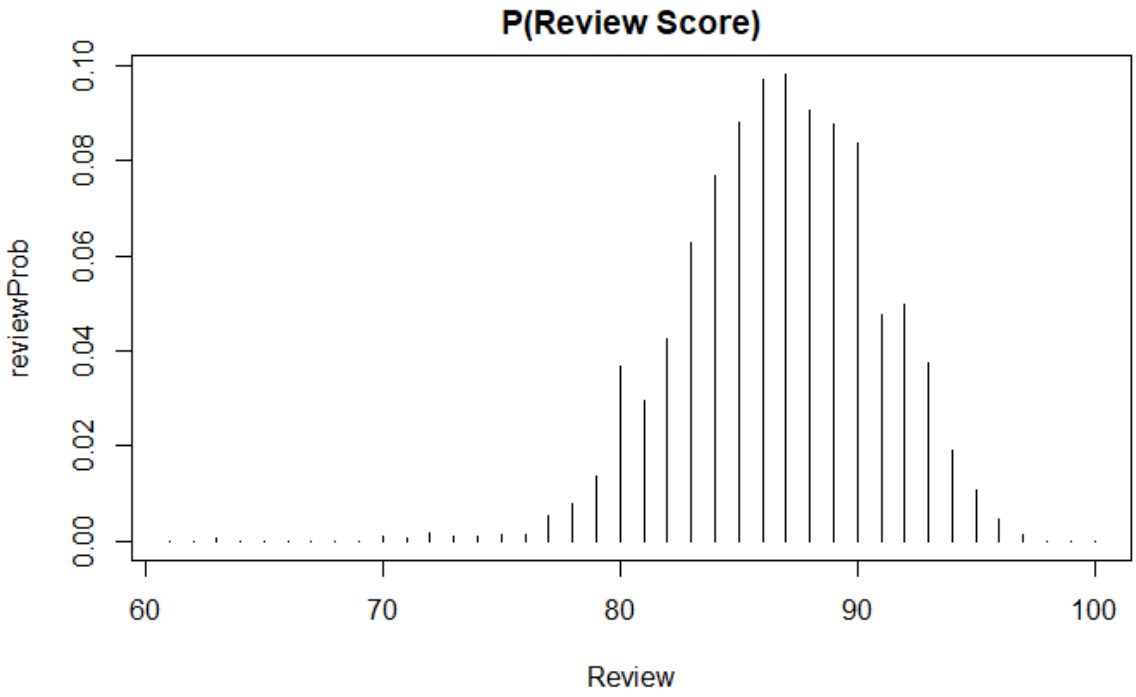


Histogram of the frequency of each review score.

Already one can see a roughly normal distribution, albeit with a bit of a skew, and tailing. Notably a peak can be seen in the mid to high 80 score.

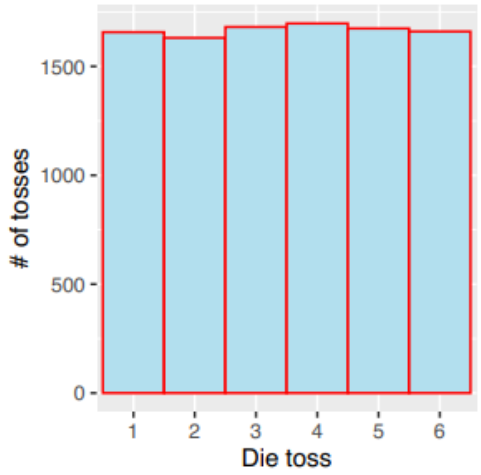
Further from this the probability of any given score can be derived, that is. This is simply the frequency of a score divided by the total number of reviews.

```
reviewCounts <- rePlot$counts
reviewProb <- reviewCounts / sum(reviewCounts)
length(reviewProb)
plot(61:100, reviewProb, type = 'h')
```

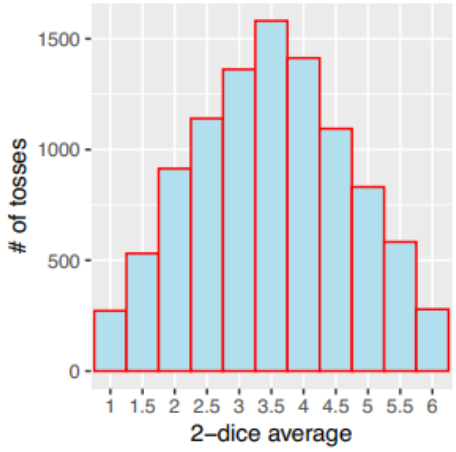


Sampling

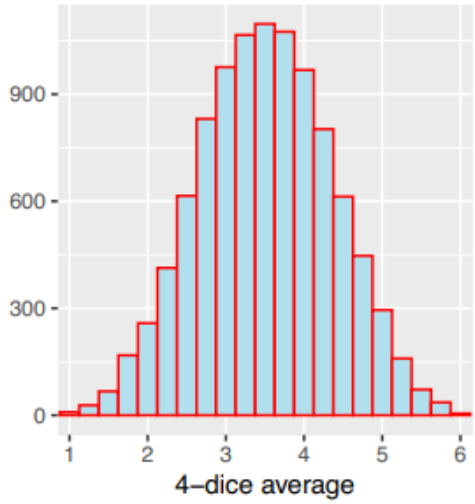
Now comes the part where we begin looking towards getting a normal distribution. The centre of a normal distribution is the mean of the data. However the mean is 86.7 which intuitively may appear strange, as looking at our data no individual measurement is equal to 86.7. This can explained by taking a closer look at what the mean is, it is an expected value. That is to say it is the average value over a number of trials. A more obvious example may be to look at rolling a 6 sided die. When rolling there is a 1 in 6 chance of any given face turning up, however the mean of the faces is (1 +2 + 3 + 4 + 5 + 6) / 6. Which is equal to 3.5, which appears impossible. However if one were to roll 2 dice and take the average of the both rolls one can see 3.5 is indeed possible, and as the number of dice creases that expected value of 3.5 becomes all the more prominent.



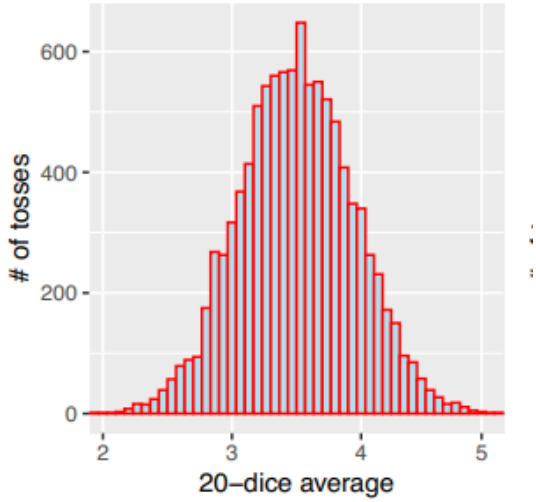
Frequency of outcome of 1 die toss. [3]



Average result of 2 dice tosses. [3]



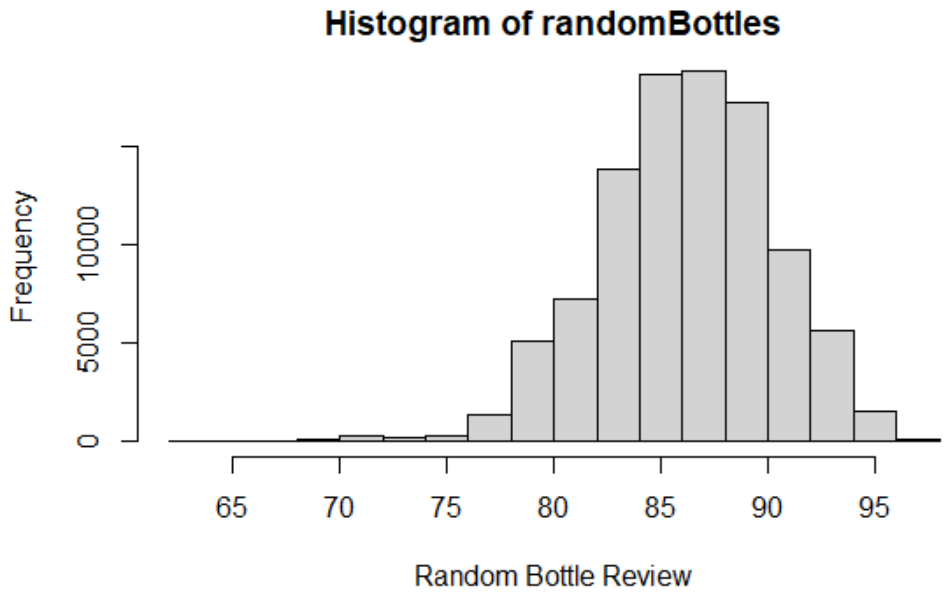
Average result of 4 dice tosses [3]



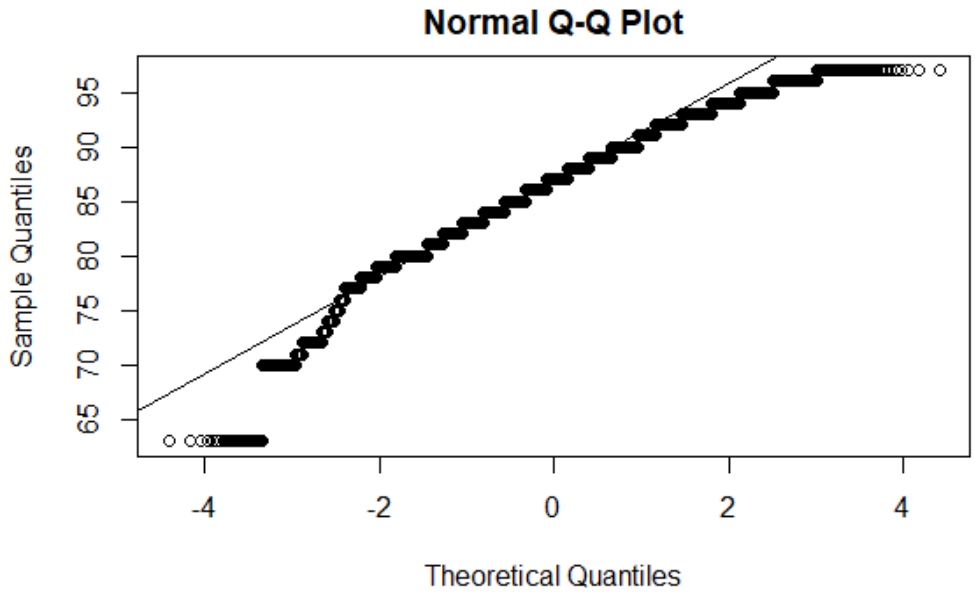
Average result of 20 dice tosses. [3]

This concept is going to be applied to the whisky data to obtain a normal distribution. Suppose a bar shelf with an infinite supply of each whisky in the dataset and one were to pull 100000 bottles down and record their ratings. Something resembling the below plot would be obtained. This can be represented in R as such:

```
randomBottles <- sample(review, size = 100000, replace = TRUE)
hist(randomBottles, xlab = "Random Bottle Review")
qqnorm(randomBottles)
qqline(randomBottles)
```

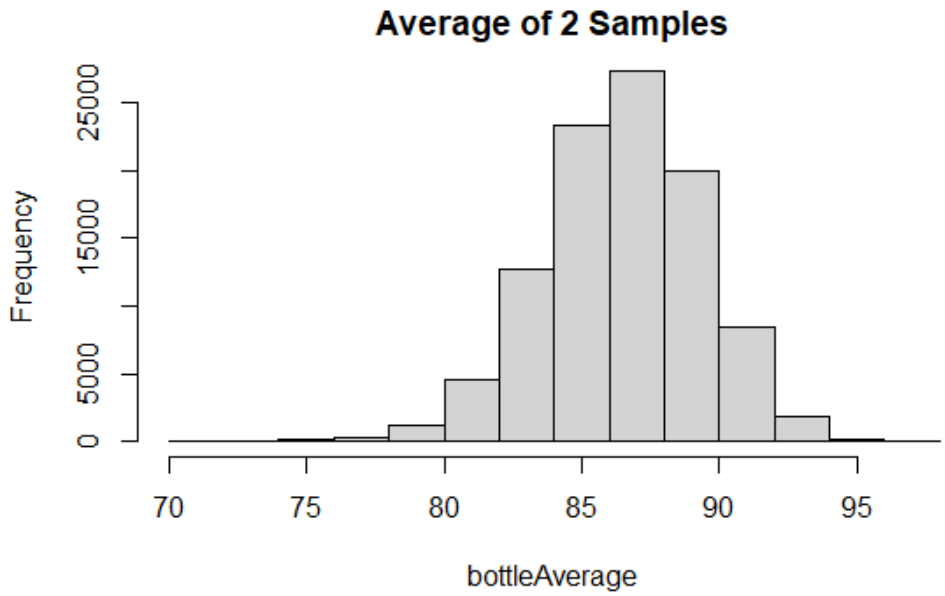


A QQ Plot is also used to gain further insight into the distribution shown. One can see some tailing due to the fall off in values at the tails.



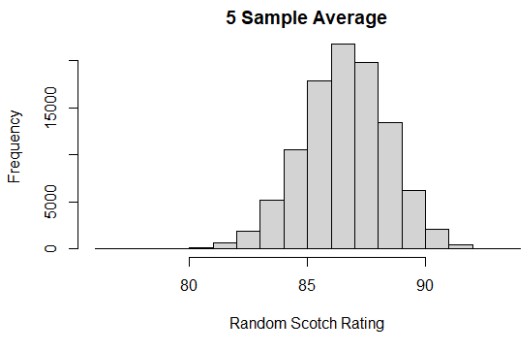
Similar to the dice example, a two samples can be taken and averaged. That is to say take 2 bottles, average their score, and repeat for 100000 pairs.

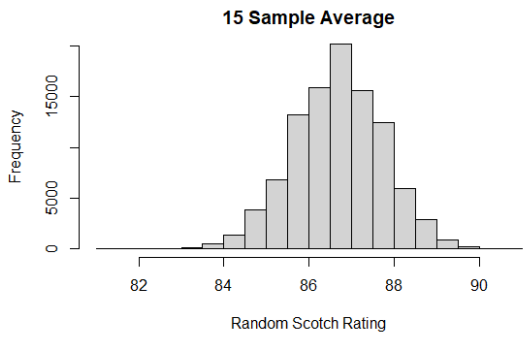
```
bottleAverage <- (sample(review, size = 100000, replace = TRUE) +
                  sample(review, size = 100000, replace = TRUE)) / 2
hist(bottleAverage, main = "Average of 2 Samples")
```

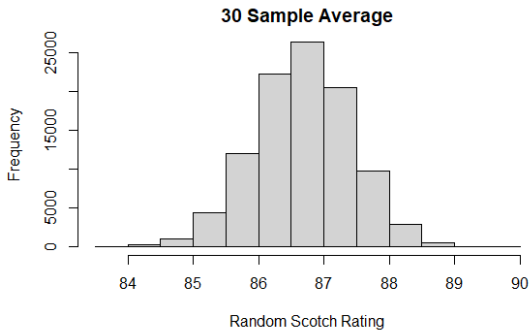


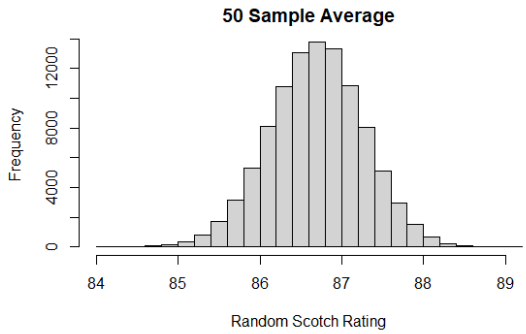
Now taking to its logical conclusion lets sample the data even further for n = 5, 15, 30, and 50.

```
eachSample <- 100
totalSamples <- 10000000
scotchSamples <- matrix(sample(review, size = totalSamples,
                               replace =TRUE),ncol = eachSample)

hist(rowMeans(scotchSamples[,1:5]), main = '5 Sample Average', xlab = 'Random Scotch Rating')
hist(rowMeans(scotchSamples[,1:15]), main = '15 Sample Average', xlab = 'Random Scotch Rating')
hist(rowMeans(scotchSamples[,1:30]), main = '30 Sample Average', xlab = 'Random Scotch Rating')
hist(rowMeans(scotchSamples[,1:50]), main = '50 Sample Average', xlab = 'Random Scotch Rating')
hist(rowMeans(scotchSamples[,1:100]), main = '50 Sample Average', xlab = 'Random Scotch Rating')

s <- rowMeans(scotchSamples)
mean(s)
var(s)
sd(s)
```









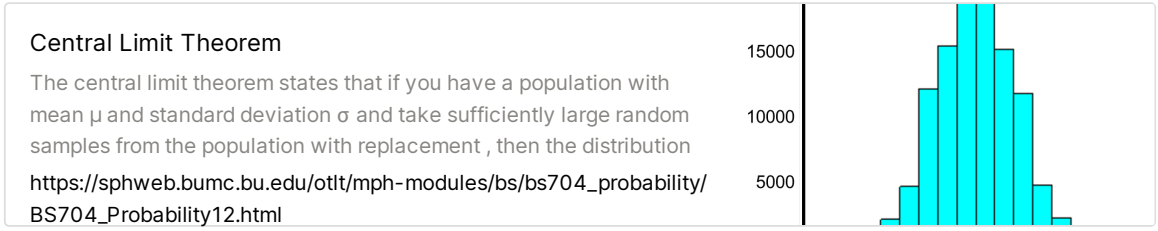
As the number of sampled bottles increases, the distribution approaches normal. Most notably for n = 50, the data is undoubtable normally distributed.

Closing Thoughts

With such an arbitrary dataset displaying normal properties I am happy to say the central limit theorem has been verified in my mind. I certainly hope you can take the concepts of this discussion and apply it to other datasets. The code supplied here is pretty much everything you need with only a couple minor changes here and there. Keep an eye for a further dive on this dataset focusing on price, finding the magic point to maximise price, and quality.

References

1. LaMorte, W.W, 2016:07:24, Central Limit Theorem. Retrieved from:



2. thatdataanalyst, 2018:06:13, 2,2k+ Scotch Whisky Review. Retrieved from:

3. Hurley, C, Central Limit Theorem 2020.