Statistics Assignment

Giampietro Ciancio 965991
Riccardo Valenti 979784

Part 1 (Probability)

Question 1

The coach of basketball team orders to his team to shoot 10 times from inside the area (2 points shots with 50% chance of scoring) and record the percentage of scoring. On the other hand, many basketball players prefer to shoot from outside the area (3 points shots with 30% chance of scoring). The assistant coach observed the players shooting and reported to the coach that 20% players did not follow the rules and decided to shoot from outside the area. Thus, the coach tested one player and asked to report his scoring percentage. The player reported 4 baskets out of 10 shots.

What is the probability that this player did not follow the coach instructions and decided to shoot from outside the area?

Let F the event of scoring 4 baskets out of 10 shots and let C the event of not following coach instructions.

We have that $P(C) = 0.2$. We may write the question as follows.

$$P(C \mid F) = \frac{P(F \mid C)P(C)}{P(F)}$$

In order to compute the probability of scoring 4 baskets from outside the area we can use binomial distribution Bin $\sim (10, 0.3)$.

$$P(F \mid C) = \binom{10}{4} 0.3^4 0.7^6 \approx 0.2001$$

We should also compute the probability of scoring 4 baskets using Bin $\sim (10, 0.3)$ and Bin $\sim (10, 0.5)$.

$$P(F) = P(F \mid C)P(C) + P(F \mid \overline{C})P(\overline{C}) = \binom{10}{4} 0.3^4 0.7^6 0.2 + \binom{10}{4} 0.5^4 0.5^6 0.8 \approx 0.2041$$

Thus we may rewrite the Bayes Formula as follows.

$$P(C \mid F) \approx \frac{0.2001 \cdot 0.2}{0.2041} = 0.196$$

Question 2

Coca Cola sells bottle to 10 different supermarkets in different countries, the price (in millions of dollars) of each contract by the selling of the bottles is very unpredictable due to difficult market conditions. The random variable have a Gaussian distribution of this form

$X_i \sim \mathsf{N}(100, 15^2)$.

The current exchange rate is 0,94€ for 1\$. The price of each contract is independent and identically distributed. Coca Cola wants to estimate the probability that the total amount of money for all contracts will be grater than €950mln

The linear combination of independent random variable having a Normal distribution also have a normal distribution. From this we can derive that the total amount of money from the contracts will be the sum of the 10 contracts stipulated.

$S = X_1 + X_2 + \cdots + X_n \sim \mathsf{N}(10 \cdot 100, 10 \cdot 15^2)$.

To compute the total amount of money in € we need to make a linear transformation of this form:

$$Y = 0,94 \cdot S$$
$$E(Y) = 0,94 \cdot E(S)$$
$$V(Y) = 0,94^2 \cdot V(S)$$
$$Y \sim \mathsf{N}(0.94 \cdot 10 \cdot 100,\ 0.94^2 \cdot 10 \cdot 15^2)$$
$$Y \sim \mathsf{N}(0.94 \cdot 10 \cdot 100,\ 0.94^2 \cdot 10 \cdot 15^2)$$
$$Y \sim (940, 1988.1)$$

Thank to the property of closure under linear transformation, Y is still a normal distributed random variable. From this we can compute the probability that the total amount from the contracts is greater than €950mln.

$$P(Y > 950) = 1 - P(Y < 950) = 1 - \Phi(\frac{950 - 940}{\sqrt{1988.1}}) = 1 - \Phi(0,22) = 0,4129$$

Alternatively, using R.

```
1-pnorm(950, 940,sqrt(1988.1))
```

```
## [1] 0.4112717
```

We can conlcude that the probability that the total amount earned from the contracts is greater than €950mln is 0.4112717.

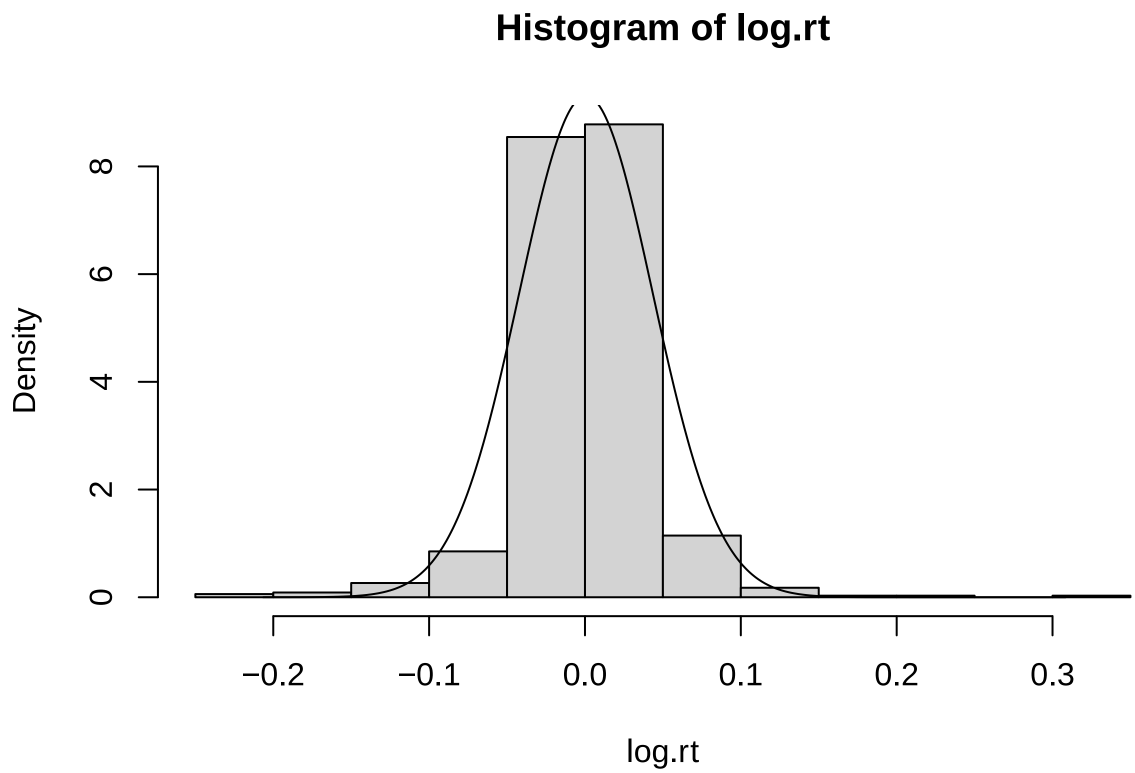
Problem 2 (Inference)

```
mlog
```

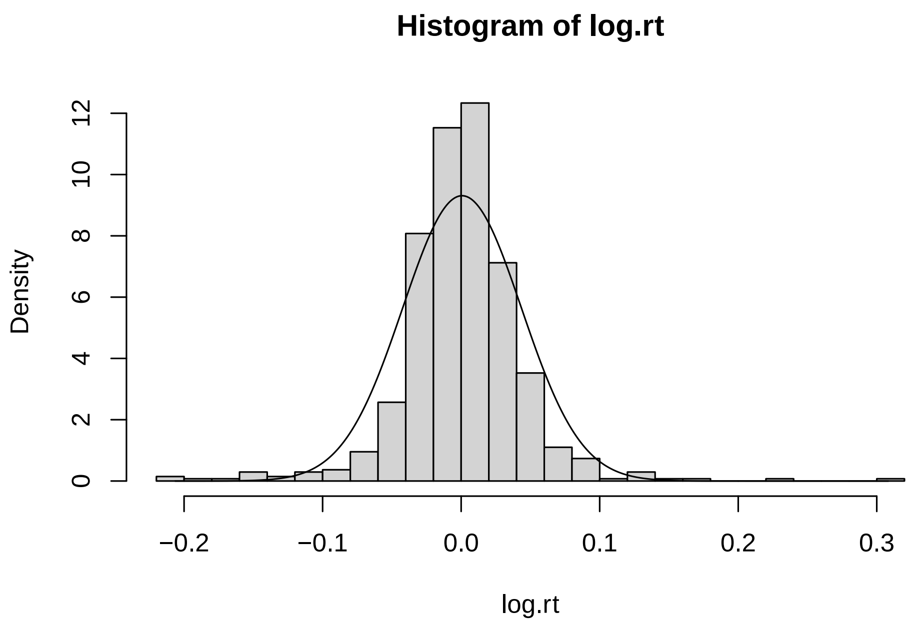
```
## [1] 0.000648585
```

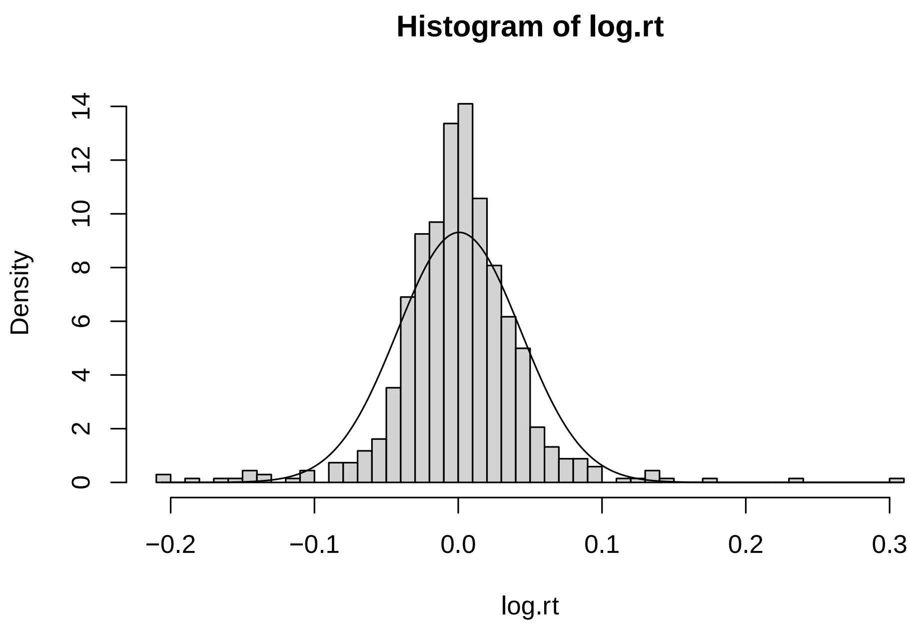
```
vlog
```

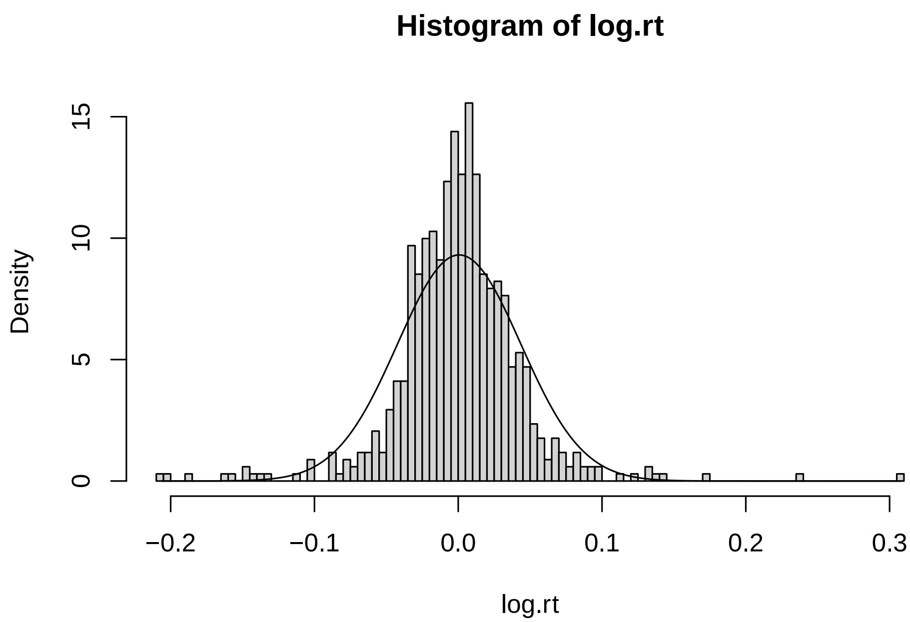
```
## [1] 0.001836494
```



From the previous plot we may observe a similar shape between the Empirical Data and the Theoretical Distribution. Both graph show that we are much more likely to get log return close to 0, which is approximately the mean of the data, and the symmetry between positive and negative log return with respect to the mean (Empirical Distribution) or μ (Theoretical Distribution). We may wrongfully conclude that the log return are approximately Gaussian. It is important to study more in detail the Empirical Distribution of our data. To do so it is necessary to add more bins in our graph. We'll show the case with 20, 50 and 100 bins.

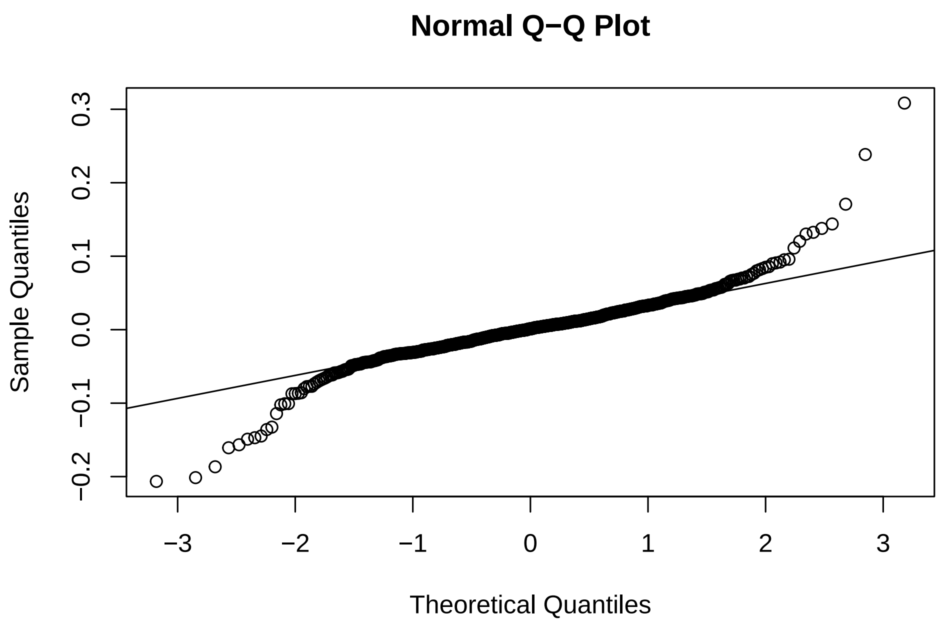

Histogram of log.rt
Density
log.rt
0
2
4
6
8
10
12
−0.2
−0.1
0.0
0.1
0.2
0.3


Histogram of log.rt
Density
log.rt
0
2
4
6
8
10
12
14
−0.2
−0.1
0.0
0.1
0.2
0.3



By increasing the number of bins we observe how much the Empirical Distribution deviates from the Theoretical One. The last graph shows that log return close to 0 are way more frequent than expected from the empirical distribution. The shape of the Gaussian is well explained by the relative high variance of the Data which is caused by the non negligible amount of extreme (both positive and negative) log return. We observe that is way more common to get very low or very high log return than predicted by the Gaussian. The Gaussian would require fatter tails to better approximate the log return of the stock. This observation is surely consistent with the nature of stocks: it is not so uncommon to experience large fall or spikes within the stock market.

Using a QQ plot with the Gaussian as the Theoretical Distribution we can observe in another fashion the difference between Empirical and Theoretical Distribution.



The straight line indicates a perfect correspondence among the empirical and theoretical quantiles. The points, however, reinforce our previous conclusion: the log return tend to present more frequently very high or very low value than expected by Gaussian.Thus, all the points above and below the line indicates an important deviation from the theoretical Gaussian quantiles.

```
t.test(log.rt, mu=0, alternative="two.sided", conf.level=0.99)
```

```
##
##  One Sample t-test
##
## data:  log.rt
## t = 0.39495, df = 680, p-value = 0.693
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  -0.00359330  0.00489047
## sample estimates:
##   mean of x
## 0.000648585
```

The code written above is use to test the null hypothesis (H_0) that the true mean is equal to 0. The output of the code includes the alternative hypothesis (H_1), which is, intuitively, true mean is not equal to 0.

The output also provide for the 99 percent confidence interval which in this case is (-0.00359330, 0.00489047). The confidence interval tells us that the 99% of the intervals generated by this technique included the true mean.

A two-sided significance test rejects the null hypothesis exactly when the claim falls outside the corresponding confidence interval for μ. The confidence interval includes the null hypothesis, thus leading us to accept H_0 at a significance level $\alpha = 1 - 0.99 = 0.01$.

The high p-value, in conclusion, provides an empirical support for H_0. The high p-value (greater than α), which is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct, support that no statistical relationship and significance exists in a set of given single observed variable. Hence, we accept H_0.

Source

Giampietro Ciancio, Riccardo Valenti. 2022. "Stats Assignment." 2022. https://github.com/giampietrociancio/StatsAssignment.