I confirm that the following report and associated code is my own work, except where clearly indicated.

1 Abstract

This report pays attention on using Monte Carlo simulation to investigate the correlation between life expectancy and infant deaths

2 Introduction

The report carries out some simulations in different scenarios and does some analysis to answer the question 'Is there any correlation between life expectancy and infant deaths?' The data used in this research is gapminder dataset from R package 'dslabs' (Rafael, 2018). Simulations, parametric and non-parametric tests as well as the calculation of effect size and power were conducted in R Studio version 3.5.1. (R Core Team, 2015).

3 Methods

By analysing the question and the dataset, we could use Pearson correlation test , a correlation coefficient test, as the parametric test and Spearman Rank Correlation as the non-parametric test. These two tests are used to evaluate the association between infant deaths and life expectancy. The scenarios I considered are the below three:

- Different continents: Africa, Americas, Asia, Europe and Oceania.
- Sample size: large vs. small.
- Standard Deviations: large vs. small.

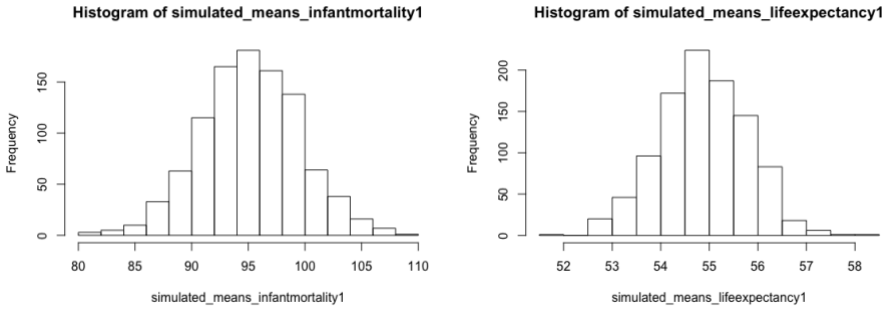
In order to conduct these simulations in different scenarios, Monte Carlo method was used to randomly sample from the distribution. In this case, we use rnorm() to simulate normal distribution.

4 Results

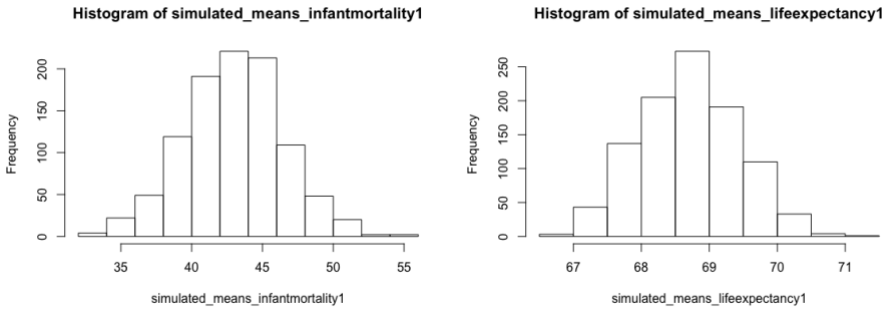
Different Continents

In this scenario, I first compute the mean and standard deviation of infant mortality and life expectancy respectively grouped by five continents. Then the simulation would be presented in each continent using the mean and standard deviation.

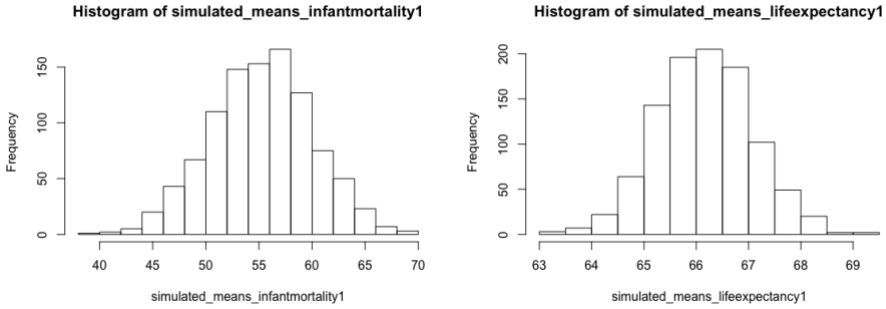
The below two figures are the frequency of mean distributions of both infant mortality and life expectancy of Africa. From the figure, we can see that the distribution is normal distribution and the simulated means conform to one from the original dataset. In order to visualise the data, the scatter plots of the data can be found in Appendices. Also, shapiro-wilk normality test for both variables was conducted as preliminary test to check the assumptions, the p-value of infant mortality if 0.2796 and of life expectancy is 0.1308. From this result, the two p-values are both greater than 0.05. Thus, the normality can be assumed.



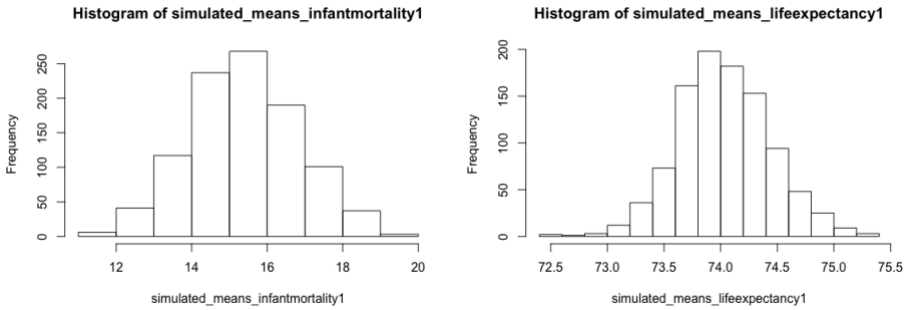
The next two figures are the frequencies of means of Americas. The shapiro-wilk normality test also shows p-value of these two variables are 0.8938 and 0.6309 respectively, which represents the normality assumption confirmed.



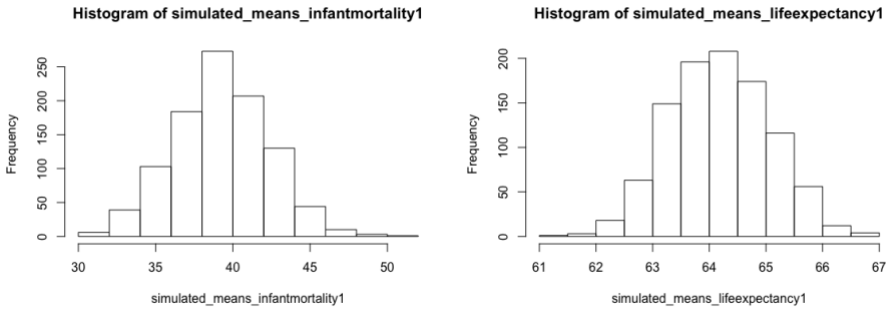
The following two figures are the frequencies of means of Asia. The shapiro-wilk normality test also displays p-value of these two variables are 0.5102 and 0.1389 respectively, which conform to normal distribution.



In addition, frequencies of Europe mean are as below. The p-values are 0.3737 and 0.8311. Therefore, the data follow normal distribution.



Finally, the distribution of Oceania means are as below. The distribution is also normal distribution due to the fact that p-value are 0.9701 and 0.16.



By using Pearson Correlation Test and Spearman Rank Correlation Test, we can get the confidence intervals of correlation coefficient at 95%. And also the p-value for each pairs of variables are greater than 0.05 shows that they are not significantly correlated.

Continent	2.5% CI	97.5% CI	p-value of Pearson Correlation Test	p-values of Spearman Rank Correlation
Africa	0.2564	0.1349	0.5325	0.4021
Americas	0.1538	0.2383	0.664	0.859

Asia	0.2028	0.1900	0.9476	0.9573
Europe	0.2414	0.1506	0.641	0.4933
Oceania	0.0061	0.3730	0.05755	0.06786

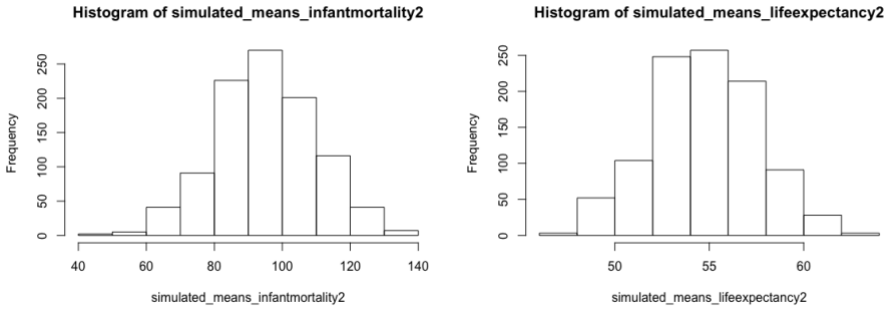
Due to the fact that the data all follow normal distributions, the effect size of each continent is the absolute value of correlation coefficient from Pearson Correlation Test. From the result, the variables in Oceania has small effect while other only have trivial effect. From the figure, we can see that all five powers are even less than 0.1, which means I only have 10% or less change of finding a statistically significant difference.

Continent	Effect Size	Power
Africa	0.0631	0.0957
Americas	0.0440	0.0718
Asia	0.0067	0.0503
Europe	0.0472	0.0752
Oceania	0.1906	0.4784

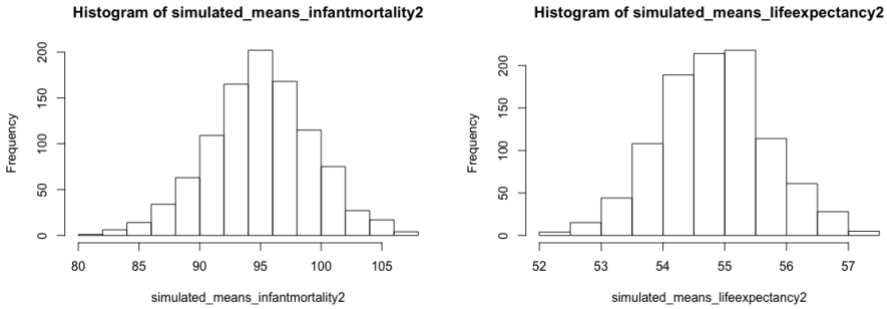
Sample Size

In this scenario, I also used the mean and standard deviation of infant mortality and life expectancy respectively. Then the simulation would be conducted in sample size 10, 100 and 1000 using the mean and standard deviation.

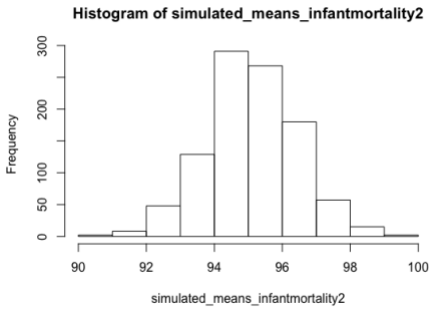
The first two histograms are the frequencies of means of 10 samples. The p-value of shapiro-wilk test is 0.2798 and 0.4758, which means the normality assumption accepted.

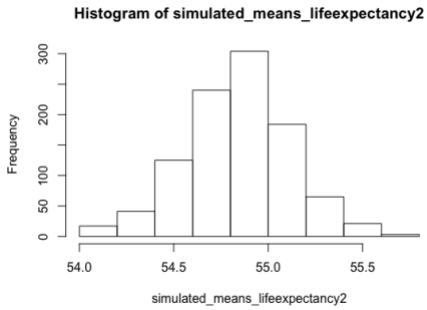


Then is the frequencies of means of 100 samples. The normality test shows p-values are 0.4049 and 0.8047, which are greater than 0.05.



The last two are the histogram of frequencies of simulated means of sample size 1000, which is the larger sample size compared to the previous two sizes. The p-value of these two are 0.7236 and 0.7663, which conform to normality.





The confidence intervals of correlation coefficient at 95% are as below. And also the p-value for each pairs of variables are greater than 0.05 shows that the two variables are not significantly correlated.

Sample Size	2.5% CI	97.5% CI	p-value of Pearson Correlation Test	p-values of Spearman Rank Correlation
10	-0.5450	0.7016	0.7227	0.3488
100	-0.0538	0.3312	0.1524	0.3128
1000	-0.0674	0.0566	0.8649	0.927

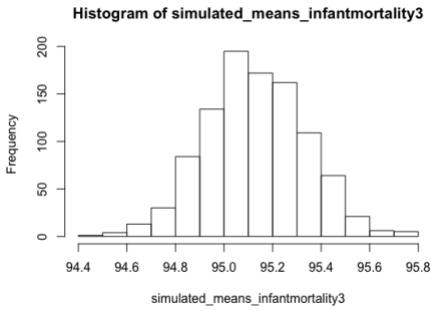
From the result, the variables with smaller samples has small effect while the larger samples only have trivial effect. From the figure, we can also find that powers of sample size 10 and 100 are even less than 0.1, which means I only have less than 10% change of finding a statistically significant difference. However, in sample size 100, the power is 30%.

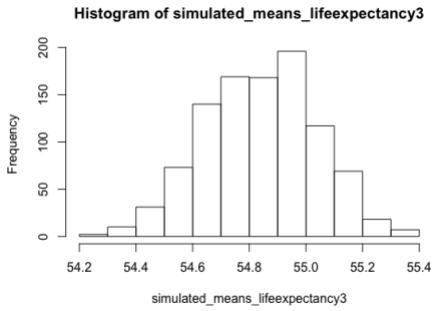
Sample Size	Effect Size	Power
10	0.1289	0.0638
100	0.1442	0.3004
1000	0.0053	0.0533

Standard Deviations

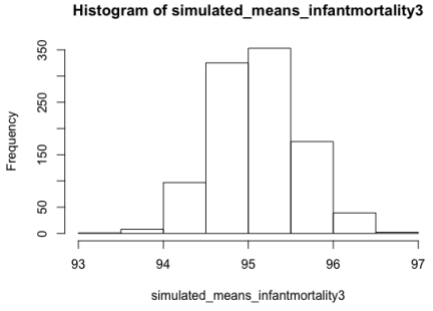
In this scenario, I only computed the mean of infant mortality and life expectancy respectively. Then the simulation would be conducted in with standard deviations of 2, 5 and 10 using the mean and sample size 10.

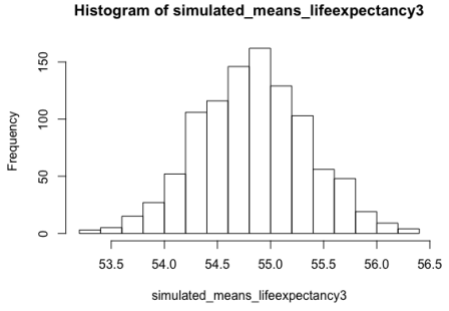
The following two plots are the frequencies of standard deviation of 2. The normality test show that p-value is greater than 0.05, which accept the distribution is normal distribution.



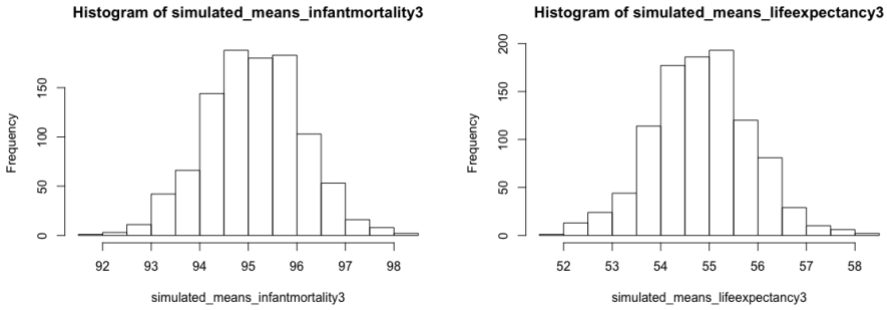


The next two distributions are the means of standard deviation of 5. The p-values are 0.2192 and 0.5311, which are also normal distribution.





The last two are the histogram of means of standard deviations of 10. The normality tests passed as the p-values are 0.9594 and 0.3916.



The confidence intervals of correlation coefficient for different standard deviations at 95% are as below. And also the p-value for each pairs of variables are greater than 0.05 shows that the two variables are not significantly correlated.

Standard Deviation	2.5% CI	97.5% CI	p-value of Pearson Correlation Test	p-values of Spearman Rank Correlation
2	-0.2079	0.1849	0.9062	0.9923
5	-0.2036	0.1892	0.9413	0.9228
10	-0.0610	0.3247	0.1739	0.0890

The effect size of each continent is the absolute value of correlation coefficient from Pearson Correlation Test. From the result, the variables with standard deviation 10 has small effect while others only have trivial effect. From the figure, we can see that powers of smaller standard deviations are even less than 0.1, which means I only have 10% or less change of finding a statistically significant difference. But with standard deviation 10, the power of data is 0.276.

Standard Deviation	Effect Size	Power
2	0.0119	0.0514
5	0.0075	0.0505
10	0.1371	0.2760

5 Conclusion

All in all, from all three different scenarios before, we can conclude that the life expectancy and infant mortality has almost no correlation. Further studies should focus on if any other factors in the original dataset has some correlation with life expectancy.

6 References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New Jersey: Lawrence Erlbaum.

Kassambara, A. (2018) ggpubr: 'ggplot2' based publication ready plots. https://CRAN.R-project.org/package=ggpubr/

R Core Team. (2015) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rafael A.I. (2018) dslabs: Data Science Labs. R package version 0.5.1. https://CRAN.R-project.org/package=dslabs/

Wickham, H., Francois, R., Henry, L., Muller, L. and R Studio (2018) dplyr: A Grammar of Data Manipulation. R package version 0.7.8. https://CRAN.R-project.org/package=dplyr/