Investigating Diabetes Mellitus by The Central Limit Theorem

Gehan Lyu*

Lancaster University, Lancaster, LA1 4YW, United Kingdom

*Corresponding Author E-mail: m18837125328@163.com

Abstract. In recent several years, Diabetes Mellitus (DM) has emerged as one of the most serious global healthcare problems. This project is to investigate DM by the Central Limit Theorem (CLT), which is one of the most important results in Probability Theory and Statistics and is the reason the normal distribution plays such a significant role. Firstly, use R to simulate the process of CLT and prove it by Normality Testing. Then investigate the probability of getting DM and the relationship between DM and Diabetic Complications.

Keywords: Central Limit Theorem (CLT), Diabetes Mellitus (DM), Probability Theory, Statistics.

1. Introduction

In 2000, it is considered that there were 171 million diabetics worldwide, and it will climb to 366 million by 2030. Besides, developing countries are more likely to have a rapid growth in the number of diabetics, where the number of people with DM is estimated to triple increase. Diabetes results in an increased risk of many complications such as chronic kidney disease (CKD) and cardiovascular disease [5].

This project focus on this global disease by the CLT. In 1733, Abraham de Moivre firstly came up with an early CLT, and Pierre-Simon Laplace consummated de Moivre's theorem and brought the CLT to bloom [1]. This theorem interprets that, under many conditions, independent random variables summed together will converge to a normal distribution as the number of variables increases. This statistical logical thinking is widely applied when we use sample statistics to estimate population parameters and the averages of samples will form a normal distribution. Based on this information, we can also estimate the probability of samples taking on extreme values that deviate from the population mean.

2. Theory

CLT and the corresponding theorems such as the Weak Law of Large Numbers (WLLN) will be stated as follows.

To investigate when averages of a sequence of independent identically distributed (i.i.d) random variables converge in probability to the expectation μ of each random variable in the sequence. Chebyshev's Inequality is needed so we will start from a more general result, Markov's Inequality.

Markov's Inequality [6]. If V is a non-negative random variable then for any a > 0,

$$P(V \geq a) \leq \frac{E[V]}{a}.$$

Proof. Let V be a continuous random variable and have density function f(v), then for any a > 0,

$$\begin{aligned} E[V] &= \int_0^\infty t f(t)\, dt \\ &\geq \int_a^\infty t f(t)\, dt \\ &\geq \int_a^\infty a f(t)\, dt \\ &= a \int_a^\infty f(t)\, dt \\ &= a\, P(V \geq a) \end{aligned}$$

Based on Markov's Inequality, for a random variable Y with E[Y] = μ and Var[Y] = σ^2, by $V = (Y - \mu)^2$ (so $E[V] = \sigma^2$) and $a = \epsilon^2$ gives

$$P((Y-\mu)^2 \geq \epsilon^2) \leq \frac{\sigma^2}{\mu^2}.$$

Chebyshev's Inequality [6]. If Y is a non-negative random variable with expectation μ and finite variance σ^2, then for any $\epsilon > 0$,

$$P(|Y-\mu|^2 \geq \epsilon^2) \leq \frac{\sigma^2}{\mu^2}.$$

Thus, the Weak Law of Large Numbers (WLLN) can be concluded as follows.

The Weak Law of Large Numbers [4]. Suppose X_1, X_2...is a sequence of i.i.d. random variables with expectation μ and finite variance σ^2, then for any $\epsilon > 0$,

$$P\left(|\overline{X}_n - \mu| \geq \epsilon\right) \to 0, \text{ as } n \to \infty$$

Proof. Using Chebyshev's Inequality, we have for any $\epsilon > 0$

$$P\left(|\overline{X}_n - \mu| \geq \epsilon\right) \leq \frac{1}{\epsilon^2} Var[\overline{X}_n] = \frac{1}{\epsilon^2}\frac{\sigma^2}{\mu^2}$$

The Central Limit Theorem [1]. Suppose X_1, X_2...is a sequence of i.i.d. random variables with expectation μ and finite variance σ^2, then for any $-\infty < x < \infty$,

$$P\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq x\right) \to \Phi(x), n \to \infty$$

where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\Phi(x)$ is the cumulative distribution function for the standard Normal distribution, $N(0,1)$, evaluated at x.

Whereas the WLLN only shows that $\overline{X}_n$converges to μ the CLT gives the stronger information that the deviations of $\overline{X}_n$ from μ scaled by $\sqrt{n}$ following an $N(0, \sigma^2)$ distribution in the limit. To use the CLT for reasonably large n we can assume that

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{\mu}\right)$$

$$S_n = \sum_{i=1}^{n} \overline{X}_i \sim N(n\mu, n\,\sigma^2)$$

3. Experiment and Results

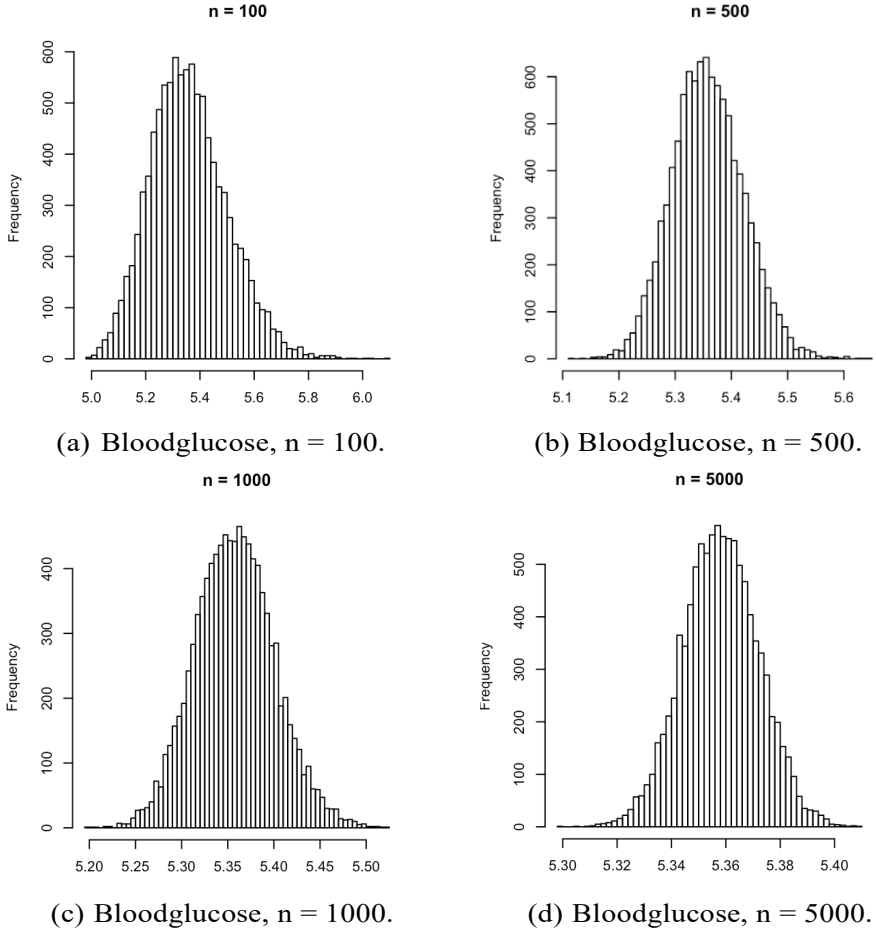
In this project, the experiment includes three parts. Firstly, we would like to use R to simulate the process of CLT and prove it by normality testing. Then we estimate the probability of getting DM by the index, blood glucose. Finally, to investigate the relationship between DM and diabetic complications, we split our data [7] into two categories, diabetics and nondiabetics, estimate the probability of getting diabetic nephropathy by the index, urea.

3.1 Normality Testing – Skewness and Kurtosis.

To simulate the process of CLT, we could process blood glucose data [7] in Rstudio, E[X] = 5.36 and Var[X] = 2.06, and randomly pick n samples (n =100, 500, 1000, 5000) form this distribution X_1, X_2, X_3... X_n. Calculate the mean $\overline{X}_n$of these n samples and then repeat this process m times to get $\overline{X}_1$, $\overline{X}_2$... $\overline{X}_m$. In this way, the means of these samples obey the Normal distribution.

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{\mu}\right)$$

Fig. 1 shows the results of the repeating process, which shows it is more and more approximate to a normal distribution a $n \to \infty$.



(a) Bloodglucose, n = 100. (b) Bloodglucose, n = 500.

(c) Bloodglucose, n = 1000. (d) Bloodglucose, n = 5000.

Figure 1. Bloodglucose.

In Statistics, normality tests are used to determine whether a data set is modeled for a normal distribution. To evaluate normality, skewness and kurtosis should be imported. Skewness is a measure of the asymmetry of the probability distribution of a random variable about its mean. Kurtosis illustrates the height and sharpness of the central peak, relative to that of a standard bell curve. If skewness is approaching 0 and kurtosis is approaching 3, the distribution is approximately symmetric, in other words, the data set is normally distributed [2].

Table 1. Skewness and Kurtosis.

n	X_{mean}	S_{mean}	X_{var}	S_{var}	*skewness*	*kurtosis*
100	5.36	5.36	2.06	0.0202	0.435	3.3
500	5.36	5.36	2.06	0.004	0.197	3.13
1000	5.36	5.36	2.06	0.00185	0.137	3.08
5000	5.36	5.36	2.06	0.000194	0.00196	3.01

According to the results from Table 1, skewness $\rightarrow$ 0, and kurtosis $\rightarrow$ 3, we can clearly realize that as $n \rightarrow \infty$, the distribution is more approximate to the normal distribution. Hence, we have verified the CLT by normality tests.

3.2 Calculations.

Based on Fig. 2, the criteria for diagnosis of diabetes, we could roughly deduce that the fasting plasma glucose (FPG) of diabetics is greater than 7.0 mmol/L [5], where fasting means no caloric intake for at least 8 hours. Our data collected satisfies all the requests. Excluding NA value, we totally get valid 9498 data. The frequency of this data set is shown in Fig. 3.

Table 1.3 Criteria for the diagnosis of diabetes

Any one of the following:

1. Symptoms of diabetes (polyuria, polydipsia, unexplained weight loss) plus random plasma glucose concentration 200 mg/dL (11.1 mmol/L).
2. FPG >126 mg/dL (7.0 mmol/L) (fasting = no caloric intake for at least 8 hours)
3. 2-hour plasma glucose 200 mg/dL during an oral glucose tolerance test (OGTT) (75 g)

In the absence of unequivocal hyperglycemia with acute metabolic decompensation, these criteria should be confirmed by repeat testing on a different day.

Figure 2. Criteria for diagnosis of diabetes [5]

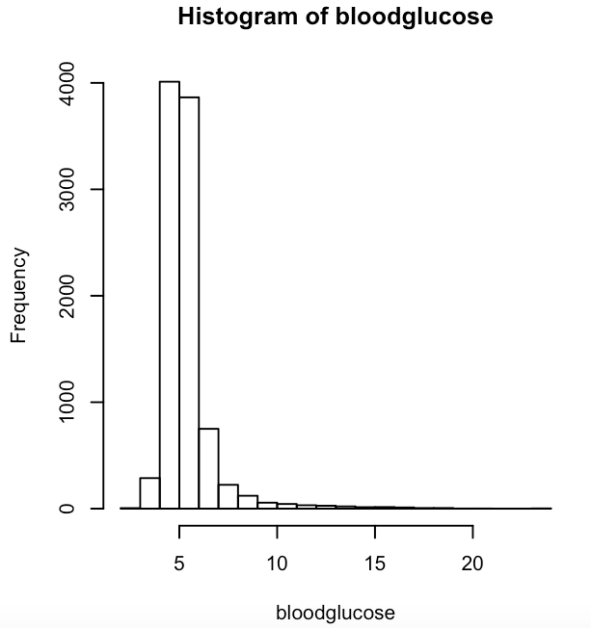


Figure 3. Frequency

Referring to CJFH Reference 1[7], the standard level of blood glucose is between 3.61 and 6.11 mmol/L. In this case, we intend to calculate the probability of our samples' average blood glucose is greater than 6.11 mmol/L. According to our previous results, E[X] = 5.36 and Var[X] = 2.06, we randomly choose 30 samples each time and the means of these samples obey the Normal distribution by the CLT.

$$\overline{X}_{30} \sim N\left(5.36, \frac{2.06}{30}\right)$$

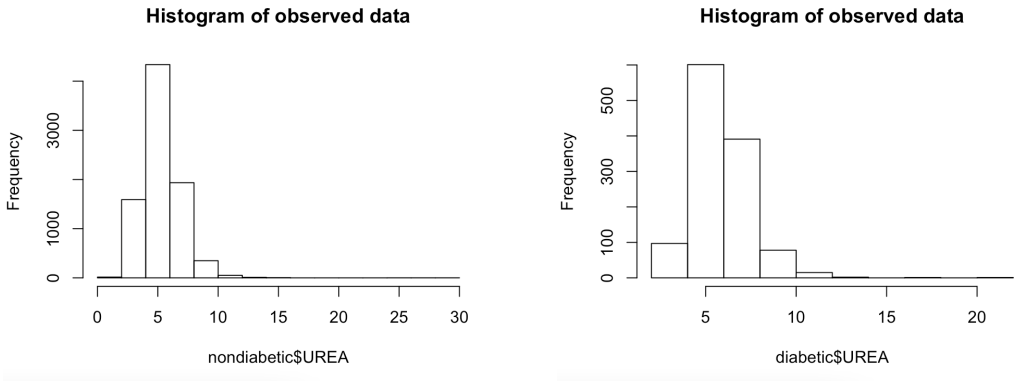
Calculate $P\left(\overline{X}_{30} > 6.11\right)$

$$\begin{aligned} P\left(\overline{X}_{30} > 6.11\right) &= 1 - P\left(\overline{X}_{30} \leq 6.11\right) \\ &= 1 - P\left(\frac{\overline{X}_{30} - 5.36}{\sqrt{\frac{2.06}{30}}} \leq \frac{6.11 - 5.36}{\sqrt{\frac{2.06}{30}}}\right) \\ &= 1 - \Phi(2.86) \\ &= 0.0021 \end{aligned}$$

3.3 Diabetic Complications.

As mentioned before, diabetes leads to a high risk of many complications such as chronic kidney disease (CKD) [5]. Urea is the end product of amino acid catabolism in the body. The determination of serum urea is one of the commonly used indexes to measure kidney function [3]. Previous investigations have shown higher blood urea nitrogen (BUN) was associated with DM. For this project, we will analyze the association between urea and the risk of incident DM by the CLT. Firstly, we split our data [7] into two categories, diabetics and nondiabetics. Secondly, estimate the probability of diabetics and nondiabetics getting diabetic nephropathy by the index, urea. Finally, we

could deduce the relationship between DM and diabetic complications. Fig.4 illustrates histograms of diabetics and nondiabetics.



(a) Nondiabetic. (b) Diabetic.

Figure 4. Histogram of observed data.

Referring to CJFH Reference 1 [7], the standard level for urea is between 2.78-7.85 mmol/L.

For diabetics, E[Y] = 5.85 and Var[Y] = 2.56, we randomly select 30 samples each time and the means of these samples obey the Normal distribution by the CLT.

$$\overline{Y}_{30} \sim N\left(5.85, \frac{2.06}{30}\right)$$

Calculate $P\left(\overline{Y}_{30} > 7.85\right)$

$$\begin{aligned} P\left(\overline{Y}_{30} > 7.85\right) &= 1 - P\left(\overline{Y}_{30} \leq 7.85\right) \\ &= 1 - P\left(\frac{\overline{Y}_{30} - 5.85}{\sqrt{\frac{2.56}{30}}} \leq \frac{7.85 - 5.85}{\sqrt{\frac{2.56}{30}}}\right) \\ &= 1 - \Phi(6.85) \\ &= 0.00000000000378 \end{aligned}$$

For nondiabetics, E[Z] = 5.31 and Var[Z] = 2.53, we repeat the same process for diabetics and the means of these samples obey the Normal distribution by the CLT.

$$\overline{Z}_{30} \sim N\left(5.31, \frac{2.53}{30}\right)$$

Calculate $P\left(\overline{Z}_{30} > 7.85\right)$

$$\begin{aligned} P\left(\overline{Z}_{30} > 7.85\right) &= 1 - P\left(\overline{Z}_{30} \leq 7.85\right) \\ &= 1 - P\left(\frac{\overline{Z}_{30} - 5.31}{\sqrt{\frac{2.53}{30}}} \leq \frac{7.85 - 5.31}{\sqrt{\frac{2.53}{30}}}\right) \\ &= 1 - \Phi(8.75) \\ &\approx 0 \end{aligned}$$

Rstudio cannot distinguish the numbers less than e^{-16}. Thus, even though $P\left(\overline{Y}_{30} > 7.85\right)$ is small, it is quite greater than $P\left(\overline{Z}_{30} > 7.85\right)$. Therefore, there is a relationship between DM and diabetic complications.

4. Discussion

4.1 Data.

The data used was picked up through professional procedure and all participants have requested no caloric intake for as least 8 hours. This data set was collected in 2009 so it seems a little bit out of

date. Besides, the China Health and Nutrition Survey (CHNS) [7] surveyed several provinces from china but the data is still bias to estimate population parameters. That is the reason why the CLT is considered.

4.2 Limitation and Further Study

The main limitation of this project is that the significant increase of urea will appear in the later period of chronic kidney disease (CKD). Thus, $P(\overline{Y}_{30} > 7.85)$ is so small and close to 0. Another reason our results are not so obvious is that we are applying the CLT which requires us to calculate the mean of n samples. When $n \to \infty$, it is more likely the average is close to the expectation.

For further study, there are majority factors of DM such as obesity, which is worthwhile to investigate their relevance.

5. Conclusion

In conclusion, this project has simulated the process of CLT and prove it by normality testing. Then we have estimated the probability of having DM by the index, FPG. Finally, the relationship between DM and diabetic complications has been shown through comparison.

References

[1] William J Adams. The life and times of the central limit theorem, volume 35. Sci. American Mathematical Soc., 2009.

[2] Roser Bono, Jaume Arnau, Rafael Alarc´on, and Maria J Blanca. Bias, precision, and accuracy of skewness and kurtosis estimators for frequently used continuous distributions. Symmetry, 12(1):19, 2020.

[3] Pei Feng, Guangli Wang, Qian Yu, Wei Zhu, and Chongke Zhong. First-trimester blood urea nitrogen and risk of gestational diabetes mellitus. Journal of cellular and molecular medicine, 24(4):2416– 2422, 2020.

[4] Hans Fischer. A history of the central limit theorem: from classical to modern probability theory. Springer Science & Business Media, 2010.

[5] Vivan Fonseca. Diabetes: improving patient care. Oxford University Press, USA, 2010.

[6] BK Ghosh.Probability inequalities related to markov's theorem. The American Statistician, 56(3):186–190, 2002.

[7] Shengkai Yan, Jiang Li, Shuang Li, Bing Zhang, Shufa Du, Penny Gordon-Larsen, Linda Adair, and Barry Popkin. The expanding burden of cardiometabolic risk in china: the china health and nutrition survey. Obesity Reviews, 13(9):810–821, 2012.