About Normality Distribution of Age in Patients with Diabetes Mellitus

Ioan GOLET, Mircea Miron JIVULESCU, and Maria Anastasia JIVULESCU *

Dept. of Mathematics, "Politehnica" University of Timişoara, Victoriei Square, No. 2, 300006, Timişoara, Romania.
E-mails: ioan.golet@mat.upt.ro; mircea.jivulescu@gmail.com; maria.jivulescu@mat.upt.ro

* Author to whom correspondence should be addressed; Tel.: +40256403098; Fax: +40256403099

Received: 28 November 2011 /Accepted: 29 February 2012 / Published online: 10 March 2012

Abstract

The age variable from a sample of patients with diabetes mellitus has been analyzed with the iteratively applied method of moving average. Graphically, we found a normal frequency distribution of patients' age. We were concerned by confirming the normality distribution of age, due to its importance, since many statistical tests require the normality as a prerequisite. We tested the above graphical conclusion using a concordance test (Kolmogorov-Smirnov).

Keywords: Diabetes mellitus; Normal distribution; Biostatistics.

Introduction

Over the last few years studies on the diabetic disease have increased proportionally to the number of people affected by this disease [1-3]. For example, The Department of Health and Human Services, USA, reported in 2008 that 68% of the adult incident cases (i.e. cases diagnosed within the past year) of diabetes melitus were diagnosed between the age of 40 and 64 years. About 15% were diagnosed before the age of 40 and about 17% were diagnosed at age 65 or older [4-5]. Regarding the prevalence of diabetes melitus among the Romanian population the WHO studies [6] from 2007 estimated that around 5% of Romanians have the disease and 1% of them are under 35. The projections indicate that in 2030 the percentage will increase to 6.22% of the entire population. The study we conducted analyzed a sample of patients affected by diabetes melitus and focused on understanding the age when the disease was firstly diagnosed.

The aim of our study has been to verify the normal distribution of age in patients with diabetes melitus from Timis County, Romania.

Material and Methods

Data was anonymously collected from general practitioners in Timiş County, Romania. The dataset included the gender, age and disease variables for 356 patients diagnosed with diabetes melitus.

The histogram of age in this dataset has been plotted, using SPSS (PASW Statistics version 18) package.

Susequently, the age variable was analyzed using the method of ***moving average*** [7]. This method is primarily for the smoothing of time series, in which each observation is replaced by a weighted average of the observation and its near neighbors. Moving averages are often used to eliminate the seasonal variation or cyclic variation from time series and hence to emphasize the trend terms. The same method can be iteratively applied to the age variable X_j from our dataset. In this way it is possible to create new series of data Y_j of the form $Y_j = \sum_{k=-1}^{k=1} X_{j+k}/3, j = 2\ldots354$

After, six iterations, we concluded that, graphically, the result was a normal frequency distribution of the age of patients. This result was then tested for normality. For this aim, according to theory, for large sample sizes it is advised to use the Kolmogorov-Smirnov test instead of the Shapiro-Wilk test.

The Kolmogorov-Smirnov test is a goodness-of-fit test for many statistical distributions [8-9]. The test relies on the fact that the value of the sample cumulative density function is asymptotically normally distributed. To apply the Kolmogorov-Smirnov test, it is necessary to calculate the cumulative frequency (normalized by the sample size) of the observations as a function of class. Then, to calculate the cumulative frequency for a true distribution (most commonly, the normal distribution). The following step is to find the greatest discrepancy between the observed and expected cumulative frequencies, which is called the "D-statistic" [10] and to compare it against the critical D-statistic for that sample size. If the calculated D-statistic is greater than the critical one, then the null hypothesis that the distribution is of the expected (normal) form is rejected. For our case, we formulate the following hypothesis regarding the distribution of the sample, which is denoted by F(x).

$$H_0: F(x) = F_0(x) \text{ for all } x$$

$$H_A: F(x) \neq F_0(x) \text{ for all } x$$

Here, $F_0 = F_0(x) = \Phi[(x-\mu)/\sigma]$, and μ, σ are default sample mean and respectively the standard deviation.

The empirical cumulative distribution function (CDF) can be constructed algorithmically [11], by following the subsequent steps. First, the observations are sorted into ascending order: $x_1 < x_2 < \ldots < x_m$, where m is the number of distinct values of *X*. The the empirical CDF becomes:

$$\hat{F}(x) = \begin{cases} 0, & -\infty < x < x_1 \\ \dfrac{\sum_{i=1}^{m} f_i I(x < x_{k+1})}{\sum_{i=1}^{n} f_i}, & x_k < x < x_{k+1}, k = 1,\ldots,m-1 \\ 1, & x_m \le x < \infty \end{cases}$$

where I is the characteristic function, $I_A(x) = \begin{cases} 1, x \in A \\ 0, \text{otherwise} \end{cases}$.

The test statistic is calculated based on differences between the empirical cumulative distribution and the theoretical cumulative distribution. For each *i = 1, …, m*, we denote by

$D_i := \hat{F}(x_{i-1}) - F_0(x_i), \hat{D}_i := \hat{F}(x_i) - F_0(x_i)$.

The test statistic is:

$$Z = \sqrt{\sum_{j=1}^{n} f_j} \max_i \{|D_i|, |\hat{D}_i|\}$$

The two-tailed probability level is estimated using the first three terms of the Smirnov (1948) [10] formula.

$$p = \begin{cases} 1, & 0 \le Z < 0.27 \\ 1 - \dfrac{\sqrt{2\pi}}{Z}(Q + Q^9 + Q^2 5), Q = e^{\frac{-\pi^2}{8Z^2}} \quad x_k < x_{k+1}, & 0.27 \le Z < 1 \\ 2(Q - Q^4 + Q^9 - Q^{16}), Q = e^{-2Z^2}, & 1 \le Z < 3.1 \\ 0, & Z \ge 3.1 \end{cases}$$

The case when $p < \alpha$ rejects the null hypothesis.

Results and Discussion

The youngest patient diagnosed is 16 years old and the oldest is 90. The mean age is 62.45 years, and the standard deviation is 12.27. The histogram of the dataset presented in Figure 1 has the peak around 60 years old, while the rest of the data is almost symmetrical distributed around it. Intuitively, we may conclude that the age variable is distributed normally around this age. Our result is strengthened also by a good matching with the superimposed bell-shaped curve.

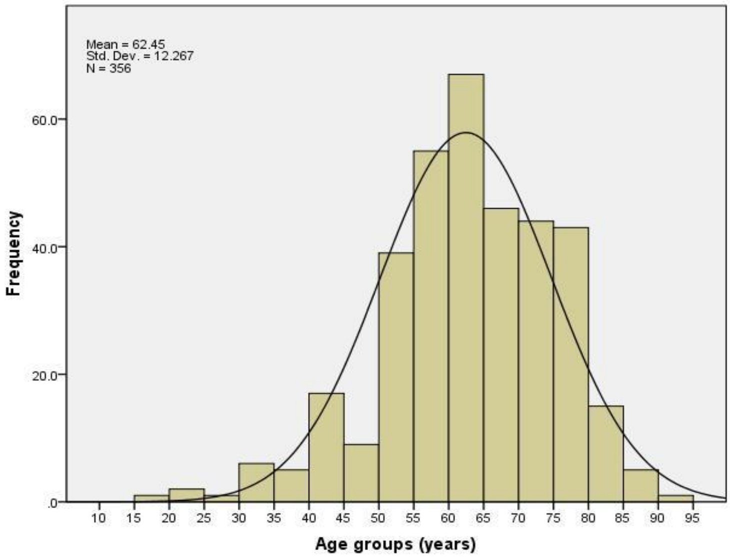

Figure 1. Histogram of age distribution superimposed by the normal distribution

The first iteration of the moving average method preserved the non-homogeneity of the dataset (Figure 2), so further iterations were required in order to smoothen the distribution. As it can be seen in the Figure 3, the sixth iteration produces a distribution roughly similar with the Gaussian distribution.

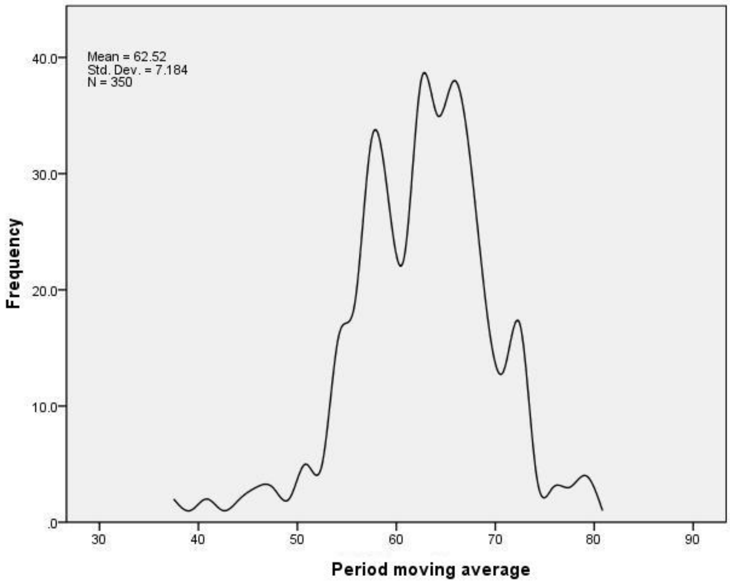

Figure 2. First iteration for the moving average method with a span of 3

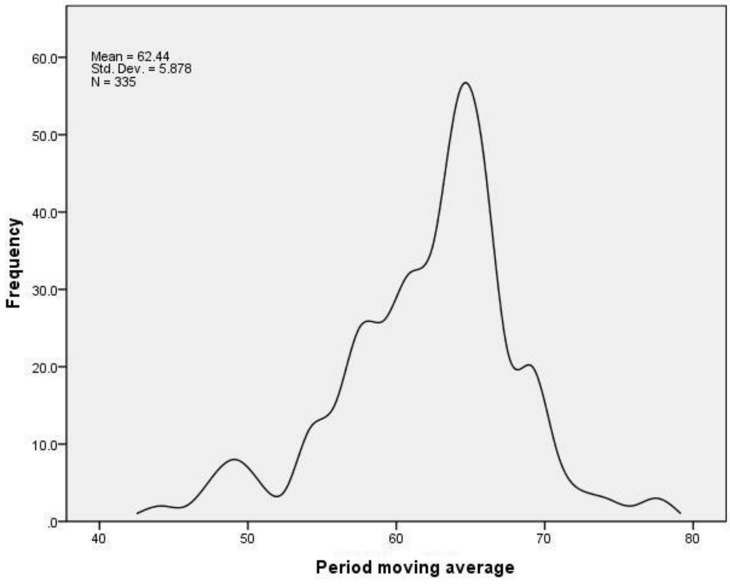

Figure 3. The sixth iteration for the moving average method with a span of 3

By applying the above steps, the two-tailed asymptotic significance value obtained for our dataset is 0.187. A significance value greater tha 0.05 indicates normality of the distribution.

The P-P (Probability-Probability plot or percent-percent plot) and Q-Q plot (Quantile-Quantile) for age in the investigated sample are represented in Figures 4 and 5, respectively.

As can be seen from Figure 4 the cumulative distribution functions represented graphically against each other certify that our data set agrees with a normal distribution. Also, the representation of the quantiles (Figure 5) of the two probability distributions against each other emphasize this similarity.

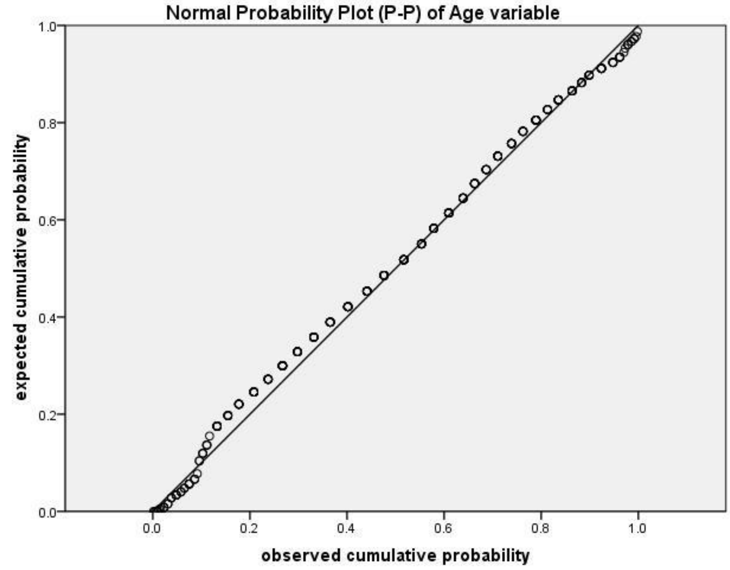

Figure 4. Probability plot of age variable distribution

Considering the obtained histogram of age, we might intuitively conclude that the age variable is normally distributed around a mean value of 60 years. This result would also be strengthened by a good matching with the superimposed bell-shaped curve. Still, further investigation of normality in any given dataset is highly advisable.

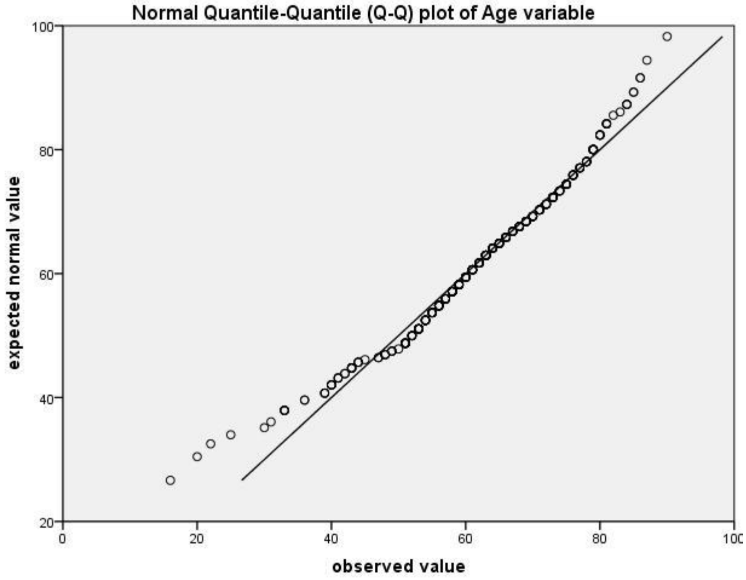

Figure 5. Quantile-Quantile (Q-Q) plot of age variable distribution

By applying the Kolmogorv-Smirnov test, we should confirm that the age variabe has a normal distribution, which should then be reconfirmed using a Q-Q plot.

Conclusions

The article presents a study on assessing normal distribution of the age variable in persons affected by diabetes mellitus. A normal distribution can be intuitively noted by simple methods such as heuristical observation of histogram and/or moving average method. The statistical software SPSS also allows for a rigorous proof of our assumption by performing more sophisticated methods such as the Kolmogorov-Smirnov test or Q-Q plot.

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgements

The authors would like to thank to the organizers of the conference “Applied Biostatistics in Life Sciences. Joint Meeting of ISCB National Groups and IPA Cross-border Co-operation Program HU-SRB/0901/221/088”, Nov. 4-5 2011, Szeged, Hungary, where this paper was presented as poster.

References

1. Aubert R. Diabetes in America, 2nd ed. DIANE Publishing: Cowie CC, Eberhardt MS; 1995. Chapter 6, Sociodemographic Characteristics of Persons with Diabetes; p. 85-90.
2. Wild S, Roglic G, Green A, Sicree R, King H. Global Prevalence of Diabetes. Diabetes Care 2004;27(5):1047-1053.

3. Australian Institute of Health and Welfare 2011. Diabetes prevalence in Australia: detailed estimates for 2007–08. Diabetes series no. 17. Cat. no. CVD 56. Canberra: AIHW.
4. Centers for Disease Control and Prevention, Atlanta, USA. Distribution of Age at Diagnosis Among Adult Incident Cases Aged 18-79 Years, United States [Internet]. 2008 [updated 2010 Feb 5; cited 2011 October 3]. Available from: http://www.cdc.gov/diabetes/statistics/age/fig1.htm
5. Cowie CC, Rust KF, Byrd-Holt DD, Eberhardt MS, Flegal KM, Engelgau MM, Saydah SH, Williams DE, Geiss LS, Gregg EW. Prevalence of Diabetes and Impaired Fasting Glucose in Adults in the U.S. Population. Diabetes Care June 2006;29:1263-8.
6. WHO Media centre. Diabetes Fact sheet N°312 [Internet] [updated 2011 August; cited 2011 October 3]. Available from: http://www.who.int/mediacentre/factsheets/fs312/en/
7. Dodge Y. The Concise Encyclopedia of Statistics. Springer & Business Media, LLC. 2008; 283-287.
8. Morris H. DeGroot Probability and statistics. Second edition. Addison-Wesley Pub Co, 1986; 552-560.
9. Riffenburgh RH. Statistics in Medicine, 2nd ed., Elsevier Academic Press 2006;371-374,565.
10. Smirnov NV. Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics 1948;19:279-281.
11. Mood AM, Graybill FA, Boes DC. Introduction to the theory of statistics, 3rd ed., McGraw-Hill, 1974; 506-518.