Deneysel Araştırma

Normal Dağılıma Uygunluğu Değerlendirmek için Açık Kaynak Web Tabanlı Yazılım: Normal Dağılımı İnceleme Yazılımı

Ahmet Kadir ARSLAN[1,a], Zeynep TUNÇ[1], Cemil ÇOLAK[1]

İnönü Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıp Bilişimi Anabilim Dalı, Malatya, Türkiye

ÖZET

Amaç: Bu araştırmada tek ve çok değişkenli normal dağılıma uygunluğu kolay bir şekilde test edebilecek ve kullanıcıların yapacakları çalışmalarında daha doğru sonuçlar elde etmesini sağlayacak yeni kullanıcı dostu web tabanlı yazılım geliştirmek amaçlanmıştır.

Gereç ve Yöntem: Açık kaynak kodlu bir R paketi olan Shiny, önerilen web yazılımını geliştirmek için kullanıldı. Geliştirilen yazılımda tek değişkenli normal dağılıma uygunluk için Shapiro-Wilk ve Anderson-Darling testleri, çok değişkenli normal dağılıma uygunluk için ise Mardia'nın çarpıklık-basıklık, Henze-Zirkler ve Doornik-Hansen testleri kullanıldı. Normal dağılıma uygunluk için verilen çıktılarda test istatistiklerine ilaveten grafiksel yöntemler de sunulmuştur. Örnek uygulama olarak simulasyon ile türetilen iki değişkenli, her bir değişkenin standart normal dağılıma sahip olduğu ve 1000 gözlemli veri seti için normal dağılıma uygunluk test edilerek bulgular değerlendirilmiştir.

Bulgular: Türetilen veri setinde her bir değişken Anderson-Darling ve Shapiro-Wilk testlerine göre normal dağılmıştır (sırasıyla x_1 ve x_2 değişkenleri için p =0.91 ve p =0.707; p =0.756 ve p =0.573). Ayrıca türetilen veri seti Mardia'nın çarpıklık-basıklık, Henze-Zirkler ve Doornik-Hansen testlerine göre iki değişkenli normal dağılım göstermiştir (respectively p =0.826, p =0.831 and p =0.868).

Sonuç: Geliştirilen yazılım tek değişkenli ve çok değişkenli normal dağılıma uygunluk analizlerini kolayca yapabilen ve kullanıcıların yapacakları çalışmalarında daha doğru sonuçlar elde etmesini sağlayan yeni kullanıcı dostu bir web tabanlı yazılımdır. İlerleyen çalışmalarda, en iyi yönteme karar vermede kullanılan ölçütlerden Tip I ve Tip II hata türlerinin yazılıma eklenmesi planlanmaktadır.

Anahtar Sözcükler: *Normal Dağılıma Uygunluk, Simülasyon, Web Tabanlı Yazılım.*

ABSTRACT

Open Source Web Based Software to Evaluate Normal Distribution: Normality Assessment Software

Objective: In this study, it was aimed to develop a new user-friendly web-based software that would easily test single-variable univariate and multivariate normal distribution suitability and enable users to get more accurate results in their studies.

Material and Method: Shiny, an open source R package, was used to develop the proposed web software. In the developed software, Shapiro-Wilk and Anderson-Darling tests were used for the uniformity of univariate distribution, and Mardia's skewness-kurtosis, Henze-Zircon and Doornik-Hansen tests were used for multivariate normal distribution. Outputs for conformity to normal distribution were supported by using graphical methods. In practice, for the data set where each variable consisting of two variables derived by simulation has a standard normal distribution and the variables contain 1000 observations, the normal distribution conformity analysis has been performed.

Results: In the derived data set, each variable is normally distributed according to the Anderson-Darling and Shapiro-Wilk tests. (for x1 and x2 variables, respectively p =0.91 and p =0.707; p =0.756 and p =0.573). In addition, the derived data set showed two-variable normal distribution according to Mardia's skewness-kurtosis, Henze-Zircon and Doornik-Hansen tests.

Conclusion: The developed software is a new user-friendly web-based software that can easily perform univariate and multivariate normal distribution conformity analysis and enable users to get more accurate results in their work. In further studies, Type I and Type II error types are planned to be included in the software in order to determine the best method.

Keywords: *Normal Distribution, Simulation, Web-Based Software.*

Bu makale atıfta nasıl kullanılır: Arslan AK, Tunç Z, Çolak C. Normal Dağılıma Uygunluğu Değerlendirmek için Açık Kaynak Web Tabanlı Yazılım: Normal Dağılımı İnceleme Yazılımı. Fırat Tıp Dergisi 2020; 25 (2): 62-68.

How to cite this article: Arslan AK, Tunc Z, Colak C. Open Source Web Based Software to Evaluate Normal Distribution: Normality Assessment Software. Firat Med J 2020; 25 (2): 62-68.

Verileri analiz etmede ilk olarak önemlilik/hipotez testlerinin varsayımlarının sağlanıp sağlanmadığı kontrol edilmelidir. Eğer veri seti dağılımsal varsayımlara sahip değilse, bu varsayımlar altında yapılacak olan tüm istatistiksel analizler geçersiz olacaktır ve çoğu zamanda yanlış sonuçlar elde edildiği için yanlış yorumlar elde etmemize neden olacaktır (1).

Güçlü olarak kabul edilen çoğu istatistiksel testler veri setinin normal dağılıma uygun olması varsayımı altında çalışmaktadır. Bu nedenle istatistiksel analizlere başlanmadan veri setinin normal dağılıp dağılmadığı araştırılmalıdır. Bu nedenle normal dağılım istatistik alanında en çok üzerinde durulan konulardan biri oluştur. De Moivre tarafından 1733 yılında temelleri binom dağılımın limit şekli olarak atılmış olsa da 1800'lü yılların başlarında Laplace ve Gauss tarafından geliştiri-

[a]Yazışma Adresi: Ahmet Kadir ARSLAN, İnönü Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıp Bilişimi Anabilim Dalı, Malatya, Türkiye
Tel: 0422 341 0660 e-mail: arslan.ahmet@inonu.edu.tr
Geliş Tarihi/Received: 02.08.2019 Kabul Tarihi/Accepted: 07.11.2019
*Bu çalışma IEEE 3. Uluslararası Multidisipliner Araştırmalar ve Yenilikçi Teknolojiler Sempozyumu'nda sunulmuştur (11-13 Ekim 2019, Radisson Blu Hotel, Ankara).

rilerek son halini almıştır. Genel olarak normal dağılım, veri setinde yer alan bağımlı değişken sayısına göre 'tek değişkenli normal dağılım' ve 'çok değişkenli normal dağılım' olmak üzere 2 kategoride sınıflandırılmaktadır. Tek değişkenli normallik varsayımında örneklemde yer alan her bir sürekli değişkene ait ölçümlerin normal dağılımdan gelmesi varsayımı vardır (2, 3). Sürekli tek bir değişkenin normalliği tanımlayıcı, grafiksel ve istatistiksel yöntemlerle incelenebilir (4).

Bağımlı değişkenlerin sürekli değişken olduğu tek ve çok değişkenli parametrik testler tek değişkenli normallik varsayımına, çok değişkenli parametrik testler ise ayrıca çok değişkenli normallik varsayımlarına dayandırılır (5, 6). Birden çok değişkenin aynı andan normallik koşulunu sağlaması anlamına gelen çok değişkenli normal dağılım, örneklemdeki her bir değişkenin tek değişkenli normallik şartını sağlamasına ek olarak değişkenlerin tüm doğrusal kombinasyonlarının ve her türlü elde edilebilen ikili kombinasyonlarının da her bir grup ve alt grup için normal oluğunu varsaymaktadır (6). Tek değişkenli normallik varsayımının incelenmesinde olduğu gibi çok değişkenli normal dağılımın varlığının araştırılması için de betimsel, grafiksel ve istatistiksel metotlar bulunmaktadır. Çok değişkenli normallik varsayımlarını incelemek için betimsel yöntem olarak çok değişkenli çarpıklık ve basıklık katsayıları kullanılabilmektedir (7). Grafiksel yöntem olarak çoklu saçılma grafiğinin yanı sıra marjinal grafikler, Kowalski'nin çizgi grafikleri çok değişkenli Q-Q grafikleri tercih edilebilir (5). Ayrıca tek değişkenli uyum iyiliği testlerinden olan Cramer-von Mises (8) testinden elde edilmiş Koziol's testi (9), Shapiro-Wilk elde edilmiş türetilmiş Mudholkar-Srivastava-Lin testi (10), Anderson-Darling testinden elde edilmiş Paulson-Roohan-Sullo testi (11), ki-kare testi ve Kolmogorov-Smirnov testinden elde edilmiş bazı testler kullanılabilmektedir (12).

Araştırmalarda parametrik testler tercih edilmesine rağmen çok az sayıdaki çalışmalarda tek değişkenli ve bilhassa çok değişkenli normallik koşullarının gerçekleşip gerçekleşmeme durumunun kontrol edildiği sonucuna ulaşılmaktadır (13). Bu araştırma kapsamında çok değişkenli normal dağılıma uygunluğu kolay bir şekilde kontrol edecek ve kullanıcıların yapacakları çalışmalarında daha doğru sonuçlar elde etmesini sağlayacak yeni kullanıcı dostu bir web tabanlı yazılım geliştirmek amaçlanmıştır.

GEREÇ VE YÖNTEM

Veri seti

Bu çalışmada yapılan yazılımın çalışma şeklini gösterebilmek ve çıktılarını değerlendirebilmek adına iki değişkenden oluşan her bir değişkenin standart normal dağılıma sahip olduğu ve değişkenlerin 1000 gözlem içerdiği veri seti IBM SPSS Statistics sürüm 25.0'ın model sekmesi (simülasyon) ile türetilmiştir.

Normallik analizleri

Bu web tabanlı yazılım araştırıcıların sahip olduğu veri setlerini tek ve çok değişkenli normal dağılıma uygunluğunu test etmek amacıyla geliştirilmiştir. Geliştirilen yazılımda tek değişkenli normal dağılıma uygunluk Shapiro-Wilk ve Anderson-Darling testleri ile test edilirken, çok değişkenli normal dağılıma uygunluk ise Mardia'nın çarpıklık-basıklık, Henze-Zirkler ve Doornik-Hansen testleri ile test edilmektedir. Bu testler dışında yer alan ve normallik sınamasında kullanılabilen grafiksel metotlardan olan Q-Q, çekirdek yoğunluk kestirim ve 3-boyutlu yüzey grafikleri de araştırmacıların kullanımı amacıyla yazılımda yer almaktadır. Ayrıca geliştirilen yazılım ile veri seti yapısına bağlı olarak veride nitel değişken(ler)in varlığı durumunda alt grup analizlerinin yapılmasına olanak sağlanmıştır.

Shapiro-Wilk Testi

Normallik testlerinden olan Shapiro-Wilk testi ilk defa Samuel Shapiro ve Martin Wilk tarafından 1965 yılında yayınlanmıştır (14). Bu test için kurulacak sıfır hipotezi bir örneklemden elde edilen veri serisinin normal dağılım gösteren bir topluma ait olmasıdır. Yani $x_1, x_2, \dots, x_n$ serisinin normal dağılım gösteren bir topluma ait olmasıdır.

Shapiro-Wilk test istatistiği olan W şu şekilde elde edilir:

Küçükten büyüğe doğru sıralanmış $x_1, x_2, \dots, x_n$ için normal dağılım değerleri olan a_i değerleri şöyle bulunur.

$$(a_1, a_2, \dots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

Burada;

$m^T = (m_1, m_2, \dots, m_n)^T$

olup $m_1, m_2, \dots, m_n$ standart normal dağılımdan örneklem olarak bulunmuş bağımsız ve aynı dağılımı gösteren rastgele değişkenlerin sıra istatistikleri değerlerine ait beklenen değerlerdir. *V* ise bu sıra istatistikleri için elde edilen kovaryans matrisidir. Son olarak W istatistiği ise

$$W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

olarak elde edilir (14).

W test istatistiği $0 < W \leq 1$ aralığında değişim göstermektedir. W test istatistiğinin 1'e yakın değerler alması değişkenin normal dağılıma sahip olduğunu, 0'a yakın değerler alması ise değişkenin normal dağılıma sahip olmadığını gösterir (14).

Anderson-Darling Testi

Anderson-Darling testi deneysel dağılım fonksiyonu istatistiklerine dayandırılarak 1974 yılında Anderson ve Darling tarafından geliştirilen bir testtir. Bu testte uygulanacak verilerin ham veri olması gerekir. Anderson-Darling testi gözlenen birikimli dağılım fonksiyonunun beklenen birikimli dağılım fonksiyonuna uyumunu karşılaştırabilmek için kullanılan bir testtir (14).

Anderson-Darling testi büyüklük sırası gösteren x_i gözlemlerinin deneysel yığılımlı olasılık dağılımının yığılımlı standart normal dağılıma uygun olup olmadığını test eder. Bu işlem için her bir gözlemin standart dönüşümleri yapılarak z_i gözlemleri elde edilir. Böylece elde edilen standart z_i değerlerinin olasılıklarının standart normal dağılıma uygunluğu test edilir. Bu nedenle Anderson-Darling testinden bir A^2test istatistiği hesaplanarak verilerin normal dağılıma uygunluğu test edilir. A^2 Anderson-Darling test istatistiği

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln\left(F_0(Z_i)\right) + \ln\left(F_0(Z_{n-1+i})\right)\right]$$

olarak elde edilir.

Burada yer alan $F_0(Z_i)$ değeri; standart normal dağılımın Z_i noktasındaki yığılımlı olasılık değerini, n ise örnek gözlem sayısını göstermektedir (14). A^2 test istatistiği küçük örneklem genişliğine sahip veri setleri için düzeltilerek kullanılır. Düzeltilmiş A^2 test istatistiği ise

$$A^2_{Düz} = A^2\left(1 + \frac{0{,}75}{n} + \frac{2{,}25}{n}\right)$$

ile elde edilir (14).

Mardia' nın çarpıklık-basıklık testi

Mardia tarafından önerilen ve çok değişkenli çarpıklık ve basıklık testleri, tek değişkenli çarpıklık ve basıklık ölçümlerini içeren t istatistiğinin gücünü genişletme çalışmaları ile 1970 yılında geliştirilmiştir. Ortalama değeri μ ve kovaryans matrisi $\in$ olan p değişkenli bir evrende X için Mardia evrene ait çok değişkenli çarpıklık ve basıklık ölçümlerini aşağıdaki gibi vermiştir (15, 16).

$$\beta_{1,p} = E\{(X-\mu)'\in^{-1}(Y-\mu)\}^3$$
$$\beta_{2,p} = E\{(X-\mu)'\in^{-1}(X-\mu)\}^2$$

Burada X ve Y bağımsız olarak özdeş olarak dağılır.

Çok değişkenli çarpıklık ve basıklık katsayıları sırasıyla $(\hat{\gamma}_{1,p})$ ve $(\hat{\gamma}_{2,p})$ katsayılarına bağlı olarak hesaplanır. $(x_1, x_2, \ldots, x_n)$ verisinin ortalama değeri μ ve kovaryans matrisi $\in$ olan p değişkenli bir evrenden rastgele alınarak elde edilen bir örnek olduğu kabul edilirse, buna ait çok değişkenli çarpıklık ve basıklık katsayıları;

$$b_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{(x_i-\bar{x})'S^{-1}(x_j-\bar{x})\right\}^3$$

$$b_{2,p} = \frac{1}{n^2}\sum_{i=1}^{n}\{(x_i-\bar{x})'S^{-1}(x_i-\bar{x})\}^2$$

şeklinde elde edilir. Burada;

$$S = \sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'/n$$

dir.

Elde edilen katsayılara bağlı olarak çarpıklık ve basıklık testi için sırasıyla elde edilecek test değerleri

$$A = \frac{nb_{1,p}}{6} \sim \chi^2_F$$

olup F= $\frac{1}{6}p(p+1)(p+2)$ dir.

$$B = \frac{b_{2,p} - p(p+2)}{\{8p(p+2)/n\}^{1/2}} \sim N(0,1)$$

ile hesaplanır (16).

Henze-Zirkler

Henze-Zirkler testi Epps ve pulley tarafından 1983 yılında geliştirilen tek değişkenli normalliğin belirlenmesi için önerilen testin çok değişkenli normal dağılım durumu için genelleştirilmiş halidir. Ayrıca 1988 yılında Baringhaus ve Henze tarafından önerilen testin de genel halidir (17).

Çok değişkenli normal dağılımı belirlemek için kullanılan Henze-Zirkler test istatistiği düzleştirme parametresi olan β nın bir fonksiyonu olup

$$\beta = \beta_p(n) = \frac{1}{\sqrt{2}}\left(\frac{2p+1}{4}\right)^{\frac{1}{p+4}} n^{\frac{1}{p+4}}$$

dir.

Henze-Zirkler test istatistiği ise;

$$T_\beta(n) = n\left[\frac{1}{n^2}\sum_{j,k=1}^{n}exp\left(-\frac{\beta^2}{2}\left\|Y_j-Y_k\right\|^2\right) - 2(1+\beta^2)^{\frac{p}{2}}\frac{1}{n}\sum_{j=1}^{n}\exp\left(-\frac{\beta^2}{2(1+\beta^2)}\left\|Y_j\right\|^2\right) + (1+2\beta^2)^{\frac{p}{2}}\right]$$

ile elde edilir. Burada yer alan

$$\left\|Y_j-Y_k\right\|^2 = (X_j-X_k)'S_n^{-1}(X_j-X_k)$$
$$\left\|Y_j\right\|^2 = (X_j-\overline{X_n})'S_n^{-1}(X_j-\overline{X_n})$$

dir (17).

Doornik-Hansen

Doornik ve Hansen tarafından 1994 yılında geliştirilen bu test Bowman ve Shenton tarafından 1975 yılında önerilen çarpıklık ve basıklık katsayılarına bağlı geliştirilen tek değişkenli normallik testi test istatistiğinin çok değişkenli halidir. $X' = (X_1, X_2, \ldots, X_n)$ bir veri kümesi, p boyutlu bir vektörde n gözlemli pxn boyutlu bir matris olduğunda örneklem ortalaması $\bar{X} = n^{-1}(X_1 + X_2 + \cdots + X_n)$ ve kovaryans matrisi $S = n^{-1}\tilde{X}'\tilde{X}$ olmak üzere burada $\tilde{X}$;

$$\tilde{X} = (X_1 - \bar{X}, \ldots, X_n - \bar{X})$$

dir.

Standart sapmaların tersleriyle köşegen (diyagonal) matris;

$$V = diag(S_{11}^{-\frac{1}{2}}, \ldots, S_{pp}^{-\frac{1}{2}})$$

korelasyon matrisi C=VSV şeklinde oluşturulur. Verilerin dönüştürülmesi ile ilgili pxn boyutlu matris

$$R' = H\wedge^{-1/2}H'V\tilde{X}'$$

şeklinde tanımlanır.

Burada $\wedge = diag(\lambda_1, \ldots, \lambda_n)$, köşegen (diyagonal) elemanlarda C'nin özdeğerli matrisidir. H'nin sütünları özvektörlerdir. Öyle ki $H'H = I_p$ ve $\wedge = H'CH$ dir. C ve V için popülasyon değerlerini kullanarak çok değişkenli normal dağılım, bağımsız standart normal dağılımlara dönüştürülebilir. Dönüştürülen gözlem değerlerinden tek değişkenli eğrilik ve diklik katsayıları hesaplanabilir.

Doornik-Hansen testi için test istatistiği,

$$E_p = Z_1'Z_1 + Z_2'Z_2 \sim \chi^2_{(2p)}$$

Burada $Z_1' = (z_{11,\ldots,}z_{1p})$ ve $Z_2' = (z_{21,\ldots,}z_{2p})$ değerlerinin elde edilmesi ve eğrilik değeri $\sqrt{b_1}$ in z_1 şekline transformasyonu D'Agostino (1970) tarafından belirtildiği gibi aşağıdaki şekilde yapılır.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad m_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^i, \quad \sqrt{b_1} = \frac{m_3}{m_3^{3/2}},$$

$$b_2 = \frac{m_4}{m_2^2},$$

$$B = \frac{3(n^2 + 27n - 70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}$$

$$\omega^2 = -1 + \{(2(B-1)\}^{\frac{1}{2}}$$

$$\delta = \frac{1}{\{log(\sqrt{\omega^2})\}^{\frac{1}{2}}}$$

$$y = \sqrt{b_1}\left\{\frac{\omega^2 - 1(n+1)(n+3)}{2.6.(n-2)}\right\}^{1/2}$$

$$z_1 = \delta \log\{y + (y^2+1)^{1/2}\}$$

Diklik değeri b_2, bir gama dağılımından χ^2 dağılımına dönüştürülür ve sonra Wilson-Hilferty küp kök transformasyonu kullanarak standart normal z_2 değerine aşağıdaki şekilde dönüştürülür.

$$\delta = (n-3)(n+1)(n^2 + 15n - 4)$$

$$a = \frac{(n-2)(n+5)(n+7)(n^2+27n-70)}{6\delta}$$

$$c = \frac{(n-5)(n+5)(n+7)(n^2+2n-5)}{6\delta}$$

$$k = \frac{(n+5)(n+7)(n^3+37n^2+11n-313)}{12\delta}$$

$$\alpha = a + b_1 c$$

$$\chi = (b_2 - 1 - b_1)2k$$

$$z_2 = \left\{\left(\frac{\chi}{2\alpha}\right)^{1/3} - 1 + \frac{1}{9\alpha}\right\}(9\alpha)^{1/2}$$ (18, 19).

BULGULAR

Geliştirilen web tabanlı yazılım

Web tabanlı bu uygulamayı oluşturmak için, R programlama dili temelinde interaktif web tabanlı uygulamaların tasarlanmasına izin veren Shiny 1.0.5 sürümü kullanılmıştır. Geliştirilen yazılım ayrıca İngilizce dil seçeneğini de içermektedir. Yazılıma ait ana ve alt menüler aşağıda açıklanmıştır.

Dosya yükleme

Bu web tabanlı uygulamanın ilk aşamasında, veri kümesini içeren dosya yüklenir. Veri analizinde farklı uzantılara sahip en yaygın kullanılan dosya türlerinden olan MS Excel (.xls / .xlsx) ve SPSS (.sav) dosya türleri ile yükleme yapılır.

Ek olarak bu menü yüklenen dosyadaki değişkenlerin tipinin ve rolünün belirlenmesini sağlayacak olan 'Değişkenlerin tipini ve rolünü belirleyiniz' sekmesini içermektedir.

Normallik analizleri

Web tabanlı uygulamanın bu menüsünde 'Sayısal değişken(ler)' sekmesi ile normallik açısından analiz etmek istediğimiz değişkenleri seçebiliriz. Yapılacak analizlerin Tip-I hata düzeyi gerektiğini belirleyen bir 'anlamlılık düzeyi' sekmesi vardır. Bu sekme ile analizler 0,01-0,05-0,10 anlamlılık düzeylerinde yapılabilmektedir. Son olarak "Alt grup analizi" sekmesi işaretlenerek gruplar açısından da normallik analizlerini yapmak mümkün olacaktır. Şekil 1, "Normallik Analizleri" menüsünü göstermektedir.

Şekil 1. *Normallik analizleri menüsü.*

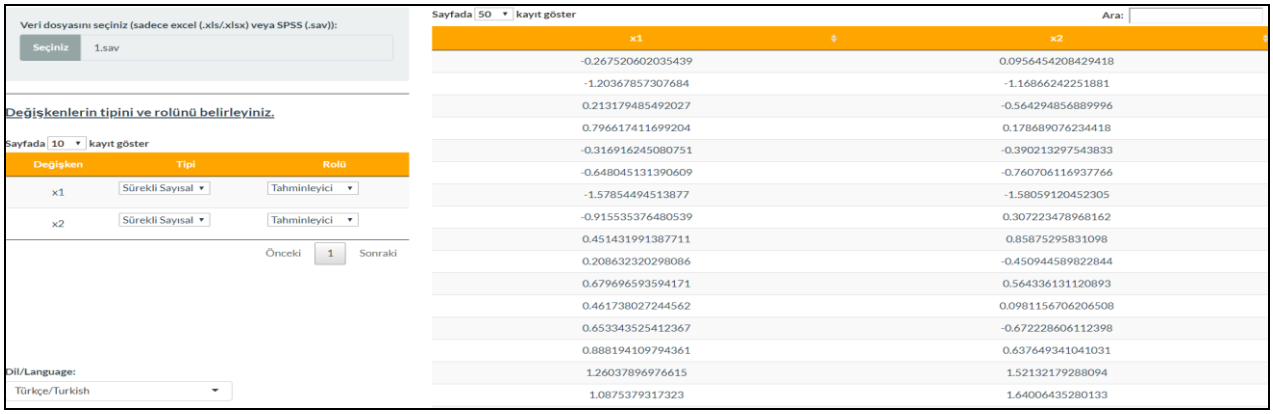
Bu menü ile tek değişkenli ve çok değişkenli normallik analizleri yapılabilmektedir. Tek değişkenli normal dağılıma uygunluk Shapiro-Wilk ve Anderson-Darling testleri ile test edilirken, çok değişkenli normal dağılıma uygunluk ise Mardia'nın çarpıklık-basıklık, Henze-Zirkler ve Doornik-Hansen testleri ile test edilmektedir. Ayrıca testlere ek olarak araştırmacıların kullanımı amacıyla grafiksel metotlardan olan Q-Q, çekirdek yoğunluk kestirim ve 3-boyutlu yüzey grafikleri de yazılımda sonuç çıktıları arasında yer almaktadır.

Geliştirilmiş interaktif web uygulamasının erişilebilirliği ve alıntılanması

Geliştirilen web tabanlı yazılıma **http://biostatapps.inonu.edu.tr/NDY/** adresinden ücretsiz olarak erişilebilir. Bilimsel çalışmalarda yazılımın kaynak gösterilmesi **"**Arslan, A. K. , Tunc, Z. & Colak, C. Normal Dağılım İnceleme Yazılımı [Web-tabanlı yazılım]" şeklinde yapılabilir. Bu yazılımın geliştirilmesinde shiny (20), shinyBS (21) ve shinythemes (22) paketleri kullanıldı.

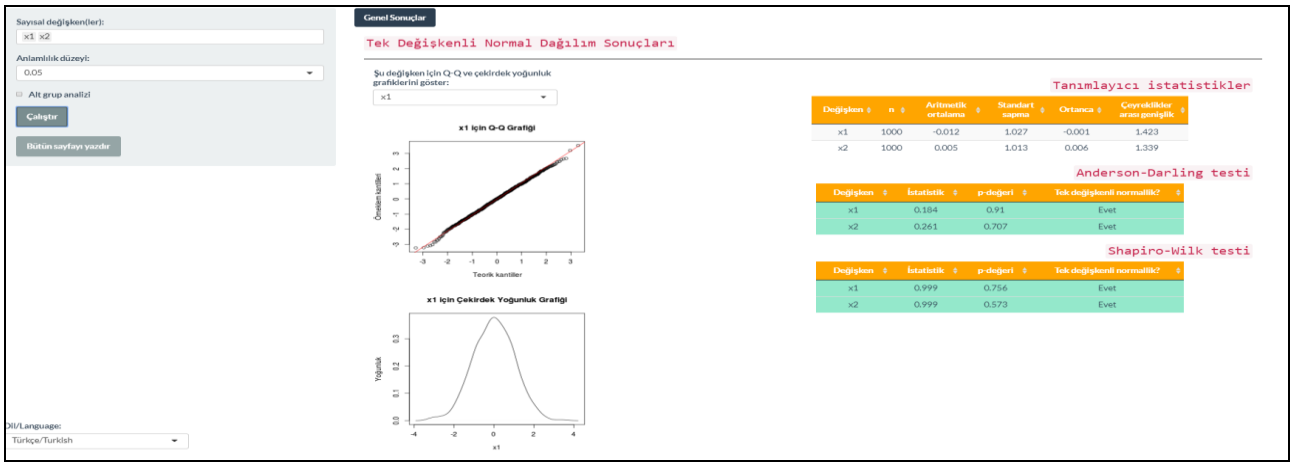
Deneysel bulgular

Yazılımın çalışma şeklini gösterebilmek ve çıktılarını inceleyebilmek için iki değişkenden oluşan her bir değişkenin standart normal dağılıma sahip olduğu ve değişkenlerin 1000 gözlem içerdiği veri seti üzerinde normal dağılıma uygunluk analizi gerçekleştirilmiştir. İlk olarak veri seti yazılıma yüklenmiştir. Şekil 2 yüklenen dosyaya ilişkin görüntüleri göstermektedir.



Şekil 2. *Yüklenen dosyaya ilişkin görüntüler.*

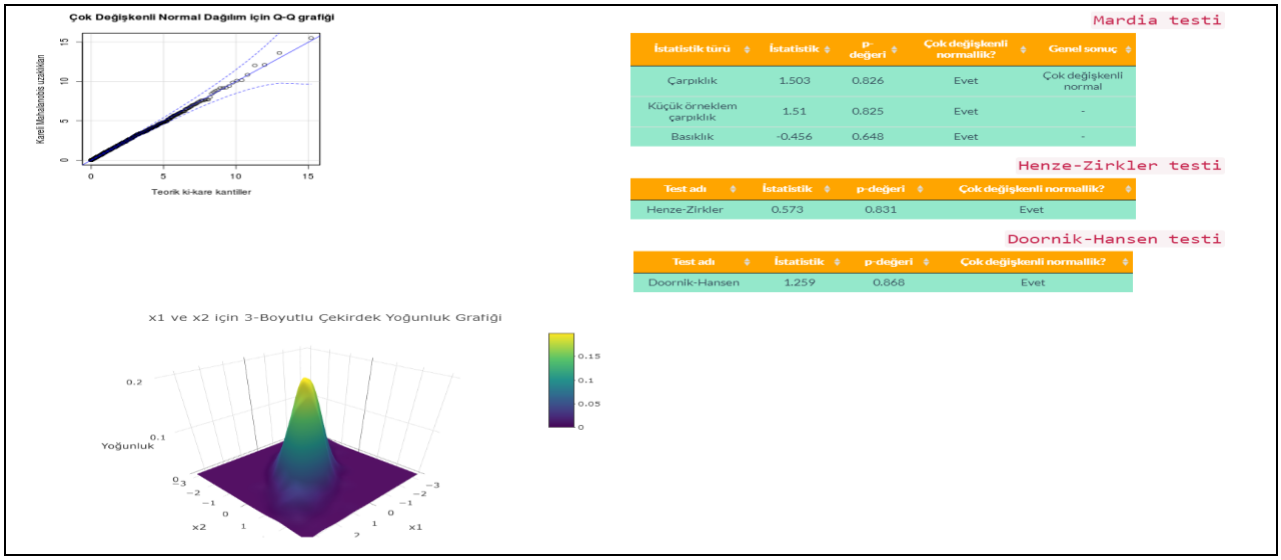
Sonra "Normallik Analizleri" sekmesi ile normalliği kontrol edilmek istenen değişkenler "sayısal değişkenler" sekmesi ile seçilir. Anlamlılık düzeyi belirlenerek normallik analizi yapılır. Şekil 3 Tek değişkenli normal dağılım sonuçlarına ait çıktıları göstermektedir.



Şekil 3. *Tek değişkenli normal dağılım sonuçlarına ait çıktılar.*

Bu sonuçlara göre iki değişkenden oluşan veri setindeki her bir değişken ayrı ayrı yazılımda normal dağılım göstermektedir.

Şekil 4 ise çok değişkenli normal dağılım' a ait sonuç çıktılarını göstermektedir.



Şekil 4. *Çok değişkenli normal dağılım' a ait sonuç çıktıları.*

Bu sonuçlara göre 2 değişkenden oluşan veri seti yazılımda yer alan üç çoklu normal dağılım yöntemine göre de normal dağılmaktadır.

TARTIŞMA

Tek değişkenli veri analizlerinde olduğu gibi çok değişkenli veri analizlerinde de çok değişkenli normallik varsayımı, yaygın olarak kullanılan çok değişkenli istatistiksel yöntemlerin temelini oluşturur. Çok değişkenli varyans analizi, diskriminant analizi, kanonik korelasyon analizi ve maksimum olasılık faktör analizi gibi çok değişkenli analizler, tek ve çok değişkenli normallik varsayımını gerektirir (23).

Tek değişkenli veriler için normallik varsayımını kontrol etme genel bir uygulama olarak düşünülürken, çok değişkenli veriler için çok sayıda test olduğu halde normallik varsayımını kontrol etme genelde uygulanmamaktadır (19). Fakat normallik koşulu sağlanmadan kullanılacak istatistiksel yöntemlerle elde edilen sonuçlar gerçek durumu yansıtmayabilir. Bu yüzden araştırıcılar verilerini analiz etmeden önce mutlaka normallik varsayımı sağlayıp sağlamadığını kontrol etmesi gerekir. Eğer veriler normal dağılıma uygunluk göstermiyorsa uygun dönüşümlerle normallik varsayımı sağlanmaya çalışılır veya parametrik olmayan yöntemler kullanılabilir (24).

Bu çalışmanın amacı araştırıcıların çok değişkenli normal dağılıma uygunluğu kolay bir şekilde kontrol ederek çalışmalarında daha doğru sonuçlar elde etmesini sağlayacak yeni kullanıcı dostu bir web tabanlı yazılım geliştirmektir. Yazılımda çok değişkenli normalliği test etmek için 3 ayrı testten yararlanılmıştır. Bu testler Mardia'nın çarpıklık-basıklık, Henze-Zirkler ve Doornik-Hansen testleridir. Normalliğe dair sonuç çıktıları araştırıcıların kolayca anlayabileceği şekilde grafiklerle desteklenerek verilmektedir.

İstatistiksel analizler için yaygınca kullanılan IBM SPSS Statistics (25), Minitab (26), MedCalc (27) gibi yazılımlarda çoklu normal dağılım yapılmasına imkan sağlayan bir modül bulunmamaktadır. Bu yazılımlara ek olarak Stata (28) programı çoklu normal dağılıma uygunluğu test edebilmektedir. Ancak açık kaynak web tabanlı bir yazılımda tek ve çok değişkenli normal dağılıma uygunluğu değerlendiren az sayıda yazılım vardır. Bunlardan biri Türkiye'de geliştirilen ve MVN: a web-tool for assessing multivariate normality (ver. 1.6) adı verilen yazılımdır (29).

Ayrıca bu çalışmada geliştirilen yazılımın çıktılarını değerlendirebilmek adına iki değişkenli her bir değişkenin standart normal dağılıma sahip olduğu ve 1000 gözlemli veri seti, IBM SPSS Statistics sürüm 25.0 ile türetilmiştir (25). Analiz sonuçlarına göre 2 değişkenin her biri tek değişkenli normal dağılıma uygunluk gösterdi. Ayrıca veri seti çok değişkenli normal dağılımı belirlemek için yazılımda yer verilen üç teste göre de çoklu normal dağılıma uygun olduğu gösterilmiştir.

Sonuç olarak, geliştirilen yazılım tek değişkenli ve çok değişkenli normal dağılıma uygunluk analizlerini kolayca yapabilen ve kullanıcıların yapacakları çalışmalarında daha doğru sonuçlar elde etmesini sağlayan yeni kullanıcı dostu bir web tabanlı yazılımdır. İlerleyen çalışmalarda, en iyi yönteme karar vermede kullanılan ölçütlerden Tip I ve Tip II hata türlerinin yazılıma eklenmesi planlanmaktadır.

KAYNAKLAR

1. Peng G. Testing normality of data using SAS. Indianapolis: Lilly Corporate Center, 2004.
2. Pituch KA, Stevens JP. Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS: Routledge. 2015.
3. Sharma S, Sharma S. Applied multivariate techniques. 1996.
4. Thode HC. Testing for normality. CRC press, 2002.
5. Mertler CA, Reinhart RV. Advanced and multivariate statistical methods: Practical application and interpretation: Routledge. 2016.
6. Mardia KV. Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Sankhyā: The Indian J Statist, Series B 1974; 36: 115-28.
7. Gnanadesikan R. Methods for statistical data analysis of multivariate observations. John Wiley & Sons 2011.
8. Darling DA. The kolmogorov-smirnov, cramer-von mises tests. The Ann Math Statist 1957; 28: 823-38.

9. Koziol JA. Assessing multivariate normality: a compendium. Commun Stat-Theor M 1986; 15: 2763-83.
10. Mudholkar GS, Srivastava DK, Thomas Lin C. Some p-variate adaptations of the Shapiro-Wilk test of normality. Commun Stat-Theor M 1995; 24: 953-85.
11. Paulson A, Roohan P, Sullo P. Some empirical distribution function tests for multivariate normality. J Stat Comput Simul 1987; 28: 15-30.
12. Ergun M, Çilingir F. İlköğretim bölümünde yapılan lisansüstü tezlerin incelenmesi: Ondokuz Mayıs Üniversitesi örneği. VI Ulusal Lisansüstü Eğitim Sempozyumu 2013; 85.
13. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965; 52: 591-611.
14. Özdamar K. Paket Programlar ile istatistiksel veri analizi. Eskişehir: Kaan Kitabevi, 2004.
15. Mardia K. Assessment of multinormality and the robustness of Hotelling's T 2 test. Applied Statistics 1975: 163-71.
16. Mardia KV. Measures of multivariate skewness and kurtosis with applications. Biometrika 1970; 57: 519-30.
17. Henze N, Zirkler B. A class of invariant consistent tests for multivariate normality. Commun Stat-Theor M 1990: 3595-617.
18. Doornik J, Hansen H. Am Omnibus test of univariate and multivariate normality. Oxford: Unpublished paper Nuffield College 1994.
19. Naczk K. Assessing tests for multivariate normality. Carleton University 2005.
20. Cheng J, Xie Y, McPherson J. shiny: web application framework for R. R package version 0.13. 2. 2016.
21. Bailey E. shinyBS: Twitter bootstrap components for shiny. R package version 061 URL https://CRAN R-project org/package= shinyBS 2015.
22. Chang W. shinythemes: Themes for Shiny. R package version 2015; 1:144.
23. Looney SW. How to use tests for univariate normality to assess multivariate normality. ASA 1995; 49: 64-70.
24. Akçadağ Hİ. Tek değişkenli ve çok değişkenli bazı normallik testlerinin karşılaştırılması [Doktora Tezi]. Selçuk Üniversitesi Fen Bilimleri Enstitüsü 2013.
25. Corp I. IBM SPSS statistics for windows, version 22.0. Armonk, NY: IBM Corp 2013.
26. Minitab I. MINITAB statistical software. Minitab Release 2000; 13.
27. Schoonjans F, Zalata A, Depuydt C, Comhaire F. MedCalc: a new computer program for medical statistics. Comput Methods Programs Biomed 1995; 48: 257-62.
28. StataCorp L. Stata data analysis and statistical Software. Special Edition Release 2007; 10: 733.
29. Korkmaz S, Goksuluk D, Zararsiz G. MVN: An R package for assessing multivariate normality. RJ 2014; 6: 151-62.

Ahmet Kadir ARSLAN 0000-0001-8626-9542
Zeynep TUNÇ 0000-0001-7956-9772
Cemil ÇOLAK 0000-0001-5406-098X