Research, Society and Development, v. 10, n. 11, e278101119317, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i11.19317

Techniques of quality of adjustment of statistical models with evaluation of probability distributions using production data of laying quails

Técnicas de qualidade de ajuste de modelos estatísticos com avaliação de distribuições de probabilidade utilizando dados de produção de codornas poedeiras

Técnicas de calidad de ajuste de modelos estadísticos con evaluación de distribuciones de probabilidad utilizando datos de producción de codornices ponedoras

Received: 08/11/2021 | Reviewed: 08/26/2021 | Accept: 08/29/2021 | Published: 08/31/2021

Antonio Augusto Carvas Sant' Anna
ORCID: https://orcid.org/0000-0002-3415-840X
Universidade Estadual do Norte Fluminense, Brazil
E-mail: augustosantanna@pq.uenf.br
Jacyara Lopes Pereira
ORCID: https://orcid.org/0000-0001-5411-4888
Universidade Estadual do Norte Fluminense, Brazil
E-mail: jacyara@pq.uenf.br
Matheus Lima Corrêa Abreu
ORCID: https://orcid.org/0000-0002-3533-7338
Universidade Federal de Mato Grosso, Brazil
E-mail: matheus.zoot@yahoo.com.br
Adolpho Marlon Antoniol de Moura
ORCID: https://orcid.org/0000-0002-4040-9041
Fundação Oswaldo Cruz, Brazil
E-mail: dofo.antoniol78@gmail.com
Elon Souza Aniceto
ORCID: https://orcid.org/0000-0002-6967-2361
Universidade Estadual do Norte Fluminense, Brazil
E-mail: elon1995@hotmail.com
Jonas Henrique Motta
ORCID: https://orcid.org/0000-0001-7124-9119
Universidade Estácio de Sá, Brazil
E-mail: motta.henri@gmail.com
Leonardo Siqueira Glória
ORCID: https://orcid.org/0000-0002-2756-5939
Universidade Estadual do Norte Fluminense, Brazil
E-mail: leonardogloria@uenf.br

Abstract
The goal of our study was to evaluate the quality of fit from different types of probability distributions for continuous data. For this, performance traits and quality of quail egg in the production of nutraceutical eggs were used as a continuous data source. The data were collected over 42 days, the experimental design was completely randomized with 7 treatments, 6 repetitions, with 252 animals allocated in 36 cages. The distributions for continuous data used were the exponential, gamma, gaussian, and lognormal. The R Open Source and SAS® University Edition software was used to perform the analysis. The graphical analysis of the traits was performed from the predicted versus observed values, Cumulative Distribution Function (CDF), and skewness-kurtosis. The fits were also evaluated by the Akaike information criterion (AIC), Bayesian information criterion (BIC), Conditional model of adjusted R-Square (R^2_{ac}), Conditional model of adjusted concordance correlation (r_{ac}), Kolmogorov-Smirnov test (KS), Cramer-von Mises test (CvM), Anderson-Darling test (AD), Watanabe-Akaike Information Criterion (WAIC) and Leave-one-out cross-validation (LOO). All the tests indicated the Gaussian distribution as the most suitable and they excluded the exponential distribution for all the evaluated characteristics.
Keywords: Animal production; Distribution continuous data; Statistical analysis; Generalized linear mixed model; Test of fit; Statistical software.

Resumo
O objetivo do nosso estudo foi avaliar a qualidade do ajuste de diferentes tipos de distribuições de probabilidade para dados contínuos. Para tanto, as variáveis desempenho e qualidade do ovo de codorna na produção de ovos nutracêuticos foram utilizadas como fonte contínua de dados. Os dados foram coletados ao longo de 42 dias, o delineamento

experimental foi inteiramente casualizado com 7 tratamentos, 6 repetições, com 252 animais alocados em 36 gaiolas. As distribuições de dados contínuos usadas foram a exponencial, gama, gaussiana e lognormal. Os softwares R Open Source e SAS® University Edition foram usados para realizar a análise. A análise gráfica das variáveis foi realizada a partir dos valores previstos versus observados, Função de Distribuição Cumulativa (CDF) e assimetria-curtose. Os ajustes também foram avaliados pelo critério de informação de Akaike (AIC), critério de informação Bayesiano (BIC), modelo condicional de R-quadrado ajustado (R^2_{ac}), modelo condicional de correlação de concordância ajustada (r_{ac}), teste de Kolmogorov-Smirnov (KS), Teste de Cramer-von Mises (CvM), teste de Anderson-Darling (AD), Critério de informação Watanabe-Akaike (WAIC) e validação cruzada de deixar um de fora (LOO). Todos os testes indicaram a distribuição gaussiana como a mais adequada e excluíram a distribuição exponencial para todas as características avaliadas.
Palavras-chave: Produção animal; Distribuição contínua de dados; Análise estatística; Modelo linear generalizado misto; Teste de ajuste; Software estatístico.

Resumen
El objetivo de nuestro estudio fue evaluar la calidad del ajuste de diferentes tipos de distribuciones de probabilidad para datos continuos. Para ello, se utilizaron como fuente continua de datos las variables rendimiento y calidad del huevo de codorniz en la producción de huevos nutracéuticos. Los datos fueron recolectados durante 42 días, el diseño experimental fue completamente al azar con 7 tratamientos, 6 repeticiones, con 252 animales distribuidos en 36 jaulas. Las distribuciones para los datos continuos utilizadas fueron exponencial, gamma, gaussiana y lognormal. Se utilizó el software R Open Source y SAS® University Edition para realizar el análisis. El análisis gráfico de las variables se realizó a partir de los valores predichos versus los observados, la Función de Distribución Acumulativa (CDF) y la asimetría-curtosis. Los ajustes también se evaluaron mediante el criterio de información de Akaike (AIC), el criterio de información bayesiano (BIC), el modelo condicional de R-Cuadrado ajustado (R^2_{ac}), el modelo condicional de correlación de concordancia ajustada (r_{ac}), la prueba de Kolmogorov-Smirnov (KS), Prueba de Cramer-von Mises (CvM), prueba de Anderson-Darling (AD), criterio de información de Watanabe-Akaike (WAIC) y validación cruzada de dejar uno fuera (LOO). Todas las pruebas indicaron la distribución gaussiana como la más adecuada y excluyeron la distribución exponencial para todas las características evaluadas.
Palabras clave: Producción animal; Distribución de datos continuos; Análisis estadístico; Modelo lineal mixto generalizado; Prueba de ajuste; Software estadístico.

1. Introduction

Instead of adapting our data to classical statistics (Analysis of Variance - ANOVA), we should use approaches according to the characteristics of the observations, such as mixed models for situations involving fixed and random effects or generalized models when we have non-normal data (Bolker et al., 2009).

From the determination of the data type (discrete, continuous, sample space, etc.), the visual assessment can be used to verify the model and the most appropriate distribution to the data seeking greater verisimilitude. Scatterplots, for example, provide some information about the presence of outliers, the relationship between variables (correlations), and the behavior of the data over time (Vonesh, 2014; Sher et al, 2017).

Silva et al. (2020) demonstrated the importance and ease of use of computational tools to analyze the data, adjust the distributions and verify the model. Making it a useful tool in reducing errors, costs, and processing time. Like other authors, they used statistical modeling techniques to forecast corn crops in the state of Mato Grosso, as they realized the need to predict events such as rain or drought, based on past situations (Silva et al., 2019).

Therefore, tools such as SAS® and R Open Source are used to identify the models that best describe reality. The GLIMMIX GOF macro of the SAS® software provides the predicted and observed values (Vonesh & Chinchilli, 1996). The fitdistrplus package of the R software produces graphs of skewness-kurtosis and Cumulative Distribution Function (CDF) (Muller et al., 2015).

The choice of Akaike (AIC) (Akaike, 1974) or Bayes (BIC) (Schwarz, 1978) information criteria for model selection must be based on their principles as statistical inference and nature of data (Anderson & Burnham, 2004). The information generated by the AIC is useful when there are not too many samples or when there is need to identify the number of model parameters, thereby penalizing the most complex models regarding the numbers of parameters of a model. The values of AIC

and BIC are similar, since both require the estimation of parameters, however, the second considers Bayesian and deviation information (Deviance Information Criterion - DIC) (Bolker et al., 2009). According to Brito et al. (2020), the importance of using statistical models was verified, to identify those that most represent and help in the prediction of human diseases (Diabetic Retinopathy) and used parameters such as AIC and BIC to evaluate the most diverse distributions studied.

The Kolmogorov-Smirnov (KS) (Massey, 1951), Cramer-von Mises (CvM) (Darling, 1957) and Anderson-Darling (AD) tests do not consider the complexity of the models, which leads researchers to always use them when the parameters is known and constant, to avoid favoring more complex models, respecting the principle of parsimony (Muller & Dutang, 2015).

The brms package of the R software provides the criteria Watanabe-Akaike Information Criterion (WAIC) (Gelman, 2014) and Leave-one-out cross-validation (LOO). With a flexible structure, the package has several distributions available, allowing the alteration of the connection functions and the use of some information we already have from the data. WAIC can be an improvement in DIC (Bürkner, 2017).

In this study, our goal was to elucidate the available techniques for assessing the fit quality of statistical models, identifying the most appropriate distributions for the analyzed traits.

2. Methodology

The choice of which methodology to use in research serves to describe your hypothesis and confront it, to obtain information that contributes to learning (Pereira et al., 2018). To obtain quality and reliable results, Lüdke & André (1986) described in the form of a guide, all the planning to be carried out and developed in research, from data collection, analysis, and interpretation of results. Based on previous studies in animal science, we developed our research to maintain the reliability of our results (Pan W., 2001; Bürkner, 2017; Muller & Dutang, 2015; Vonesh et al., 1996).

2.1 Data sampling

The data were obtained from an experiment carried out in accordance with the institutional committee for the use of animals (protocol 0059/2013), at the Experimental Poultry Farm, in the county of Itaguaí, Rio de Janeiro State, Brazil. We used 252 female Japanese quails (*Coturnix japonica*) of the Fujikura lineage. The animals were 90 days old, the average weight of 188.3 ± 4.0 g, and average laying rate of 90%. The quails were allocated in 36 laying cages with dimensions of 33 × 25 × 20 cm. The experimental design was completely randomized with 7 treatments, 6 repetitions with 6 animals per cage (experimental unit). The experimental diets were obtained from a control diet with increasing levels of organic selenium (0.10; 0.20; 0.30 and 0.40 mg) and 200 mg of DL-a-tocopheryl acetate were added per kilogram of feed as a source of vitamin E. Thus, the diets were: (1) Control diet (CD); (2) CD + 200 mg of vitamin E (VE); (3) CD + 0.20 ppm of organic selenium (SE); (4) CD + 0.10 ppm of organic selenium + 200 mg of vitamin E (SVE1); (5) CD + 0.20 ppm of organic selenium + 200 mg of vitamin E (SVE2); (6) CD + 0.30 ppm of organic selenium + 200 mg of vitamin E (SVE3); e (7) DC + 0.40 ppm of organic selenium + 200 mg of vitamin E (SVE4).

We used the nutritional requirements of Japanese quails described by NRC (1994) for formulating the diet. The exceptions were protein and calcium requirements that were based on the recommendations of Oliveira et al. (1999) and Barreto et al. (2007), respectively. The supplementation of vitamins and minerals was produced without selenium from tocopherols, this way we did not overestimate the concentrations.

The performance traits related to egg production were feed intake (Fi; g/(quail×day)), egg mass (EM; g/(quail×day)), daily egg yield per quail (DEy; eggs/(quail×day)), and feed conversion (FC; dmls, ie, intake mass/egg mass). We also evaluated egg quality using yolk mass (YM; g/egg); albumen mass (AM; g/egg); eggshell mass (ESM; g); yolk ratio (Yr; %); eggshell ratio (ESr; %); and albumen ratio (Ar; %).

At the beginning of the experiment, an adaptation phase was carried out for seven days, with the offering of the control diet to all animals. Subsequently, the experimental diets were offered in the ad libitum feeding system. The animals were submitted to a photoperiod of 17 hours, with the light controlled by a timer, and the temperature and relative humidity were recorded inside the house. The egg production trait and Fi were measured weekly. The traits YM, Am, and ESM were determined by collecting 3 eggs from each repetition (daily). As a reference, on day "zero", 80 eggs were collected randomly and followed the same protocol.

After 42 days of supplementation, the nutraceutical effects of tocopherol and selenium were assessed by analytically determining the concentration of their metabolic indicator, malondialdehyde, in the yolk (MDAY (mmol/g) of quails eggs, according to the methodologies described by Shahryar et al. (2010) and Enkvetchakul et al. (1995).

2.2 Data analysis

We used the software SAS® University Edition, and R Open-source version 3.6.3 to perform the analyses. The machine had a Linux Elementary OS operating system, with 6 GB RAM, 500 GB HD, and Intel® Core™ i7 processor. For all the studied traits, we evaluated the fit of the Gaussian, lognormal, gamma, and exponential distributions, that are suitable for continuous data.

2.3 Good of fit analysis

The SAS® %GLIMMIX_GOF macro was used to generate the R-Square Type Goodness-of-Fit Information and Model Fitting Information tables for each distribution and variable.

From the GLIMMIX procedure, we evaluated for each distribution the relationship on the Cartesian plane between observed values (y-axis) and predicted values (Pred, x-axis). From %GLIMMIX_GOF macro, the Conditional Model Adjusted R-Square (R^2_{ac}) and Conditional Model Adjusted Concordance Correlation (r_{ac}) were obtained, and corrected by fixed and random factors, as well as fitted to parameter numbers. The use this same macro followed the methodology described by Vonesh et al. (1996).

2.4 Fit analysis through tests

The R open-source software, with the *gofstat* function of the fitdistrplus package, was used to evaluate candidate distributions, these being Gaussian, lognormal, gamma and exponential by means of the tests of Kolmogorov-Smirnov (KS), Cramer-von Mises (CvM) and Anderson-Darling (AD) tests, according to the methodology described by Muller & Dutang (2015). Other outputs of the *gofstat* function such as the Akaike (AIC) and Bayesian (BIC) information criteria, were also recorded.

2.5 Bayesian method for information criteria

The function *brm* and package *brms*, that uses Stan language (Stan Development Team, 2017; Carpenter, 2017), was used to describe the model and the distribution, and from the *waic* and *loo* functions the values of the information criteria of Watanabe-Akaike Information Criterion (WAIC) and Leave-one-out cross-validation (LOO) were obtained (Bürkner, 2017).

3. Results

3.1 Coefficients (R^2_{ac} and r_{ac})

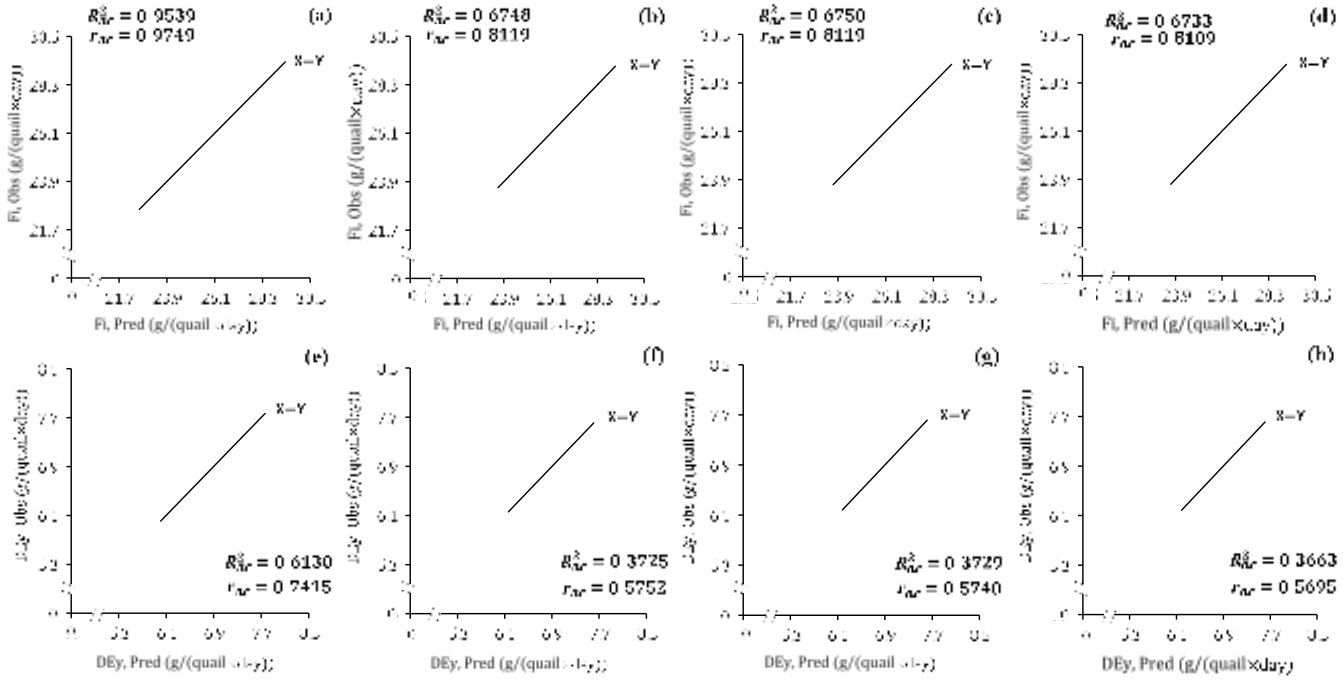
According to the results of the SAS® GLIMMIX_GOF macro, all traits had the highest R^2_{ac} and $r_{ac}$with the Gaussian

distribution (panels (a) Figures 1 to 6). The variable MDAY was the only one, among the 11 traits evaluated, that presented values of R^2_{ac} and r_{ac} with Gamma (c) and exponential (d) distributions equal to the Gaussian one.

Only the traits Fi (Figure 1; panel (a)), Yr (Figure 4; panel (a)) and Ar (Figure 5; panel (a)) presented values of R^2_{ac} and r_{ac} and rac greater than 0.9 and only with the Gaussian distribution.

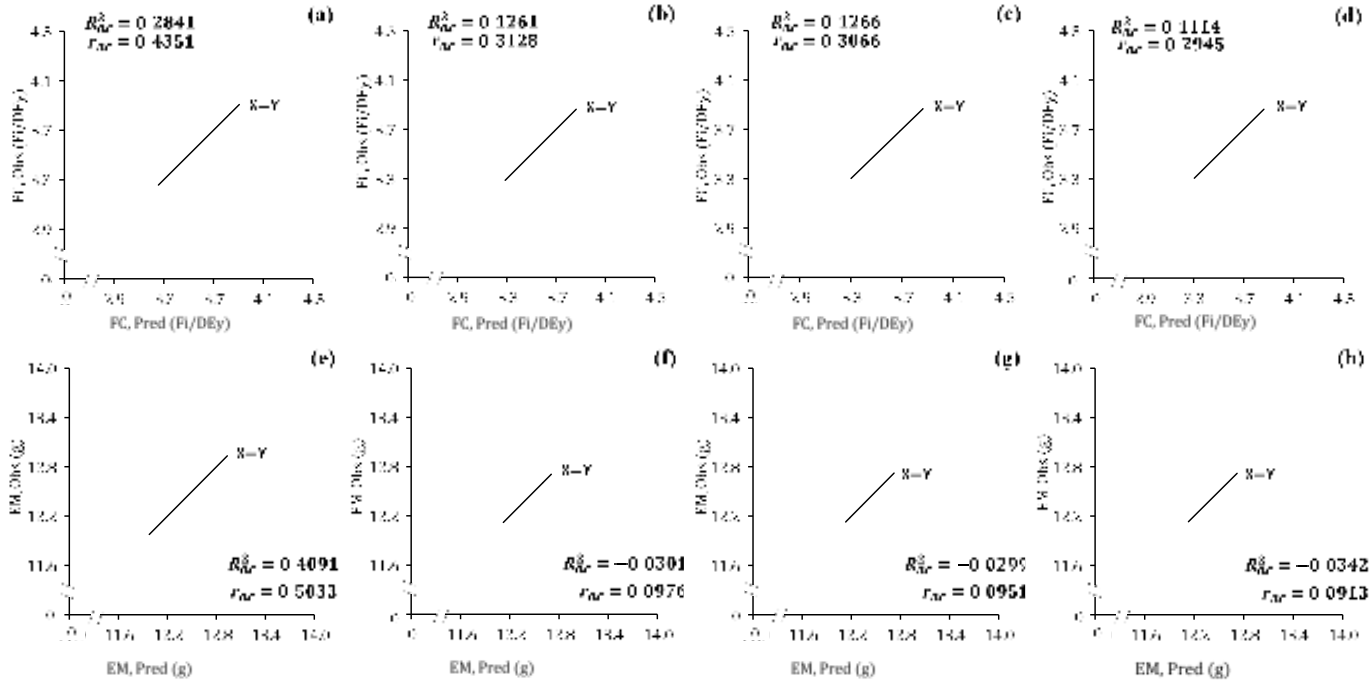
For the traits MDAY, DEy, Fi, Em, FC, YM, ESM, Yr, ESr, and Ar, the r_ac values (Figures 1 to 6) showed a positive correlation in all distributions. Except for the Am, which showed a positive correlation only for the Gaussian distribution, and negative for other distributions.

Figure 1. Observed (Obs) and predicted (Pred) of the traits feed intake (Fi) and daily eggs yield (DEy) of laying quails fitted by GLIMMIX procedure with the quantitative distributions: Gaussian (panels a and e), lognormal (panels b and f), gamma (panels c and g) and exponential (panels d and h). The observed data are represented in the graphs by the marker "○", and the solid lines correspond to the unit lines (Observed = Predicted; e.i., X=Y). For each fit, the Conditional Model Adjusted R-Square (R^2_{ac}) and Conditional Model Adjusted Concordance Correlation (r_{ac}) were generated through the Goodness-of-Fit analysis (GOF) of % GLIMMIX_GOF.



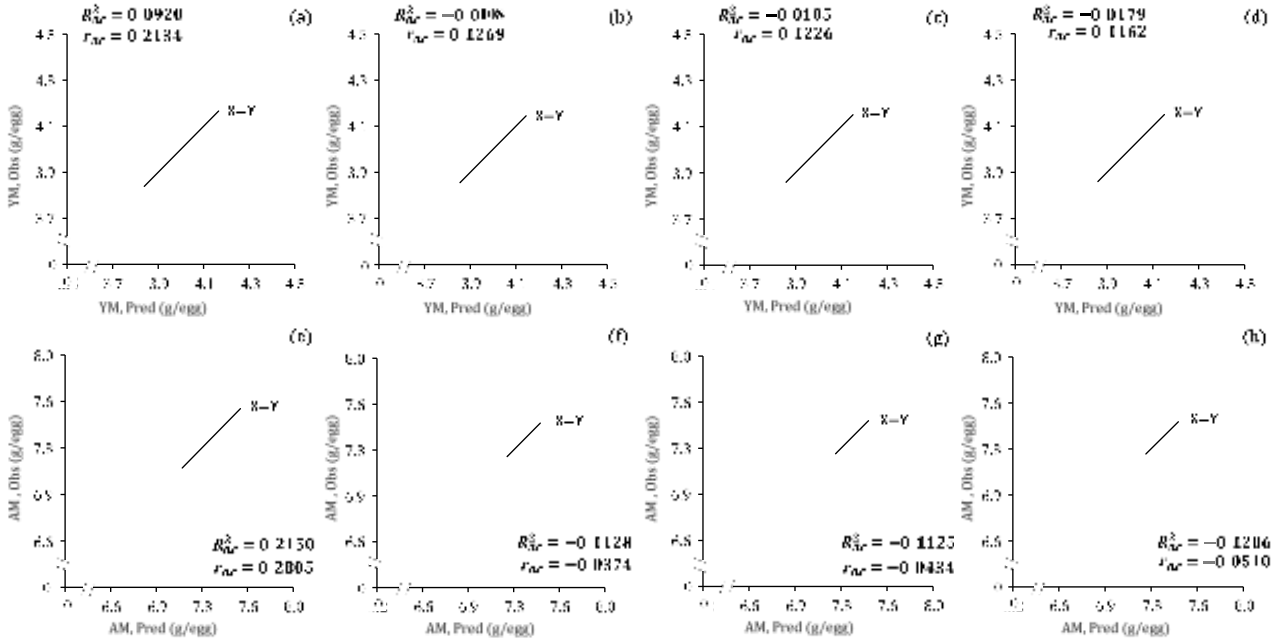
Source: Authors.

Figure 2. Observed (Obs) and predicted (Pred) of the traits feed conversion (FC) and eggs mass (EM) of laying quails fitted by GLIMMIX procedure with the quantitative distributions: Gaussian (panels a and e), lognormal (panels b and f), gamma (panels c and g) and exponential (panels d and h). The observed data are represented in the graphs by the marker "○", and the solid lines correspond to the unit lines (Observed = Predicted; e.i., X=Y). For each fit, the Conditional Model Adjusted R-Square (R^2_{ac}) and Conditional Model Adjusted Concordance Correlation (r_{ac}) were generated through the Goodness-of-Fit analysis (GOF) of % GLIMMIX_GOF.



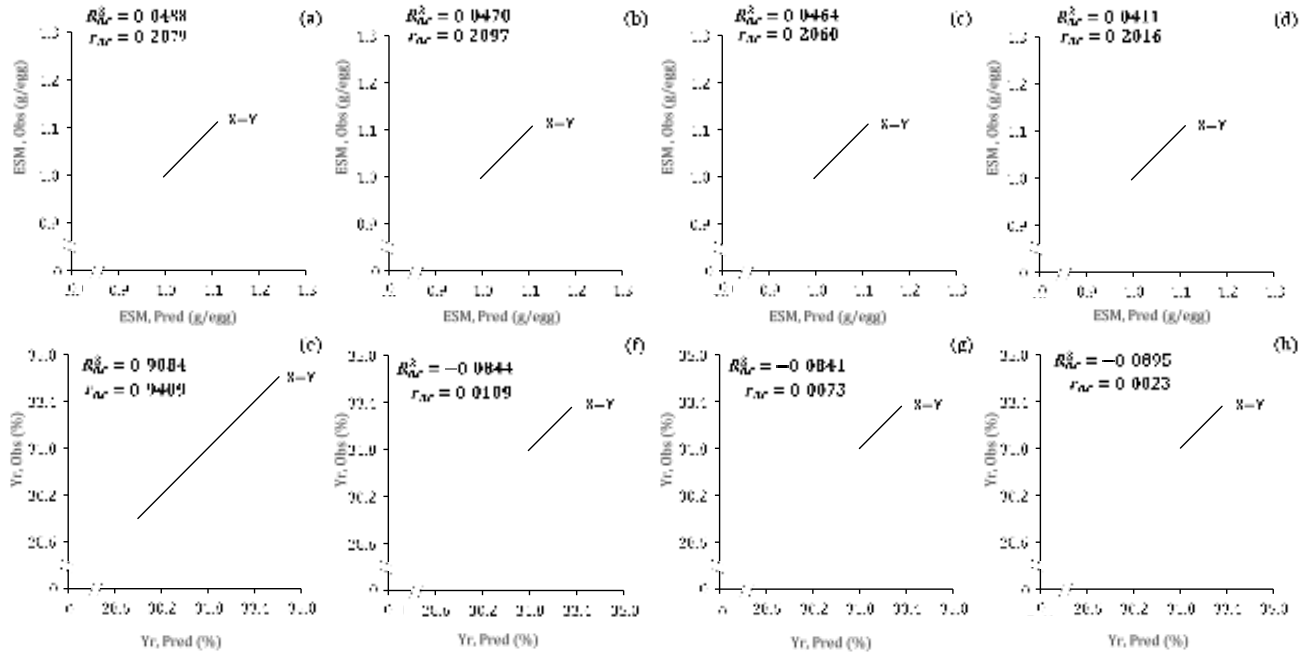
Source: Authors.

Figure 3. Observed (Obs) and predicted (Pred) of the traits yolk mass (YM) and albumen mass (AM) of laying quails fitted by GLIMMIX procedure with the quantitative distributions: Gaussian (panels a and e), lognormal (panels b and f), gamma (panels c and g) and exponential (panels d and h). The observed data are represented in the graphs by the marker "○", and the solid lines correspond to the unit lines (Observed = Predicted; e.i., X=Y). For each fit, the Conditional Model Adjusted R-Square (R^2_{ac}) and Conditional Model Adjusted Concordance Correlation (r_{ac}) were generated through the Goodness-of-Fit analysis (GOF) of % GLIMMIX_GOF.



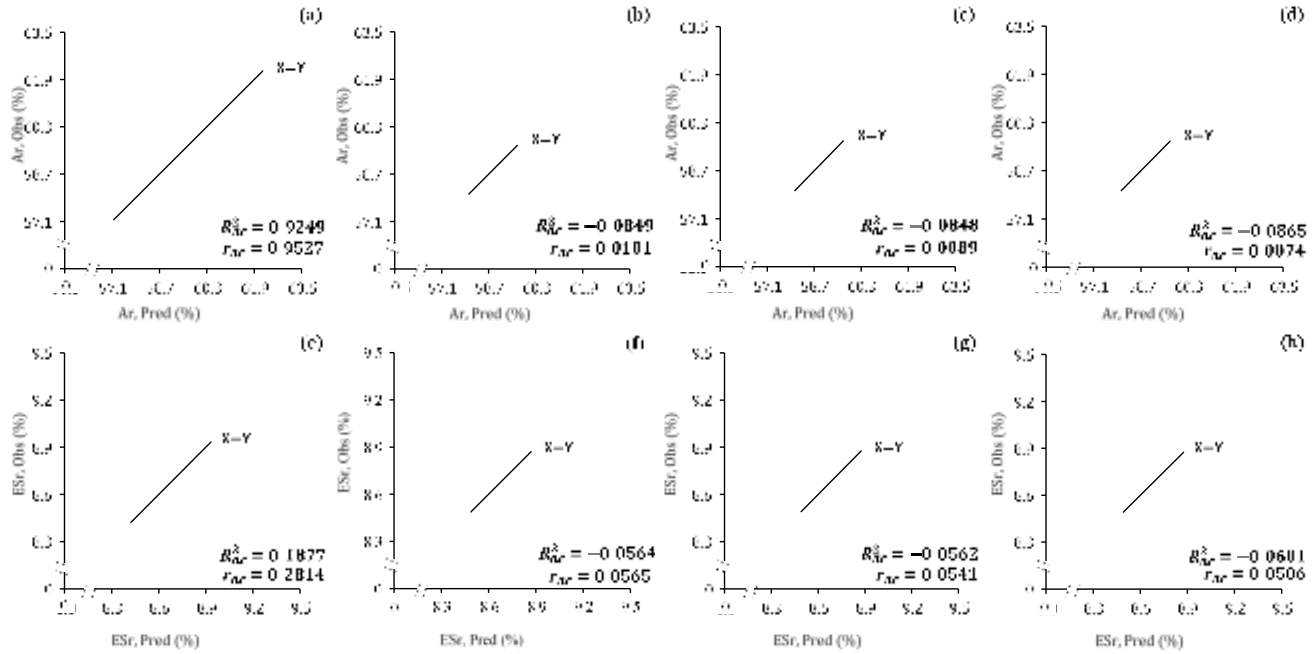
Source: Authors.

Figure 4. Observed (Obs) and predicted (Pred) of the traits eggshell mass (ESM) and yolk ratio (Yr) of laying quails fitted by GLIMMIX procedure with the quantitative distributions: Gaussian (panels a and e), lognormal (panels b and f), gamma (panels c and g) and exponential (panels d and h). The observed data are represented in the graphs by the marker "○", and the solid lines correspond to the unit lines (Observed = Predicted; e.i., X=Y). For each fit, the Conditional Model Adjusted R-Square (R^2_{ac}) and Conditional Model Adjusted Concordance Correlation (r_{ac}) were generated through the Goodness-of-Fit analysis (GOF) of % GLIMMIX_GOF.



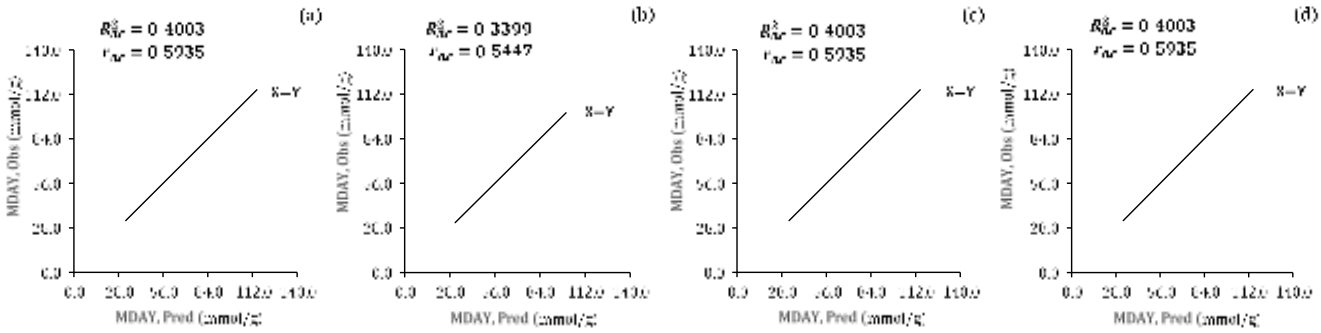
Source: Authors.

Figure 5. Observed (Obs) and predicted (Pred) of the traits albumen ratio (Ar) and eggshell ratio (ESr) of laying quails fitted by GLIMMIX procedure with the quantitative distributions: Gaussian (panels a and e), lognormal (panels b and f), gamma (panels c and g) and exponential (panels d and h). The observed data are represented in the graphs by the marker "○", and the solid lines correspond to the unit lines (Observed = Predicted; e.i., X=Y). For each fit, the Conditional Model Adjusted R-Square (R^2_{ac}) and Conditional Model Adjusted Concordance Correlation (r_{ac}) were generated through the Goodness-of-Fit analysis (GOF) of % GLIMMIX_GOF.



Source: Authors.

Figure 6. Observed (Obs) and predicted (Pred) of the variable malondialdehyde in yolk (MDAY) of laying quails fitted by GLIMMIX procedure with the quantitative distributions: Gaussian (panel a), lognormal (panel b), gamma (panel c) and exponential (panel d). The observed data are represented in the graphs by the marker "○", and the solid lines correspond to the unit lines (Observed = Predicted; e.i., X=Y). For each fit, the Conditional Model Adjusted R-Square (R^2_{ac}) and Conditional Model Adjusted Concordance Correlation (r_{ac}) were generated through the Goodness-of-Fit analysis (GOF) of % GLIMMIX_GOF.



Source: Authors.

3.2 Tests

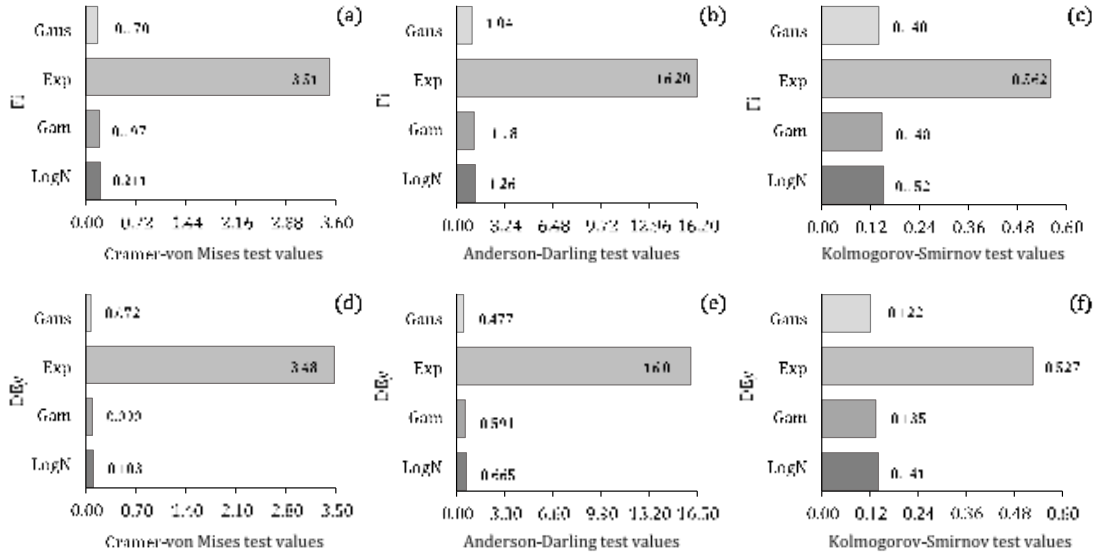
Regardless of the three tests used, ten among eleven traits analyzed with the exponential distribution, the lack of adjustment was evident, due to the much higher values in relation to the other analyzed distributions (Figures 7 to 12).

The test of Kolmogorov-Smirnov, for the traits, EM, YM, ESM, Ar, ESr, and MDAY the smaller values were obtained by lognormal distribution, while for DEy, Fi, FC, AM, and Yr the Gaussian distribution was the least biased.

In the Cramer-von Mises test, the traits MDAY, YM, ESM, and ESr, and Lognormal Ar showed lower values. For DEy, Fi, FC, Am, and Yr the Gaussian distribution was the best. In this same test, the EM traits showed the same value for the Gamma and Gaussian distributions.

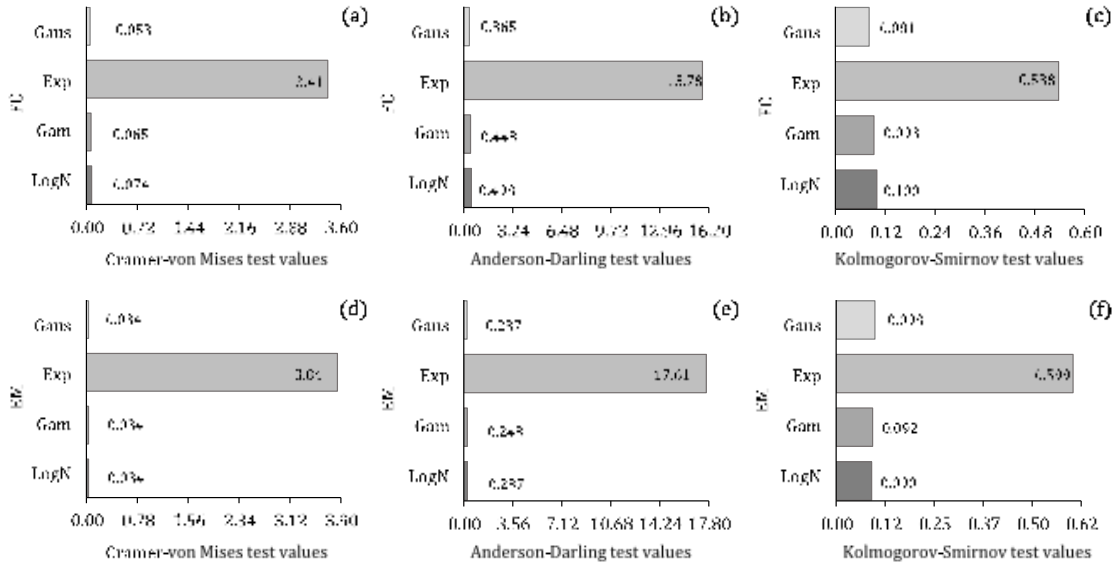
In the Anderson-Darling test, the traits MDAY, ESM, ESr, and Ar showed smaller values with the lognormal distribution. The DEy, Fi, EM, FC, AM, and Yr showed smaller values in the Gaussian distribution. The YM had the smaller value in the Gamma distribution, however the values for lognormal and Gaussian distributions were close.

Figure 7. Cramer-von Mises (panels a and d), Anderson-Darling (panels b and e), Kolmogorov-Smirnov (panels c and f) test values for the traits feed intake (Fi; panels a to c) and daily eggs yield (Dey; panels d to f) of laying quails fitted by *Fitdist* procedure with the quantitative distributions: Gaussian, Exponential, Gamma and LogNomal.



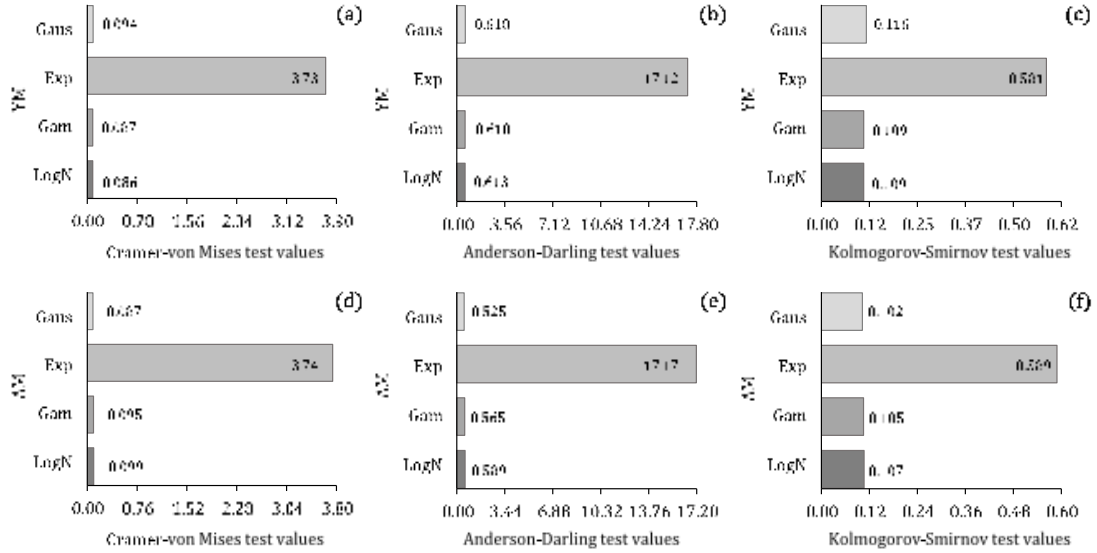
Source: Authors.

Figure 8. Cramer-von Mises (panels a and d), Anderson-Darling (panels b and e), Kolmogorov-Smirnov (panels c and f) test values for the traits feed conversion (FC; panels a to c) and eggs mass (EM; panels d to f) of laying quails fitted by Fitdist procedure with the quantitative distributions: Gaussian, Exponential, Gamma and LogNomal.



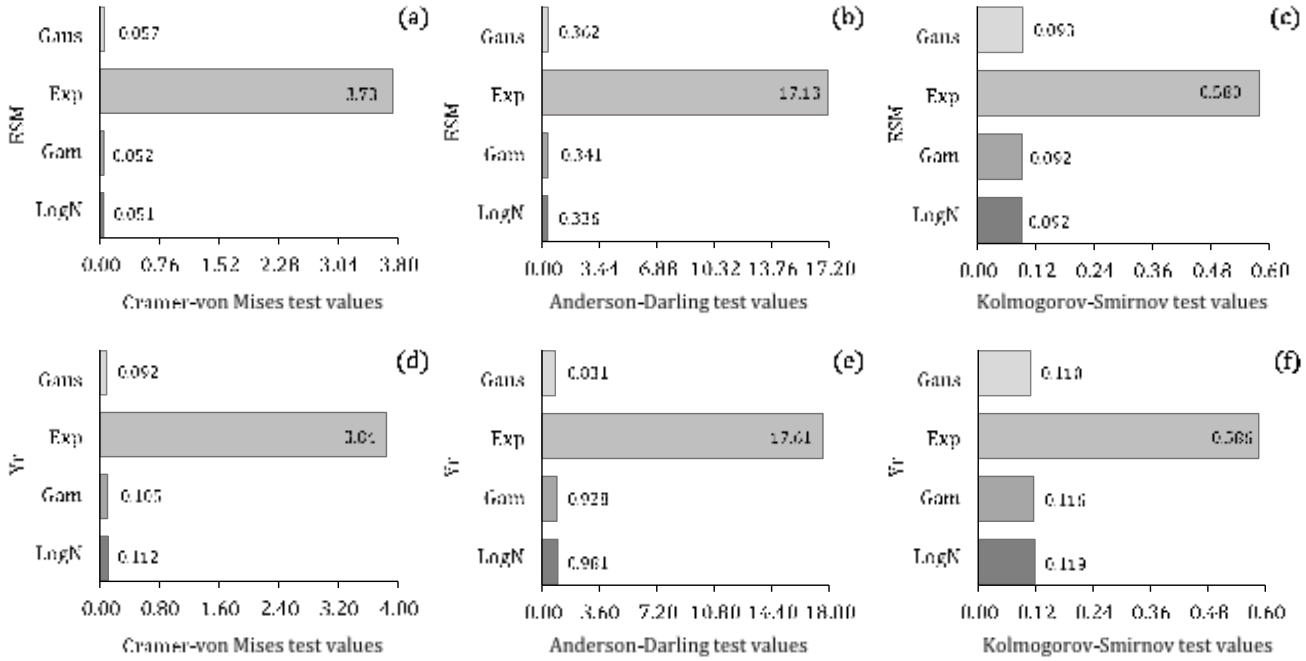
Source: Authors.

Figure 9. Cramer-von Mises (panels a and d), Anderson-Darling (panels b and e), Kolmogorov-Smirnov (panels c and f) test values for the traits yolk mass (YM; panels a to c) and albumen mass (AM; panels d to f) of laying quails fitted by Fitdist procedure with the quantitative distributions: Gaussian, Exponential, Gamma and LogNomal.



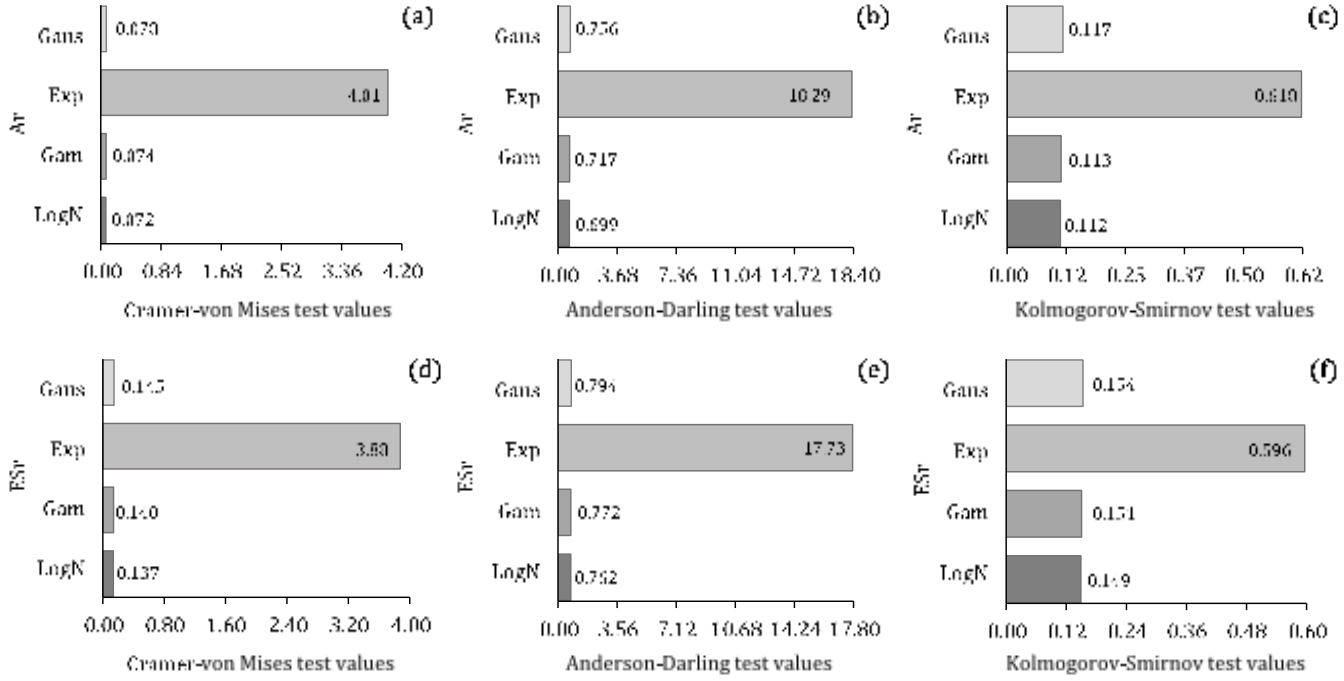
Source: Authors.

Figure 10. Cramer-von Mises (panels a and d), Anderson-Darling (panels b and e), Kolmogorov-Smirnov (panels c and f) test values for the traits eggshell mass (ESM; panels a to c) and yolk ratio (Yr; panels d to f) of laying quails fitted by Fitdist procedure with the quantitative distributions: Gaussian, Exponential, Gamma and LogNomal.



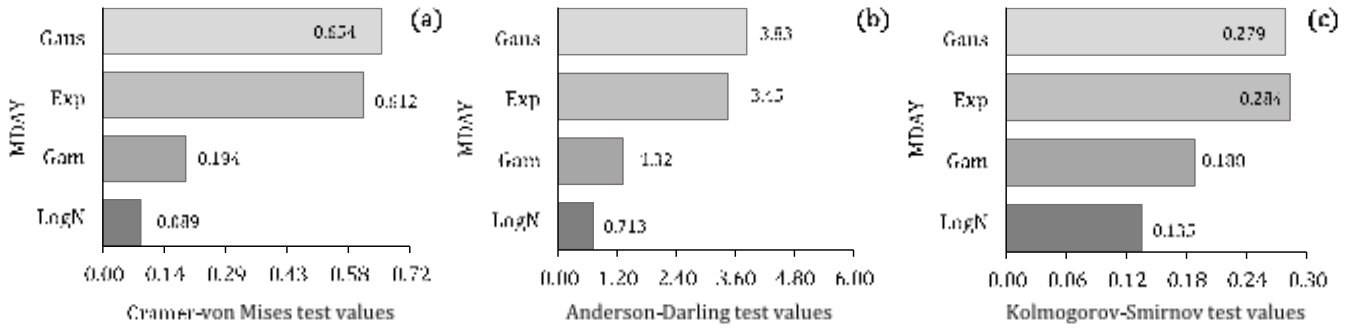
Source: Authors.

Figure 11. Cramer-von Mises (panels a and d), Anderson-Darling (panels b and e), Kolmogorov-Smirnov (panels c and f) test values for the traits albumen ratio (Ar; panels a to c) and eggshell ratio (ESr; panels d to f) of laying quails fitted by Fitdist procedure with the quantitative distributions: Gaussian, Exponential, Gamma and LogNomal.



Source: Authors.

Figure 12. Cramer-von Mises (panel a), Anderson-Darling (panel b), Kolmogorov-Smirnov (panel c) test values for the variable malondialdehyde in yolk (MDAY) of laying quails fitted by Fitdist procedure with the quantitative distributions: Gaussian, Exponential, Gamma and LogNomal.



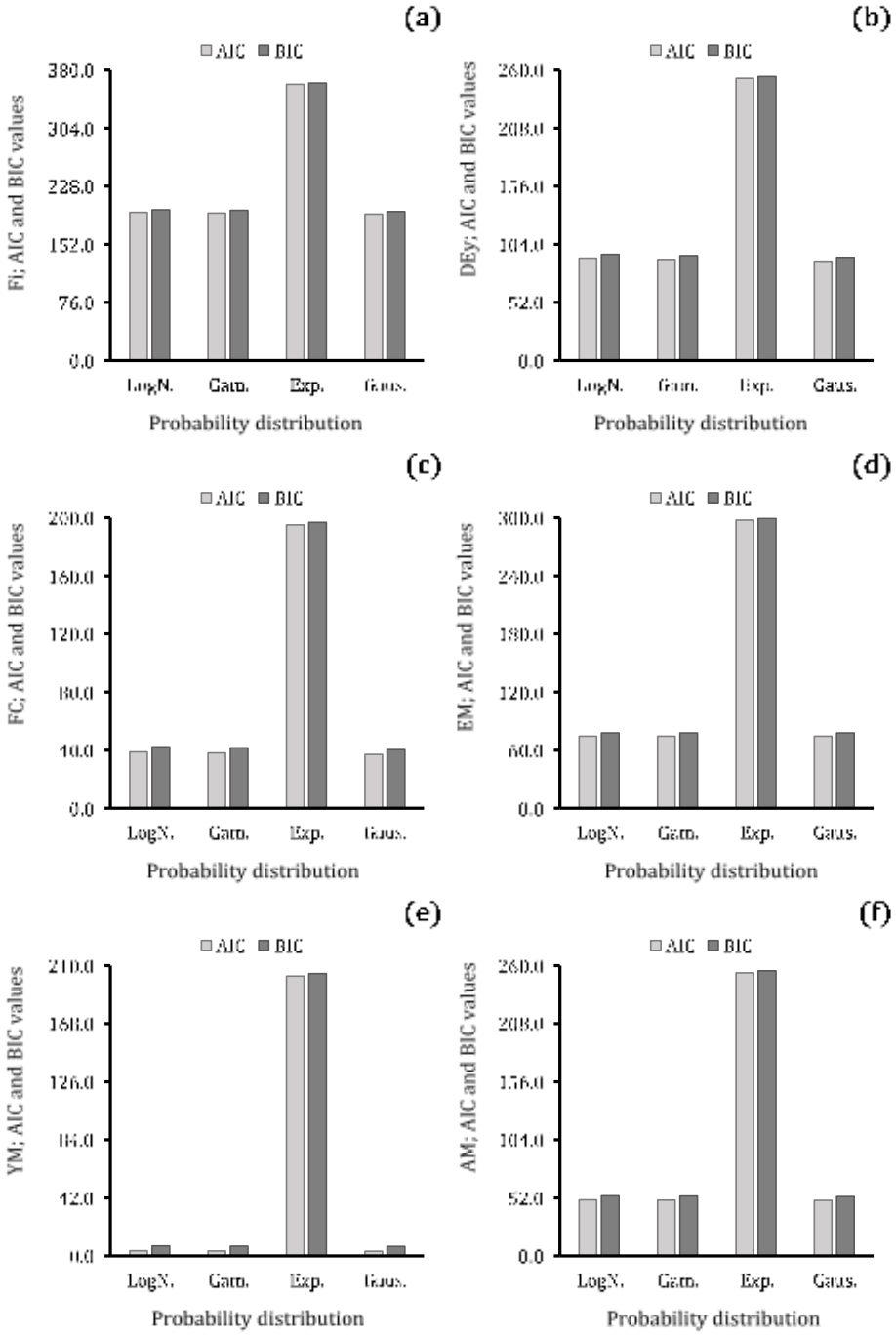
Source: Authors.

3.3 Information criteria (AIC and BIC)

The values of AIC and BIC were similar, which led to equal decisions for all traits except for the variable MDAY (Figures 13 and 14).

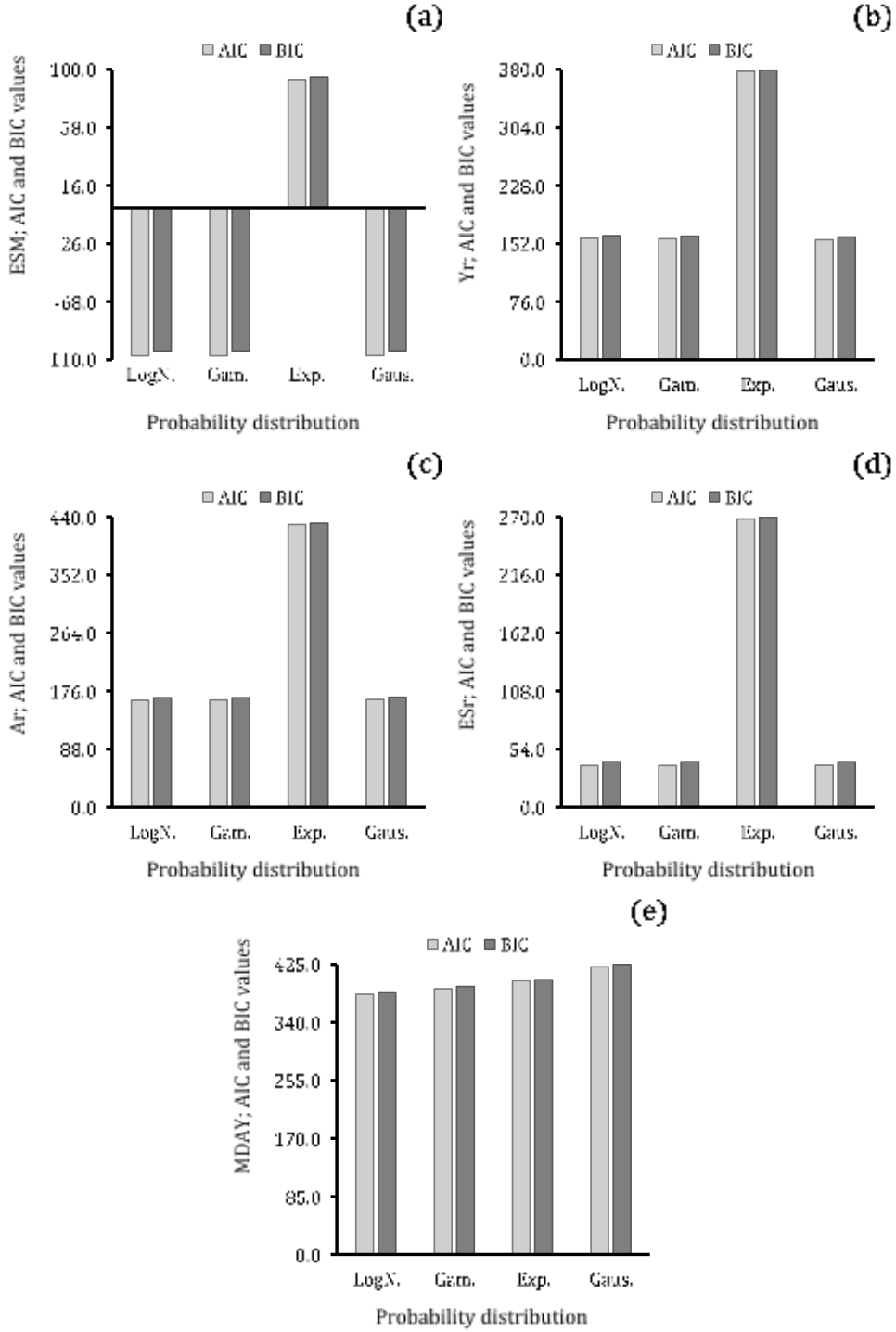
Similar to what was observed in the tests, the exponential distribution was evidently the worst than the other distributions (Figures 13 and 14), with the exception of the variable MDAY (Figure 14, panel e), which the Gaussian had the highest values of AIC and BIC.

Figure 13. Values of the Akaike (AIC) and Bayesian (BIC) information criteria from the analysis of the traits feed intake (Fi; panel a), daily eggs yield (Dey; panel b), feed conversion (FC; panel c), eggs mass (EM; panel d), yolk mass (YM; panel e) and albumen mass (AM, panel f), of laying quails with the Lognormal (LogN.), Gamma (Gam.), Exponential (Exp.) and Gaussian (Gaus.) probability distributions.



Source: Authors.

Figure 14. Figure. Values of the Akaike (AIC) and Bayesian (BIC) information criteria from the analysis of the traits eggshell mass (ESM; a), yolk ratio (Yr; b), albumen ratio (Ar; c), eggshell ratio (ESr; d), and malondialdehyde in yolk (MDAY; e) of laying quails, with the Lognormal (LogN.), Gamma (Gam.), Exponential (Exp.) and Gaussian (Gaus.) probability distributions.



Source: Authors.

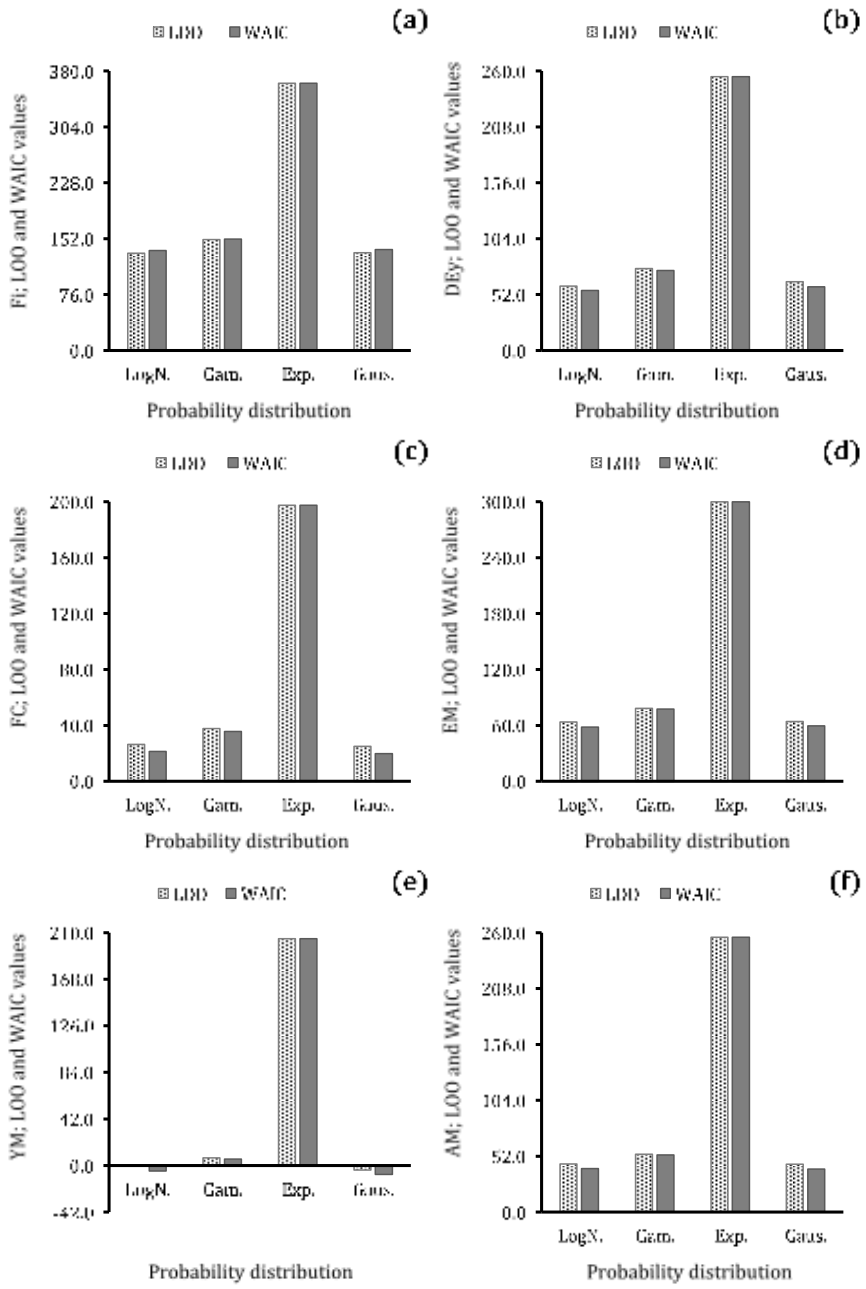
3.4 Information criteria (WAIC and LOO)

The behavior of the values of WAIC and LOO (Figures 15 and 16) were like the values of AIC and BIC. For the WAIC, the traits Fi, EM, FC, AM and Yr, showed smaller values when using the Gaussian distribution. However, the MDAY, YM, and

ESM have already indicated the gamma distribution as the most likelihood. For DEy, ESr, and Ar the smaller WAIC were found in the lognormal distribution.

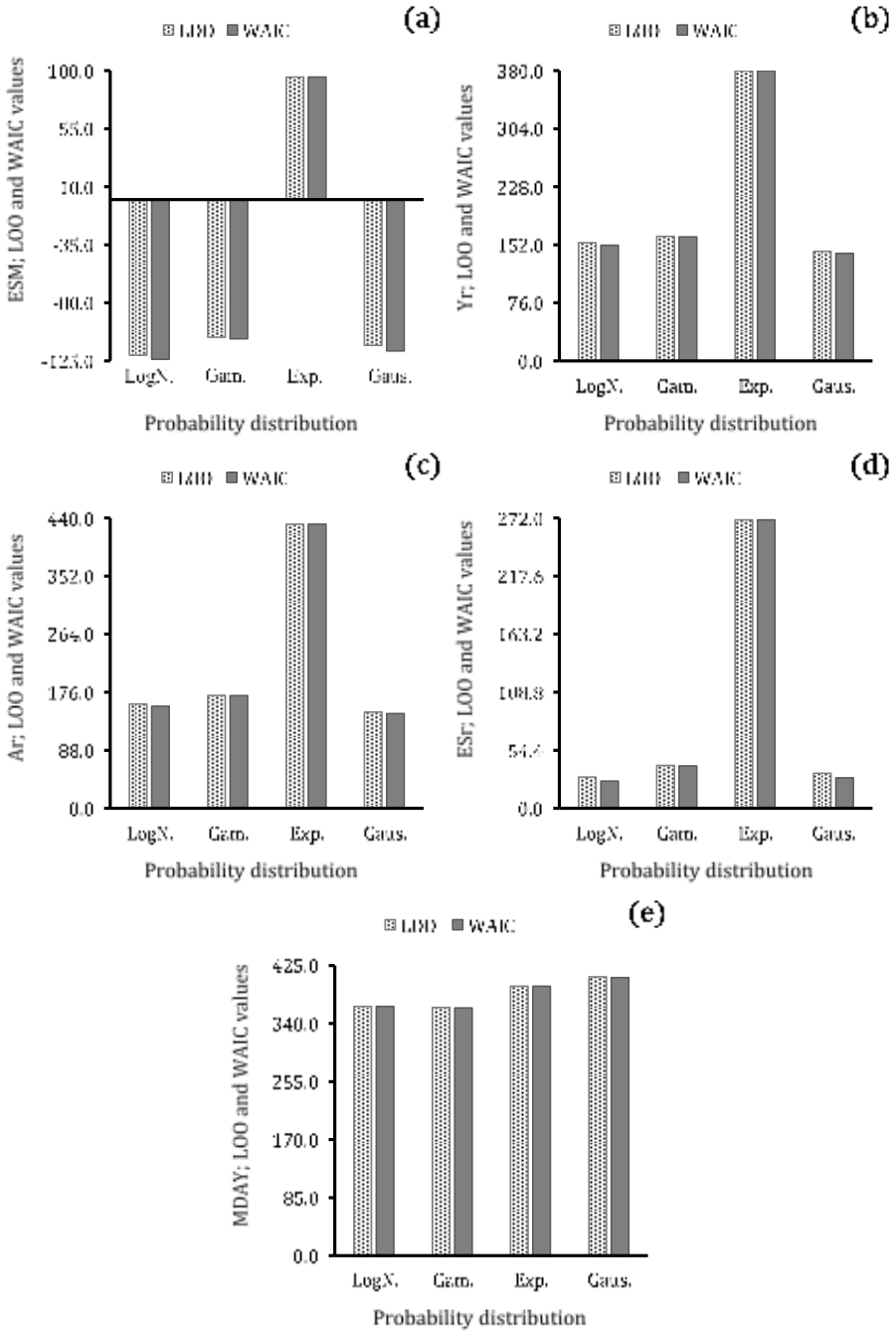
According to LOO, the Gaussian distribution was most likelihood for Fi, EM, FC, AM, and Yr, the gamma distribution for MDAY and YM, and the lognormal for DEy, ESr, and Ar. The ESM traits showed the smaller and identical LOO values for the gamma and Gaussian distributions.

Figure 15. Values of the Leave-one-out (LOO) and Watanabe–Akaike information criterion (WAIC) from the analysis of the traits feed intake (Fi; panel a), daily eggs yield (Dey; panel b), feed conversion (FC; panel c), eggs mass (EM; panel d), yolk mass (YM; panel e) and albumen mass (AM, panel f), of laying quails with the Lognormal (LogN.), Gamma (Gam.), Exponential (Exp.) and Gaussian (Gaus.) probability distributions.



Source: Authors.

Figure 16. Values of the Leave-one-out (LOO) and Watanabe–Akaike information criterion (WAIC) from the analysis of the traits eggshell mass (ESM; a), yolk ratio (Yr; b), albumen ratio (Ar; c), eggshell ratio (ESr; d), and malondialdehyde in yolk (MDAY; e) of laying quails, with the Lognormal (LogN.), Gamma (Gam.), Exponential (Exp.) and Gaussian (Gaus.) probability distributions.



Source: Authors.

4. Discussion

The choice of the most feasible distribution is made by approximating the observed data to the distributions. A non-zero asymmetry shows the lack of symmetry of the distribution concerning the observed data (Muller & Dutang, 2015). For the

MDAY characteristic, the data were equally distant in all distributions, indicating no difference between them. For the others, the exponential distribution was the one that least converged.

The data evaluation using the Akaike and Bayes information criteria are the most indicated, but they should be used with caution, because the choice of which criterion will be used depends on the characteristics of the data, the number of observations, missed observations, among others. In situations when the number of observations is relatively large, both the AIC and BIC will tend to the same model (Burnham & Anderson, 2004). In our work, we found that both criteria had the smaller values for the same distributions, and the exponential distribution had the greatest values, thus it is not indicated for the evaluated characteristics.

The R^2_{ac} and r_{ac} coefficient were used in their conditional forms (considering fixed and random effects) and corrected for the number of model parameters. R^2_{ac} indicates the degree to which the predicted value is associated with the observed value, its range is between 0 and 1, which 1 indicates the perfect fit and 0 that there is no correlation. Therefore, the higher the R^2_{ac} value, the greater the relationship between the independent variables (X_1, X_2, …, and X_n) and the dependent variable (Y) (Vonesh et al., 1996). If to increase the number of variables in the model, R^2_{ac} can increase, decrease, and even be negative (Gujarati, 2009). In this study, it was possible to observe R^2_{ac} with negative values (Figures 2 to 5). Another way to explain theses negative values is by the square soma of model (axis-x) is so much higher that observed data (axis-y), i. e, the model using some distribution fits very poorly to the data of a given variable.

The r_{ac} is a measure used to measure the degree of association between variables, that is, e.g.: X_1; X_2; …; X_n, and Y are related, their covariance, the similarity between variables (Vonesh et al., 1996). For $r = 1$ the relationship is positive and perfect; for $r = -1$ the relationship is negative and perfect; and for r = 0 there is no relationship, or the correlation is not linear (Mukaka, 2012). Both R^2_{ac} and r_{ac} are used to inform the correlation between variables, and not the fit quality of the model (Vonesh, 1996).

The Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling tests should be used when the number of parameters of the model is known, not allowing the comparison of results when the number of parameters is different (Muller and Dutang, 2015), which is opposed to the principle of parsimony (Burnham & Anderson, 2004). Anderson-Darling statistics are important when emphasizing the tails of candidate distributions (Muller & Dutang, 2015). Thus, as well as the AIC and BIC criteria, the lower their value, the better the quality of fit of the distribution to the data (D'Agostino and Stephens, 1986; Muller & Dutang, 2015). The results found in the tests corroborate the results obtained for AIC and BIC, indicating the Gaussian and lognormal distributions as the most appropriate, suggesting that both evaluations, either by tests or criteria, are efficient for choosing the distribution for the evaluated characteristics.

WAIC and LOO have great flexibility in their use. In addition to being a multi-model selection information criterion, the use of the Bayesian is the adjustment of the complete experimental model including all effects simultaneously with the evaluation of the best likelihood function that fits the data. They have a wide variety of probability distributions (gamma, binomial, exponential, Gaussian, among others), allowing the users to enter information they already have about their variables as well as the inclusion of the regression model that is more flexible manually and simple to change. Both indexes are less biased and less robust than the deviance Information Criterion (DIC), also widely used in the selection of models by the Bayesian method (Bürkner, 2017).

For MDAY, the WAIC and LOO criteria have the smaller value the gamma distribution. However, the values we found are close to the values for lognormal distribution, converging to the results obtained with the Akaike and Bayes criteria, and AD, KS, and CvM tests.

In the evaluation of the characteristics DEy, Fi, EM, FC, YM, ESM, AM, Yr, ESr, and AP, the criteria, and tests (AIC, BIC, AD, KS, CvM, WAIC, and LOO) indicated lognormal, gamma and Gaussian as possible distributions, presenting values with little variation between them, and the exponential distribution as the least fitted.

5. Conclusion

Preliminary assessments that aim to identify the statistical model that best represents the reality of the data are essential to ensure the quality of subsequent statistical analyses, to avoid under or overestimation of results, and to reduce errors resulting from the use of an inappropriate distribution.

For future work, we suggest further studies on the STAN Bayesian approach, in which a complete model with all variables and distribution together can be evaluated.

Acknowledgments

The authors wish to thank the Universidade Estadual do Norte Fluminense – Darcy Ribeiro and the Fundação Osvaldo Cruz (Fiocruz) for the partnership and promoting conditions for the development of the research. To CAPES, for the financing.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716-723.

Barreto, S. T. L, Pereira, C. A, Umigi, R. T, Rocha, T. C, Araujo, M. S, Silva, C. S. & Filho, R. A. T. (2007). Determinação da exigência nutricional de cálcio de codornas japonesa na fase inicial do ciclo. *Revista Brasileira de Zootecnia*, 36; 68-78p.

Brito, A. De L., Júnior, S. F. X., Mendonça, E. B. de, Xavier, E. F. M., Santos, T. T. de M. & Oliveira, T. A. de. (2020). Adjustment of Fragility Models and Proportional Risks Applied to Diabetic Retinopathy Data. *Research, Society and Development*, 9(8). http://dx.doi.org/10.33448/rsd-v9i8.5691.

Burnham K. P. & Anderson D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*. 33, 261-304. 10.1177/0049124104268644.

Bolker B. M., Brooks M. E., Clark C.J., Geange S.W., Poulsen J.R., Stevens M. H. H. & White J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*. 24, 127-135. 10.1016/j.tree.2008.10.008.

Bürkner P. C. (2017). brms: An R package for bayesian multilevel models using stan. *Journal of Statistical Software*. 80, 1-28. 10.18637/jss.v080.i01.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancour, M., Brubaker, M. A., Guo, J., Li, P., & Ridell, A. (2017). "Stan: A Probabilistic Programming Language.". *Journal of Statistical Software*, 76(1), 1–32. 10.18637/jss.v076.i01.

D' Agostino R. B. & Stephens M. A. (1986). Godnees-of-fit techniques. *Statistics: Textbooks and monographs*. Departament of Statistics Southern Methodist University Dallas, Texas.

Darling, D. A. (1957). The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics*, 28(4), 823-838.

Enkvetchkul B., Anthony N. B. & Bottje, W. G. (1995). Liver and blood glutathione in male broiler chickens, turkeys, and quail. *Poultry Science*. 74, 885-889. 10.3382/ps.0740885.

Gelman, A., Hwang, J. & Vehtari, A. (2014***).*** Understanding predictive information criteria for Bayesian models. *Statistics and computing*, 24(6), 997-1016.

Gujarati, D. N. (2009) Basic Endometrics. *The McGraw-Hili Companies, Inc*. (4a ed.), New Delhi.

Massey Jr. F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253), 68-78.

Mukaka M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal: the journal of Medical Association of Malawi*, 24(3), 69–71.

Laio F. (2004). Cramer-von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters. *Water Resources Research*. 40, W09308. 10.1029/2004WR003204.

Lüdke, M. & André, M. E. D. A. (1986). Pesquisa em Educação: Abordagens Qualitativas. *Temas Básicos de Educação e Ensino* (E.P.U.).

Muller, M. L. D. & Dutang, C. (2015). Fitdristrplus: An R package for fitting distributions. *Journal of Statistical Software*. 64, 1-34. 10.18637/jss.v064.i04.

Neter J. Wasserman. W., & Kutner M. H. (1985). Applied linear statistical models: regression, analysis of variance, and experimental designs. *RD Irwin, Homewood.*

Oliveira, A. M., Furlan, A. C. Murakami., A. E. et al. (1999). Exigência nutricional de lisina para codornas japonesas (Coturnix coturnix japonica) em postura**.** *Revista Brasileira de Zootecnia*, 28, 550 1050-1053p.

Pan W. (2001). Akaike's information criterion in generalized estimating equations. division of biostatistics**.** *Biometrics*. 57, 120-125. 10.1111/j.0006-341x.2001.00120.

Pereira, A. S., Shitsuka, D. M., Parreira, F. J. & Shitsuka, R. (2018). *Metodologia da Pesquisa Científica*. Universidade Federal de Santa Maria, RS.

Silva, R. B. Z da, Silva, R. N. Z. Da, Aires, F. F. Da C. & Soares, E. J. O. (2019). The use of time series models for forecast corn production in Mato Grosso state. *Research, Society and Development*, 9, n.1. http://dx.doi.org/10.33448/rsd-v9i1.1915.

Silva, G. de L. P. E, Geraldine, R. M., Santana, R. F., Bento, J. A. C., Neto, M. A. de S. & Caliari, M. (2020). Simulation of the production process of a mineral water industry by system dynamics method. *Research, Society and Development,* 9(7). http://dx.doi.org/10.33448/rsd-v9i7.4729.

Shahryar H. A., Salamatdoust R., Chekane-Azar S. & Ahadi F. M Vahddatpoor T. (2010). Lipid oxidation in fresh and stored eggs enriched with dietary ω3 and ω6 polyunsaturated fatty acids and vitamin E and A dosages. *African Journal of Biotechnology*. .9, 1827-1832. 10.5897/AJB10.1482.

Schwarz, G. (1978). Estimating the dimension of a model**.** *The annals of statistics,* 461-464.

Sher, V., Bemis, K. G., Liccardi, I., & Chen, M. (2017). An empirical study on the reliability of perceiving correlation indices using scatterplots. *In Computer Graphics Forum*. 36(3), 61-72).

Stan Development Team (2017). Stan Modeling Language: User's Guide and Reference *Manual.* URL http://mc-stan.org/manual.html.

Vonesh, E. F. (2014***).*** Generalized linear and nonlinear models for correlated data: theory and applications using SAS. *SAS Institute.*

Vonesh E. F., Chinchilli V. M. & Pu K. (1996). Goodness-of-fit in generalized nonlinear mixed-effects models**.** *Biometrics*. 52, 572-87. 0.2307/2532896.