Original Research Article

Biometrical assessment of urucum (*Bixa orellana* L) seed quantitative features in the northeastern Brazil

ABSTRACT

Urucum (*Bixa orellana* L.) is a shrub plant occurring in the caatinga biome. This study evaluated the biometrical assessment of quantitative features of Urucum in northeastern Brazil. Fruit samples were taken from an area of native vegetation of the Federal Rural Semi-Arid University (UFERSA), Mossoró, RN. Measurements on seed length and width of fruit samples were done at the plant science laboratory. Descriptive and graphical analysis were carried out using the R statistical software. Seed length and width showed a small range of variation with an excellent and reasonable values of coefficient of variation, respectively. A regular degree of symmetry and a mesokurtic distribution were observed for seed length and width. There was non-significant linear correlation between both variables which did not fit to the normal probability distribution.

Key words: Medicinal plant; descriptive statistics; inference.

1. INTRODUCTION

The Urucum (*Bixa orellana* L.), a shrub species native to tropical region of America, has gained importance in the world agricultural market every year and is cultivated in the tropical areas all over the world [1]. The fruits are ovoid capsules containing 30 to 40 seeds [2]. The Urucum produces particular toxic coloring substances with a hypolipidemic pigments that is allowed by the World Health Organization [3].

The species has multiple uses for ornamental and medicinal values; and soil restorer properties. It is a pioneer species, native to the Amazon Rainforest. Its seeds are used to produce condiments and tinctures that are widely used in the industry. The fast growth nature of Urucum allows its cultivation together with other species in degraded areas destined to reforestation [2].

Morphological analysis of fruits and seeds help to understand the germination process and vigor and viability characteristics [4]. Biometric analysis comprises an essential tool to detect genetic variability within and between populations and to define the relationships between the variability and environmental factors, thus contributing to genetic improvement programs [5].

Biometrics of fruits and seed are central in studies of species, providing data for differentiation of species within the same genus. For example, Carpanezzi and Marques [6] showed that the seeds of *Hymenaea courbaril L* weight almost twice the seeds of *Heritiera* parvifolia. Also, the seeds characteristics influence its patterns of dispersion and establishment of seedlings [7],

which are used to differentiate pioneer and non-pioneer species in tropical forests [8]. In most shrubs and trees, there is an inverse relation between seed size and the number of seeds per fruit [9].

Urucum stands out for its therapeutic and food properties. However, there is no record of its exploration in the region of Mossoró, RN.

This work evaluated the biometry of Urucum seeds as a subsidy studies for comparing their characteristics when submitted to different environments.

2. MATERIAL AND METHODS

The study was carried out in Mossoró, RN, geographic coordinates: 5º11'S and 37º20'W at 18 m of altitude, with an annual mean temperature of 27.5°C and relative humidity of 68.9% [10]. The climate in Mossoró is BSwh', hot and dry. In total, 300 seed samples were collected from the native vegetation of the Federal Rural Semi-Arid University (UFERSA), Brazil during May 2018. The morphological characterization of Uricum seed was evaluated in the plant breeding laboratory. The width and length of well-developed seeds were determined with the precision of a 0.1 mm pachymeter. The descriptive analysis and graphs were done with the software R version 3.1.1. [11].

3. RESULTS

To determine and compare quantitative aspects of the distributions of values of seeds length and width based the analysis on specialized biometric literature [12-18]. Thus, adopted the exploratory data analysis using frequency distributions, box plots, as well as the statistical estimators of the variables under study, which are the main descriptive and inferential statistical measures, such as arithmetic mean, median, total range, variance, standard deviation, standard error of mean, coefficient of variation, asymmetry coefficient, kurtosis coefficient, quartiles and interquartile deviation. We used the Pearson correlation coefficient and statistical inference such as the hypothesis tests, T-test or Z-test, at a significance level of 5% probability, based on Student's t-distribution and Normal distribution, respectively, for the construction of confidence intervals with 95% probability (Tables 1 to 3 and Figures 1 to 7).

Table 1. Frequency distribution of urucum seed length (mm).

Class (Length in mm)	f_i	X_i	f %
[2.9---------3.2)	2	3.1	0.67
[3.2---------3.4)	5	3.3	1.67
[3.4---------3.7)	9	3.5	3.00

[3.7---------3.9)	21	3.8	7.00
[3.9---------4.1)	45	4.0	15.00
[4.1---------4.4)	72	4.3	24.00
[4.4---------4.6)	84	4.5	28.00
[4.6---------4.9)	51	4.7	17.00
[4.9---------5.1)	11	5.0	3.67
Total	300	--------	100

Table 2. Frequency distribution of urucum seed width (mm)

Classes (Width in mm)	f_i	X_i	***f* %**
[1.7----------1.9)	4	1.8	1.33 %
[1.9----------2.2)	28	2.1	9.33
[2.2----------2.5)	41	2.4	13.67
[2.5----------2.8)	47	2.6	15.67
[2.8----------3.1)	58	2.9	19.33
[3.1----------3.3)	44	3.2	14.67
[3.3----------3.6)	32	3.5	10.67
[3.6----------3.9)	33	3.8	11.00
[3.9----------4.2)	13	4.0	4.33
Total	300	--------	100

Table 3. Results of the descriptive and inferential statistics of length (mm) and width (mm) of 300 urucum seeds.

Statistics or Estimator	Length	Width
Sample size (number of seeds)	300	300
Minimum value	2.94	1.65
Maximum value	5.05	4.17
Total range	2.11	2.52
Arithmetic mean	4.32	2.94
Median or second quartile	4.38	2.94
Variance	0.13	0.33
First quartile	4.13	2.51
Third quartile	4.57	3.36
Standard deviation	0.37	0.57
Standard Error	0.02	0.03
Coefficient of Variation (%)	8.44	19.40
Asymmetry Coefficient	-0.85	0.08
Kurtosis Coefficient	1.12	-0.86

Interquartile Range (IR)	0.44	0.85
Z Test for mean, at 0.1% of probability	216.00***	98.00***
Confidence interval for mean at 95% of probability	4.29 to 4.36	2.88 to 3.00
Theoretical Normal Distribution or Gaussian Probabilities, D'Agostino-Pearson's Test	*p*-value = 0.01 Fitted to normal distribution	*p*-value = 0.06 Fitted to normal distribution
Pearson's Simple Linear Correlation Coefficient (r)	+ 0.16 (p-value of Student's T test = 0.10)	

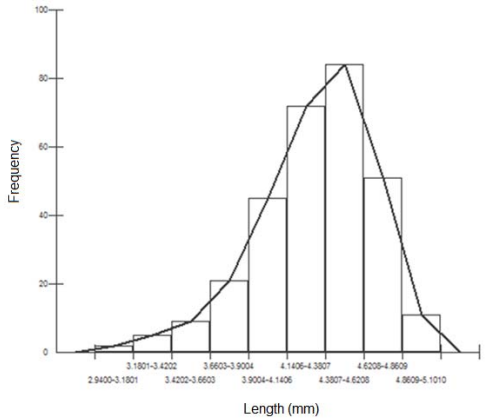

Figure 1. Histogram and polygon of frequencies representative of the length distribution (mm) of urucum seeds.

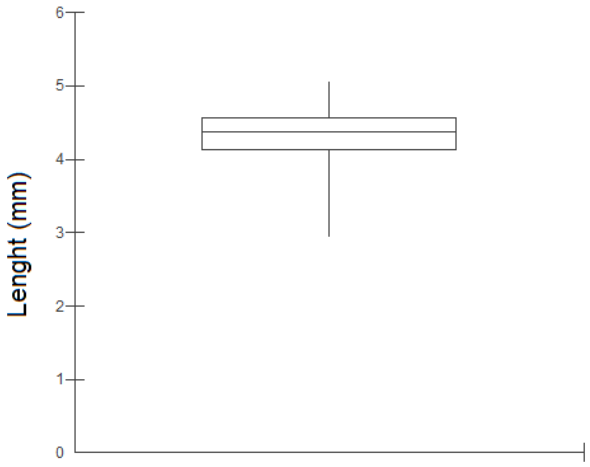

Figure 2. Box plot using median and quartiles of length of urucum seeds (mm).

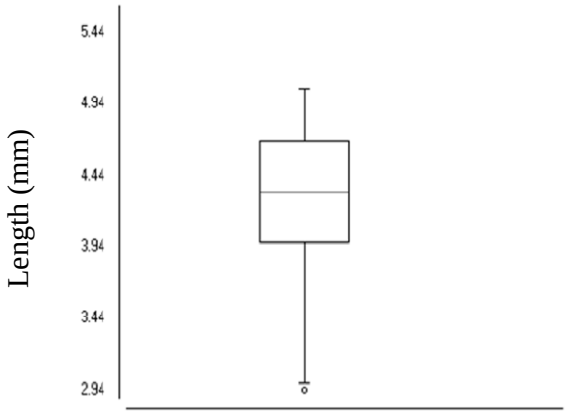

Figure 3. Box plot using the mean and standard deviation of urucum seeds length (mm).

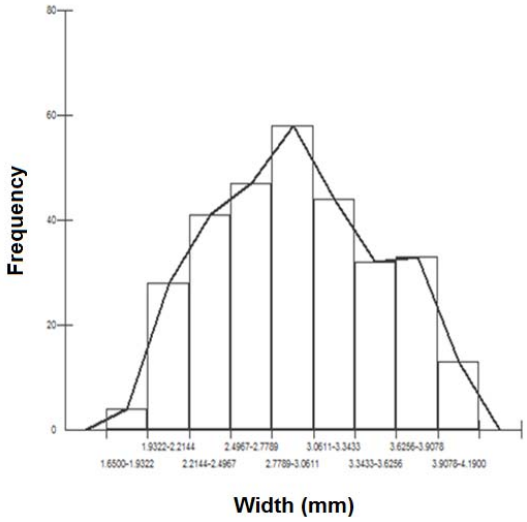

Figure 4. Histogram and polygon of frequencies of the width (mm) distribution of urucum seeds.

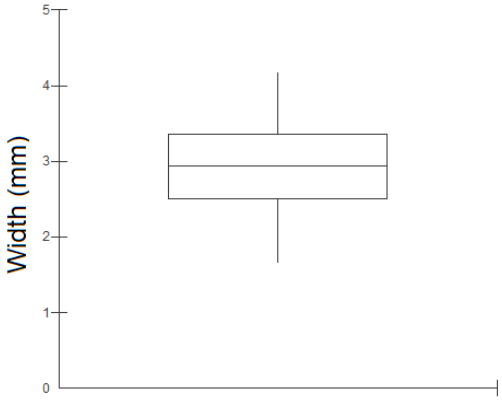

Figure 5. Box plot using median and quartiles of urucum seed width (mm).

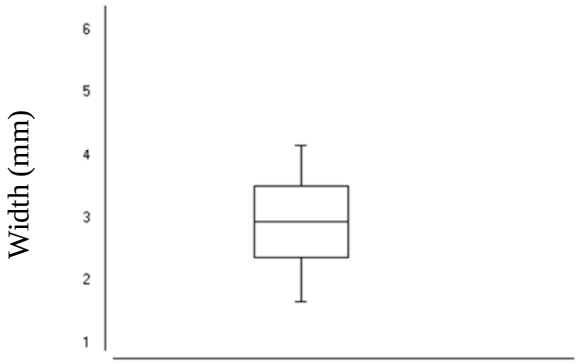

Figure 6. Box plot using the mean and standard deviation of urucum seed width (mm).

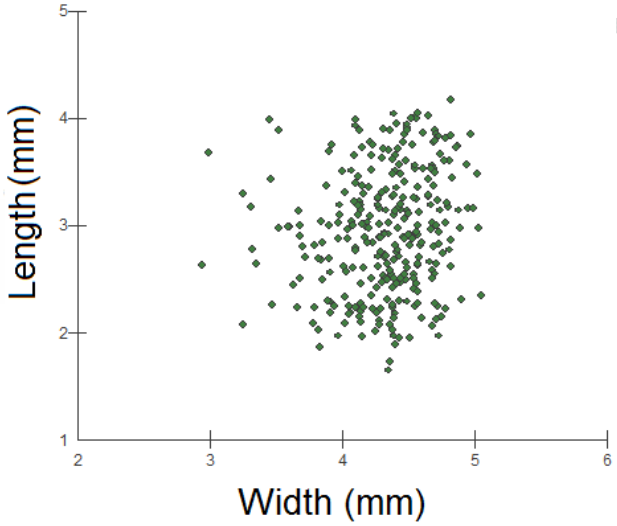

Figure 7. Pearson linear correlation dispersion diagram for length and width (mm) of 300 urucum seeds.

4. DISCUSSION

The length and width of seeds had a small magnitude of variation (Table 3) and good coefficients of variation, showing a high degree of homogeneity. Data dispersion was higher for width than length. There is not a great similarity in the location of the length and the width of the seeds, since the means 4.32 and 2.94 mm for the average length and the average width, respectively, are not very close. It is worth noting that the calculation of the average does not have to make real sense. However, it can be stated that the seeds have a center of gravity in the distribution of these variables, which shows that the length of the seeds is close to 4.32 millimeters and the width around 2.94 millimeters. The relative variation between the variables measured by the coefficients of

variation, in this case, was the most adequate, since they are independent of the magnitude of the variable, as well as of its unit of measure, being, therefore, the most correct procedure. The coefficients of variation (8.44 e 19.40% for length and width, respectively) showed a similar or close relative dispersion for the two characteristics. This coefficient was the most adequate to measure data variability since they are independent of the magnitude of variables and their units. Even though there was less relative variability for the length with a significant difference in the magnitude compared to the width [12,14,16,19].

Other results obtained, as displayed in Table 3, where that the mean and median were very close to the length and equal to the width, which indicated a reasonable and a high degree of symmetry for length and width, respectively, which were -0.85 and 0.08 for the length and width, respectively. However, the momentum coefficients of kurtosis, which were 1.12 and -0.86, respectively, showed that both length and width had a mesocuric distribution, and the characterization of the degree of asymmetry and kurtosis of a distribution according to [13,16,19, 20-21], cannot be done by looking only at measures of position or central tendency such as the average and the median, but also through the asymmetry and kurtosis coefficients, as well as through histograms and frequency polygons (Figures 1 and 4).

With respect to quantiles, the first quartile, 25% of the lowest values, showed a maximum value of 13.24 mm length and 9.78 mm for width. The 25% larger seed lengths and widths had at least 14.67 and 10.97 mm, respectively, corroborating data of Figueiredo *et al* [22]. The interquartile range verifies the dispersion of the data concerning the median and thus, identify the presence of outliers in the data. The interquartile was 0.44 mm for the length and 0.85 mm for the width of the seed (Table 3).

The degree of association or simple linear correlation, whose values range from -1 to + 1, including zero, and its magnitude is dimensionless, between length and seed width of urucum was 0.16 (Table 3), showing a non-significant positive relation, as evidenced by the result of Student's t-test [12-18,22-23].

From a graphic, visual or geometric point of view, the dispersion diagram (Figure 7), showed high dispersion which confirms the lack of correlation between length and width found in the low value of correlation coefficient (0.16). The length and width variables of urucum seed did not fit to the theoretical or special density distribution of normal or bivariate Gaussian probability so that all the conclusions obtained through statistical inference are not guaranteed or assured by this assumption (Table 3).

Reporting the quantiles comprises an excellent way to illustrate the dispersion of a distribution. Researchers are more familiar with the type of quantile called percentile because of its use in standardized tests. When a test score is reported within the 90th percentile, 90% of the scores

are smaller than it, and 10% are higher. Contrary to variance and standard deviation, quantile values do not depend on arithmetic mean or median values. When distributions are asymmetric or have outliers, which are extreme values that are not characteristic of the distribution quantile box plots can portray the distribution of data more accurately than mean and standard deviation.

High values were obtained at the Z test, both for seed length and seed width, concluding that the mean values of these characteristics were highly significant (Table 3).

Boxplots, using the median and the quartiles are a widely used chart in the biological and medical sciences, showing the median, first and third quartiles. It also shows the lowest and highest scores through the lower and upper boundaries of vertical straight lines, which originate from the first and third quartiles, respectively. According to the results obtained in Figure 2, using the median and quartiles, a strong concentration of the seed length data in millimeters around the central value in the case the median was observed, which evidenced a high homogeneity of these observations, which makes its analysis and current and subsequent interpretation, including the making of statistical inferences such as the construction of confidence intervals and the application of hypothesis test tests, as well as the adjustment of regression models for estimation and forecasting purposes.

Using the mean and standard deviation, the box plot graph, similar to the previous graph, shows the mean and standard deviation in the Box. It also shows the lowest and highest scores through the upper and lower boundary of vertical straight lines where the presence of outliers can also be verified. (Figure 3). In the case of seed length, only an atypical value occurred, showing a high similarity among the grouped observations, except for the presence of this unusual observation. Also according to the results obtained in Figure 3, most seed length data clumped around the central value, in the case the mean, which suggests a high homogeneity of these observations.

A strong concentration of seed width data around the central value (median) was observed, which evidenced a high homogeneity of data collected (Figure 5). Using the mean and standard deviation (Figure 6), there was no atypical value, showing a relevant similarity in the grouped observations and a strong concentration of the seed width data around the central value (mean), which also evidenced the high homogeneity.

In general, the results of the descriptive measures of location, variability, asymmetry and kurtosis can serve as a basis for future studies of descriptive analysis and statistical inference, for the comparison of different environments, genetic improvement studies, grouping of experiments in joint analysis, in stability analysis of cultivars, as well as in the construction of so-called components of variance [12-18, 22-23].

Standard deviation and variance are special cases of what statisticians and physicists call the central momentum (CM). Central moment comprises the mean deviation of all observations in a data set from the mean of the observations, raised to the power r. The first central moment (r = 1) measures the sum of the differences between each observation minus the sample mean (arithmetic), which is always equal to zero. The second central moment (r = 2) is the variance. The third central moment (r = 3), divided by the standard deviation to the cube (s^3), is the asymmetry. The asymmetry describes how the sample differs from the form of a symmetric distribution. A normal distribution has an asymmetry coefficient equal to zero. A distribution in which the value of the asymmetry coefficient is greater than zero has asymmetry to the right, that is, there is a long tail of larger observations at the right of the mean. However, if the asymmetry coefficient is less than zero it has asymmetry on the left, there is a long tail of smaller observations to the left of the mean. The kurtosis or flattening has its basis in the fourth central moment ($r = 4$), measuring the extent or peak at which the probability density is distributed in the tails versus the center of the distribution. The distribution is classified as heavy tail or light tail when compared to a standard normal distribution (mesokurtic). Aggregate or platykurtic (flattened) distributions have a kurtosis coefficient less than zero, compared to the normal distribution, meaning more mass of probability in the center of the distribution and less probability at the tails. In contrast, leptokurtic (tapered) distributions have a kurtosis coefficient greater than zero. Leptokurtic distributions have less mass of probability in the center and tails of relatively heavy probabilities [24].

According to Oliveira [14], the law of large numbers proves that for an infinitely large number of observations, the formula $\frac{\sum_{i=1}^{n} Y_i}{n}$ is an approximation of the population mean μ, where $Y_n = [Y_i]$ is the sample of size n of a random variable Y with expected value (E (Y)). Similarly, the variance of $y_n = \frac{\sigma^2}{n}$. Since the standard deviation is just the square root of the variance, y_n is given by $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$, which is the same as the standard error of the mean. Therefore, we have an estimate of the standard deviation of the population mean, which is the standard error.

If the conclusions based on a single sample are representative of the entire population, it is recommended using the standard error of the mean. However, if the samples limit the conclusions, it is better to use the sample standard deviation.

Large observational surveys covering large spatial scales with a substantial number of samples are likely representative of the population of interest as a whole, and the standard error of the mean should be used. Small, controlled experiments with few replicates are likely based on a

single cluster, and possibly unrepresentative of individuals, as a consequence, the standard deviation should be used to characterize or measure the degree of absolute dispersion of the samples.

5. CONCLUSION

From this study, Urucum seed length and width showed a low amplitude of variation, as well as an excellent value for the coefficient of variation, which gave a reasonable degree of homogeneity. However, seed length was less dispersed than seed width.

There was a regular to high degree of symmetry and a mesokurtic distribution for seed length and width which characteristics had no significant correlation and showed mean values with highly significant differences.

The data of urucum seed biometry did not fit to the normal distribution probability. Results from descriptive measurements regarding location, variability, asymmetry, and kurtosis can aid in future studies of descriptive analysis and statistical inference. Such studies could compare different environments or plant genotypes, and involve criteria used for grouping of experiments in statistical analysis, stability test of cultivars in multivariate analysis, as well as support the construction of so-called components of variance.

References

1. Mercadante AZ, Pfander H. Carotenoids from annatto: a review. Recent Research Developments in Agriculture and Food Chemistry. 1998; 2 (1): 79-91.
Available:htt://www.scielo.br/scielo.php?script=sci_nlinks&ref=000112&pid=S1983...lng=pt

2. Lorenzi H. Brazilian trees: manual of identification and cultivation of native tree plants of Brazil. New Odessa: Plantarum. 1998; 2 (1): 352 p. Portuguese.
Available:https://www.bdpa.cnptia.embrapa.br/consulta/busca?b=ad...%22LORENZI...

3. Silva JHV, Silva EL, Filho JJ, Ribeiro MLG, Costa FGP. Residue of annatto seed (*Bixa orellana* L.) as dye of yolk, skin, beak and ovary of laying hens evaluated by two analytical methods. Ciência e Agrotecnologia. 2006; 30 (5): 988-994. Portuguese.
Available:https://www.scielo.br/pdf/cagro/v30n5/v30n5a24.pdf

4. Mathus MT, Lopes JC. Morphology of fruits and seedlings and seed germination of Erythrina variegata L. Revista Brasileira de Sementes. 2007; 29 (3): 8-15.
Available:https://www.redalyc.org/pdf/744/74413024004.pd

5. Gusmão E, Vieira FA, Júnior EMF. Fruit biometry and murici endocarps (*Byrsonima verbascifolia* Rich, Ex. A. Juss). Revista Cerne. 2006; 12 (1): 84-91. Portuguese.
Available:https://www.researchgate.net/profile/Fabio_Vieira2/publication...

6. Carpanezzi AA, Marques LCT. Germination of jutaí-açu (*Hymenaea courbaril* L.) and jutaí-mirim (*H. parvifolia* Huber) seeds scarified with commercial sulfuric acid. Circular Técnica 19. EMBRAPA-CPATU, Belém. 1981.15p. Portuguese.
Available:https://www.embrapa.br/busca-**de**-publicacoes/-/publicacao/376633/...

7. Fenner M. Seed ecology. London: Chapman & Hall, 1993:151 p.
Available:https://link.springer.com/chapter/10.1007/978-1-4615-1619-4_4

8. Baskin CC, Baskin JM. Seeds: ecology, biogeography and evolution of dormancy and Germination. London: Academic Press, 1998. 666p.
Available:https://link.springer.com/content/pdf/bbm:978-3-642-55974-7/1.pdf

9. Carvalho JEU, Nascimento WMO, Muller CH. Physical characteristics and seed germination of fruit species native to the Amazon. Research Bulletin 203. EMBRAPA-CPATU, Belém. 1998.18p. Portuguese.
Available:https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100...

10. Carmo Filho F, Oliveira OF. Mossoró: a municipality of the northeastern semi-arid region, climatic characterization and floristic aspect. Mossoró: ESAM, 1995. 62 p. (Coleção Mossoroense, Série B). Portuguese.
Available:http://www.scielo.br/scielo.php?script=sci_nlinks&pid=S0102...

11. R Version 3.1.1. Vienna, Áustria: Foundation for Statistical Computing, 2014. (Software). RAO KVK, Schwartz SA, Nair HK, Aalinkeel R, Mahajan S, Chawda R, Nair MPN. Plant derived products as a source of cellular growth inhibitory phytochemicals on PC-3M, DU-145 and LNCaP prostate cancer cell lines. Current Science. 2004; 87 (1): 1585-1588.

12. Ferreira DF. Basic statistics. Lavras: UFLA, 2005. 625 p. Portuguese.
Available:http://www.scholar.google.com.br/citations?user=M80iQtgAAAAJ&hl=pt-BR

13. Spiegel MR, Stephens LJ. Statistics. 4 ed. Porto Alegre: Bookman, 2009. 597 p. Portuguese.
Available:https://www.estantevirtual.com.br/livros/murray-r-spiegel/estatistica/3772227637

14. Oliveira MS. Introduction to statistics. Lavras, MG: Publisher of the Federal University of Lavras (UFLA), 2009. 329 p. Portuguese.
Available:https://www.editora.ufla.br/

15. Bussab WO, Morettin PA. Basic statistics. 6. ed. São Paulo: Saraiva, 2010. 540 p. Portuguese.
Available:https://www.saraiva.com.br/livros/editoras/saraiva

16. ZAR, J. H. Biostatistical analysis. 5. ed. New York: Prentice Hall, 2010. 944 p
Available:https://books.google.com.br/books?isbn=1119168996

17. Claudio LC, Stein CE. Descriptive statistics and probability theory. 2. ed. Blumenau: Edifurb, 2011. 213 p. Portuguese.
Available:https://books.google.com.br/books?isbn=853526356X

18. Cecon PR, Silva ARS, Nascimento M, Ferreira A. Statistical methods. Viçosa: Editora UFV, 2012. 229p. Portuguese.
Available:https://www.locus.ufv.br/bitstream/handle/123456789/9211/texto%20completo.pdf?...1...

19. Fonseca JS, Martins GA. Statistical course. 6. ed. 15. reimp. São Paulo: Atlas, 2012. 320 p.
Available:https://www.passeidireto.com/arquivo/46902730/livro-texto...iipdf-estatistica/13

20. Andrade DF, Ogliari PJ. Statistics for the agricultural and biological sciences with notions of experimentation. 3 ed. Florianópolis: Editora da UFSC, 2010. 470 p. Portuguese.
Available:https://www.passeidireto.com/.../livro-estatistica-basica-para-ciencias-agrarias-e-biolog...

21. Casella G, Berger RL. Statistical inference. São Paulo: Editora Cengage Learning. 2010. 612 p. Portuguese.
Available:https://www.cengage.com.br/ls/inferencia-estatistica-traducao-da-2a-edicao-norte-americana

22. Figueiredo F, Figueiredo A, Ramos A, Teles P. Descriptive statistics and probabilities - solved and proposed problems with R. applications Escolar Editora. Lisboa: Portugal, 2007. 420 p.
Available:https://www.wook.pt › Livros em Português › Ciências Exatas e Naturais › Matemática

23. Costa GGO. Course of inferential statistics and probabilities: theory and practice. São Paulo: Editora Atlas, 2012. 370 p. Portuguese.
Available:https://www.grupogen.com.br/curso-estatistica-inferencial-probabilidades

24. Gotelli NJ, Ellison AM. Principles of statistics in ecology. São Paulo: Artmed Editora. 2010. 532 p. Portuguese.
Available:https://www.passeidireto.com/.../principios-de-estatistica-em-ecologia---gotelli-nichola...