Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 75(2) 2021 93–103

Appropriate Statistical Distributions for Yield and Important Auxiliary Characters in Apple

Anju Sharma, P.K. Mahajan, Ashu Chandel, R.K. Gupta and Geeta Verma

Dr. Y.S. Parmar University of Horticulture and Forestry, Nauni, Solan

Received 12 May 2021; Revised 06 August 2021; Accepted 10 August 2021

SUMMARY

The frequency distribution analysis of yield and auxiliary characters utilizing different probability distributions *viz.,* Dagum 3P, Fatigue Life 3P, Gamma 3P, Generalized Gamma 4P, Inverse Gaussian 3P, Log-Logistic 3P, Lognormal 3P and Log-Pearson 3 have been carried out to sample data on different morphological growth characters for selecting the best distribution. The PCA was used to identify the important tree characters which were significantly contributing towards the yield. On analysing the values of different test statistics and based on the scores of goodness of fit tests for different growth variables under study the distribution Dagum 3P was found to be best fitted to spread and yield while Log-Pearson 3 and Inverse Gaussian 3P distributions were most valid fits for number of flowers and number of spurs, respectively. Hence, the results of this study can be used to know the trend and distributional pattern of auxiliary characters of apple crop towards enhancement of yield.

Keywords: Principal component analysis, Scree plot, Apple yield, Probability distributions, Goodness of fit tests.

1. INTRODUCTION

Apple being the commercial and dominating amongst all the fruit crops in Himachal Pradesh has completely transformed the socio-economic status of farmers in the High Hills zone of the State. It accounts for nearly 79% of the total fruit production (5,65,307 MT) and is grown in about 49% of the total area under fruit crops (2,30,852 hectares) of the State. Its area and production have tremendously increased from 88,669 ha and 49,129 MT in 1999-00 to 1,12,634 ha and 4,46,574 MT in 2017-18, respectively. Although there is a substantial increase in the production and productivity of apple in the state yet there are wide variations in yield on year-to-year basis due to variable environmental situations and phenotypical developmental stages of the plant. Therefore, it is very pertinent to equip the famers with most probable yield estimates of their trees so as to plan for best marketing strategies. The form of distribution depends on crop, average yield level and many local or regional climatic conditions (Day, 1965; Gallagher, 1986). The traditional approach to forecasting crop yields is to assume normal weather and present crop yields as linear extensions of past trends. Predicting the future yield and its distribution based on the data of previous years has proved to be difficult and the results are unreliable.

Distribution fitting is the method of selecting appropriate distributions among the available distributions to be fitted on a given set of data to predict the probability or to forecast the frequency of occurrence of the magnitude of the phenomenon in a certain interval (Jha *et al.*, 2012). In practice, probability distributions are applied in many fields such as actuarial science and insurance, risk analysis, investment, market research, business and economic research, customer support, mining, reliability engineering, chemical engineering, hydrology, image processing, physics, medicine, sociology, demography etc. Knowledge of the expected behaviour of a phenomenon is very useful in a large number of problems in practical situations. Such phenomenon facilitates decision-makers in making predictions on the basis of theoretical information. Frequency analysis usually involves the fitting of a theoretical frequency

Corresponding author: Anju Sharma
E-mail address: anjusharma_uhf@yahoo.com

distribution using a selected fitting method. Various probability distributions are currently used to predict expected yield and a number of probability models have been developed to depict the distribution. The selection of choice of an appropriate distribution model mainly depends on the available yield and auxiliary data at a particular site. To find a suitable distribution that suits the data well, it is necessary to evaluate the available distribution models. The goodness of fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit summarize the difference between observed and the expected values under the given model. Such measures can be used in statistical hypothesis testing, e.g. to test whether two samples are drawn from identical distributions or whether outcome frequencies follow a specified distribution.

Crop-yield probability distributions have been extensively studied, mainly in regards to premium rate making for crop insurance programs (Yeh and Sun, 1980; Ramirez *et al.*, 2003) and for use in farm decision support systems. The results of these studies can be exploited by farm decision makers to enhance the outcomes of their own yield related decisions. Farm decision makers need to recognize that crop-yield distributions may be non-normal which will require non-normal methodology to generate useful statistics for decision making. The irregularly changing weather causes variability in crop yields. Gorski (2009) suggested log normal distribution with an inverted abscissa for obtaining a fairly good approximation of the probability distribution for actual yields. Various probability distributions have also been used to predict expected rainfall in different return periods through probability and frequency analysis of rainfall data (Amin *et al.*, 2016). The Gamma distribution was used to model the distribution of the quarterly rainfall amount and Kolmogorov – Smirnov, one sample test was used to evaluate the model fit (Dikko *et al.*, 2013).

In the present study frequency analyses of morphological data on growth characters have been performed. The objective of the study is to perform probability analysis to gauge the trend and distributional pattern of important auxiliary growth characters significantly contributing towards the enhancement of apple yield.

2. MATERIALS AND METHODS

Probability distributions are basic concepts in statistics which link results of statistical experiments and their probabilities of occurrence. The use of advanced distributions for data analysis increases the validity of models, which in turn, leads to better decisions. Out of many available distributions, eight theoretical (continuous) distributions have been used in the present study which included seven non-negative distributions viz., Dagum 3P, Fatigue Life 3P, Gamma 3P, Generalized Gamma 4P, Inverse Gaussian 3P, Log-Logistic 3P, Lognormal 3P and an advanced continuous distribution viz., Log-Pearson 3. These distributions have been tried on data pertaining growth characteristics which contributed significantly towards the yield of tree. For the purpose, data were recorded on different tree growth (morphological) characteristics like age (years), girth (m), height (m), spread (m), volume (m^3), number of main branches, number of secondary branches, number of spurs per tertiary branch, length of spurs (cm), number of flowers per tertiary branch, number of fruits per tertiary branch, fruit weight (g) and yield of tree (kg) from a random sample of 300 apple trees from a randomly selected orchards at Shimla district of Himachal Pradesh. The comprehensive analysis of data was done with the help of graphs of probability density function (PDF) and cumulative distribution function (CDF) of these distributions. Table 1 shows the equations of PDFs and CDFs of different theoretical distributions.

The goodness of fit test measures the compatibility of a random sample with a theoretical probability distribution function. It gives the distance and critical values, measured between the data and the distribution being tested. The critical value is then compared to some threshold value. The goodness of fit reports includes the test statistics and the critical values calculated for various significance levels (= 0.2, 0.1, 0.05, 0.02, 0.01). The P-value can be helpful specifically when the null hypothesis is rejected at all selected significance levels, where the P-value is criteria of uniformity between the results actually obtained in the experiment and the random chance explanation for those results. It is required to know at which level it could be accepted (Jha *et al.*, 2011).

2.1 Hypothesis Testing

H_0 : the data follow the specified distribution;

H_1 : the data do not follow the specified distribution.

The hypothesis regarding the distributional form is rejected at the chosen significance level (α) if the test statistic is greater than the critical value obtained from a table.

2.2 P-Value

The P-value, in contrast to fixed α values, is calculated based on the test statistic, and denotes the threshold value of the significance level in the sense that the null hypothesis (H_0) will be accepted for all values of α less than the P-value.

Therefore, in assessing whether a given distribution is suited to a data-set, the following tests and their underlying measures of fit have been used:

i) Kolmogorov–Smirnov test to decide if a sample comes from a hypothesized continuous distribution. It is based on the empirical cumulative distribution function (ECDF). Assume that we have a random sample $x_1, x_2, \ldots, x_n$ from some distribution with CDF $F(x)$. The empirical CDF is denoted by

$$F_n(x) = \frac{1}{n}[\text{Number of observations} \le x]$$

The Kolmogorov–Smirnov statistic (D) is based on the largest vertical difference between the theoretical and the empirical cumulative distribution function:

$$D = \underset{1 \le i \le n}{\max}\left(F(x_i) - \frac{i-1}{n}, \frac{i}{n} - F(x_i) \right)$$

ii) Anderson-Darling test to compare the fit of an observed cumulative distribution function to an expected cumulative distribution function to an expected cumulative distribution function. This test gives more weight to the tails than the Kolmogorov-Smirnov test.

The Anderson-Darling statistic (A^2) is defined as

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1).\left[\ln F(X_1) + \ln(1 - F(X_{n-i+1}))\right]$$

iii) Chi Squared test to determine if a sample comes from a population with a specific distribution.

$$\div^2 = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i}$$

3. RESULTS AND DISCUSSION

In the first stage of analysis, descriptive statistics giving primary and important information (Table 2) and correlation between the different tree characteristics (Table 3) were calculated. Since the standard deviation showed substantial differences, data were standardized. Further all variables were significantly correlated with yield. Hence, a statistical multivariate technique called principal component analysis (PCA) that uses orthogonal transformation to convert several correlated observed variables into a smaller number of linearly uncorrelated variables known as principal components was employed using SPSS windows version 22 software. A large number of variables are often measured by plant breeders, some of which may not be of sufficient discriminatory power for germplasm evaluation, characterization and management (Maji and Shaibu, 2012). Sharma *et al.* (2018) and Verma *et al.* (2018) highlighted the usefulness of multivariate techniques for determining the relative contribution of morphological characters responsible in increasing the apple and kinnow yield respectively. Prior to PCA, the data was tested for suitability using Kaiser-Meyer-Olklin (KMO) measure of sampling adequacy (0.715) and Bartlett Test of Sphericity ($p \text{ value} = 0$). The extracted components were rotated using orthogonal varimax method so that a smaller number of highly correlated variables might be put under each factor and achieve significant components. Out of 12 correlated variables included in PCA, 3 principal components with eigen values greater than 1 (i.e. 4.591, 3.177 and 1.065 respectively)were retained for further analysis (Table 4) and explained about 38.26%, 26.48% and 8.88% of variance respectively of the total variation and accounted for 73.61% of total variation of original variables. As evident from the scree plot (Fig. 1), only the first three principal components explained maximum variability. From the fourth component onwards, the line is almost flat, meaning each successive component is accounting for smaller and smaller amounts of the total variance. Maximum extraction (Table 4) was shown by the variable number of flowers (0.967) followed by height (0.929), spread (0.926) and number of spurs respectively. A perusal of Table 5 showing loadings of correlation matrix and correlation of principal components with original variables using Varimax with Kaiser Normalization rotation method showed maximum weight of spread (0.404), number of flowers (0.517) and number of spurs (0.429), in first, second and the third principal components, respectively. A better view of the data is obtained by "rotating them". Thus, by discarding the loads lesser than 0.4, the rotation component matrix suggested that first principal component consisting of spread (0.957), height (0.956), volume (0.931), length of spurs (0.623), age (0.545), and fruit weight (0.496) may be interpreted as **plant vigour**. The second principal component was dominated by number of flowers (0.955), number of spurs (0.924), and number of fruits (0.715) and hence may be interpreted as **blooming and fruiting characteristics**. The third principal component representing **vegetative characteristics** consisted of number of secondary branches (0.828), girth (0.706), and number of main branches (0.685). Hence, important morphological characters contributing significantly towards the yield of tree, depicted by principal component analysis were viz., spread, height, volume, number of flowers and number of spurs. Thus, these variables along with yield variable have been selected for fitting of the probability distributions to sample data using the distribution fitting software Easy Fit.

The maximum likelihood method has been used to estimate the parameters. Parameters of various continuous probability distributions viz., Dagum 3P, Fatigue Life 3P, Gamma 3P, Generalized Gamma 4P, Inverse Gaussian 3P, Log-Logistic 3P, Lognormal 3P and Log-Pearson 3 have been tabulated in Table 6. These statistical parameters (scale, shape and location) can be used to calculate the estimated values of different growth characteristics contributing towards the apple yield using different probability distributions at certain level of probability. To find a suitable distribution that will provide accurate estimates of yield and other tree characters, it is necessary to evaluate the available distributions.

The eight probability distributions were subjected to three goodness of fit tests viz., Kolmogorov Smirnov test, Anderson Darling test and Chi-Squared test in respect of different morphological tree growth characteristics to select one or more best fitting probability distributions. A standard procedure was followed for application of goodness of fit tests (Chowdhury and Stedinger, 1991). The distribution with the lowest statistical value is the best-fit. The probability distributions were ranked from one (best fit) to eight (least- fit) for these goodness of fit tests. The statistic value along with the rank of these distributions pertaining to different variables have been presented in Table 7.

Selection of the best-fit probability distribution is based on the total score from all the goodness of fit tests. The results of goodness of fit tests for each probability distribution used in this study have been shown in Tables 8. Based on the results of the goodness of fit tests the best-fit probability distributions were identified. On the basis of ranks the Inverse Gaussian 3P distribution for spread with rank total = 6, and height with rank total = 4, Dagum 3P for volume with rank total = 3 and yield with rank total = 3 and Log-Pearson distribution was a best fit for number of flowers and number of spurs with rank total equal to 5 and 6 respectively (Table 8). However, after analysing the goodness of fit deeply by looking at the critical value and significance, Dagum 3P was found to be best suited for spread with scale parameter $\beta = 0.313$ and shape parameters $k = 22.135$ and $\alpha = 1.290$. Similarly, Dagum 3P distribution has been observed to be optimal analysis of yield with scale parameter $\beta = 267.970$ and shape parameters $k = 0.126$ and $\alpha = 5.202$. The Log-Pearson 3 and Inverse Gaussian 3P distributions were the best fits for number of flowers with parameters $\alpha = 23.338$, $\beta = -0.150$, $\gamma = 6.573$ and number of spurs with parameters $\lambda = 11.379$, $\mu = 7.178$, $\gamma = -0.436$, respectively. The goodness of fit test results of finally identified best fit distributions and graphs of their PDFs and CDFs distributions are given in Table 9 and Fig. 2, respectively. The Log-Pearson 3 distribution for number of flowers and the Inverse Gaussian 3P distribution for number of spurs are accepted by all three tests and at all level of significances. Whereas, Dagum 3P distribution for spread and yield is accepted by Kolmogorov-Smirnov and Anderson-Darling tests. Mathematical expression for the calculation of respective variables can be obtained. Hence, either from graphs or from equations, either x_i or p_i could be easily predicted with highest degree of accuracy possible with probability analysis. The number of vertical bars are based on the total number of observations. The equation $k = 1 + \log 2N$, was used to find the number of bins (histogram), where N is the total sample size. The height of each histogram bar indicates how many data points fall into that class.

4. CONCLUSION

The study has established an evidence about the distribution of yield in particular, and important morphological characters contributing significantly towards apple yield in general. Estimation of yield and other tree growth characters in Himachal Pradesh obtained from principal component analysis were carried out by Dagum 3P, Fatigue Life 3P, Gamma 3P, Generalized Gamma 4P, Inverse Gaussian 3P, Log-Logistic 3P, Lognormal 3P and Log-Pearson 3 probability distributions on the data recorded on 300 apple trees. On analysing the values of different test statistics and based on the scores of goodness of fit tests for different growth variables under study the distribution Dagum 3P has given a close fit for spread and yield while Log-Pearson 3 and Inverse Gaussian 3P were the best fits for number of flowers and number of spurs, respectively. There is no single distribution which is best fit for frequency analysis of all tree characters. Thus, it can be concluded that fitting of appropriate statistical distribution(s), for yield and important auxiliary character(s) may provide information regarding the trend and distributional pattern of apple in geographical and agro-climatically heterogeneous state of Himachal Pradesh. The expected values of different estimates calculated using the best fit probability distributions might be used by decision-makers for making good predictions. Crop-yield distribution information, in particular, can be utilised for better on-farm decision making and can be a future avenue for research.

ACKNOWLEDGEMENTS

We would like to express our sincere thanks and gratitude to the learned reviewers for their valuable suggestions that helped us a lot in improving the quality of this manuscript.

REFERENCES

Amin, M.T., Rizwan M. and Alazba A.A. 2016. A best-fit probability distribution for the estimation of rainfall in northern regions of Pakistan. *Open Life Science,* **11**, 432-440

Chowdhury J.U., Stedinger J.R. and Lu L.H. 1991. Goodness of fit tests for regional generalized extreme value flood distributions. *Water Resources Research,* **27(7)**, 1765-1777.

Day R.H. 1965. Probability distribution of field crop yields. *Journal of Farm Economics,* **47**, 713-741.

Dikko H.G, David I.J. and Bakari H.R. 2013. Modeling the distribution of rainfall intensity using quarterly data. *IOSR - Journal of Mathematics,* **9(1)**, 11-16.

Gallagher P. 1986. US corn yield capacity and probability: estimation and forecasting with nonsymmetric disturbances. *North Central Journal of Agricultural Economics,* **8**, 109-122.

Gorski T. 2009. A probability distribution for crop yield in Poland. *Geographia polonica,* **82(1)**, 61-67.

Jha S.K., Dwivedi A. and Tiwari A . 2011. Necessity of goodness of fit tests in research and development. *International J. Comp. Sci. and Technology,* **2**, 135-141.

Jha S.K., Singh S. and Tiwari A . 2012. Identification of appropriate distribution in reliability analysis. *International Journal of Engineering Research and Technology,* **1(6)**, 1-7.

Maji A.T. and Shaibu A.A. 2012. Application of principal component analysis for rice germplasm characterization and evaluation. *Journal of Plant Breeding and Crop Science,* **4(6)**, 87-93.

Ramirez O.A., Misra S. and Field J. 2003. Crop-yield distributions revisited. *American Journal of Agricultural Economics,* **85**,108-120.

Sharma A., Mahajan P.K., Malhotra R., Belwal O.K. and Sharma S.K. 2018. Application of Multivariate Techniques in Horticulture Research-a Case Study of Apple Crop. *International Journal of Bio-resource and Stress Management,* **9(5)**, 556-561

Verma G., Sharma A. and Mahajan P.K. 2018. Study on yield and some morphological characters for optimizing kinnow yield through multivariate statistical techniques. *International Journal of Farm Sciences,* **8(2)**, 162-165.

Yeh M.H. and Sun S. 1980. An evaluation of the actuarial structure of Manitoba Crop Insurance program and its implications. *Can. J. Ag. Econ.* **28**:54-70.

Table 1. Probability density functions and cumulative distribution functions

Distribution & Parameters	Domain	Probability Density Function	Cumulative Distribution Function
Dagum 3P k- continuous shape parameter (k > 0) α - continuous shape parameter $(\alpha > 0)$ β - continuous scale parameter $(\beta > 0)$ γ - continuous location parameter	$\gamma \le x < +\infty$	$f(x) = \dfrac{\alpha k \left(\dfrac{x}{\beta}\right)^{\alpha k - 1}}{\beta \left(1 + \left(\dfrac{x}{\beta}\right)^{\alpha}\right)^{k+1}}$	$F(x) = \left(1 + \left(\dfrac{x}{\beta}\right)^{-\alpha}\right)^{-k}$
Fatigue Life(Birnbaum-Saunders) 3P α - continuous shape parameter $(\alpha > 0)$ β - continuous scale parameter $(\beta > 0)$ γ - continuous location parameter	$\gamma < x < +\infty \beta F(x)$ $= \varnothing\left(\dfrac{}{\alpha}\left(\sqrt{\dfrac{\gamma}{\beta}} - \sqrt{\dfrac{\beta}{\gamma}}\right)\right) f(x)$ $\dfrac{\sqrt{(x-\gamma)/\beta} + \sqrt{\beta/(x-\gamma)}}{\alpha(\quad \gamma)} .\gamma\alpha$ $\varnothing\left(\dfrac{}{\alpha}\left(\sqrt{\dfrac{\gamma}{\beta}} - \sqrt{\dfrac{\beta}{\gamma}}\right)\right)$	$f(x) = \dfrac{\sqrt{(x-\gamma)/\beta} + \sqrt{\beta/(x-\gamma)}}{2\alpha(x-\gamma)} .$ $\varnothing\left(\dfrac{1}{\alpha}\left(\sqrt{\dfrac{x-\gamma}{\beta}} - \sqrt{\dfrac{\beta}{x-\gamma}}\right)\right)$	$F(x) = \varnothing\left(\dfrac{1}{\alpha}\left(\sqrt{\dfrac{x-\gamma}{\beta}} - \sqrt{\dfrac{\beta}{x-\gamma}}\right)\right)$
Gamma 3P α - continuous shape parameter $(\alpha > 0)$ β - continuous scale parameter $(\beta > 0)$ γ - continuous location parameter	$\gamma \le x < +\infty$	$f(x) = \dfrac{(x-\gamma)^{\alpha-1}}{\beta^{\alpha}\tau(\alpha)} \exp(-(x-\gamma)/\beta)$	$F(x) = \dfrac{\tau_{(x-\gamma)/\beta\ (\alpha)}}{\tau(\alpha)}$
Gen. Gamma 4P k- continuous shape parameter (k > 0) α - continuous shape parameter $(\alpha > 0)$ β - continuous scale parameter $(\beta > 0)$ γ - continuous location parameter	$\gamma \le x < +\infty$	$f(x) = \dfrac{k(x-\gamma)^{k\alpha-1}}{\beta^{k\alpha}\tau(\alpha)}$ $\exp(-((x-\gamma)/\beta)^{k})$	$F(x) = \dfrac{\tau_{(x-\gamma)/\beta^{k}\ (\alpha)}}{\tau(\alpha)}$
Inv. Gaussian 3P λ - continuous parameter $(\lambda > 0)$ μ - continuous parameter $(\mu > 0)$ γ - continuous location parameter	$\gamma < x < +\infty$	$f(x) = \sqrt{\dfrac{\lambda}{2\pi(x-\gamma)^{3}}}$ $\exp\left(-\dfrac{\lambda(x-\gamma-\mu)^{2}}{2\mu^{2}(x-\gamma)}\right)$	$F(x) = \phi\left(\sqrt{\dfrac{\lambda}{x-\gamma}}\left(\dfrac{x-\gamma}{\mu} - 1\right)\right) +$ $\phi\left(-\sqrt{\dfrac{\lambda}{x-\gamma}}\left(\dfrac{x-\gamma}{\mu} + 1\right)\right) \exp(2\lambda/\mu)$
Log-Logistic 3P α - continuous shape parameter $(\alpha > 0)$ β - continuous scale parameter $(\beta > 0)$ γ - continuous location parameter	$\gamma \le x < +\infty$	$f(x) = \dfrac{\alpha}{\beta}\left(\dfrac{x-\gamma}{\beta}\right)^{\alpha-1}\left(1 + \left(\dfrac{x-\gamma}{\beta}\right)^{\alpha}\right)^{-2}$	$F(x) = \left(1 + \left(\dfrac{\beta}{x-\gamma}\right)^{\alpha}\right)^{-1}$
Lognormal 3P σ - continuous parameter $(\sigma > 0)$ μ - continuous parameter - continuous location parameter	$\gamma < x < +\infty$	$f(x) = \dfrac{\exp\left(-\dfrac{1}{2}\left(\dfrac{\ln(x-\gamma) - \mu}{\sigma}\right)^{2}\right)}{(x-\gamma)\sigma\sqrt{2\pi}}$	$F(x) = \phi\left(\dfrac{\ln(x-\gamma) - \mu}{\sigma}\right)$
Log- Pearson 3 α - continuous parameter $(\alpha > 0)$ β - continuous parameter $(\beta \ne 0)$ γ - continuous parameter	$0 < x \le e^{\gamma}, \quad \beta < 0$ $e^{\gamma} \le x \langle +\infty, \beta \rangle 0$	$f(x) = \dfrac{1}{x\|\beta\|\tau(\alpha)}\left(\dfrac{\ln(x) - \gamma}{\beta}\right)^{\alpha-1}$ $\exp\left(-\dfrac{\ln(x) - \gamma}{\beta}\right)$	$F(x) = \dfrac{\tau_{(\ln(x)-\gamma)/\beta}(\alpha)}{\tau(\alpha)}$

Table 2. Descriptive statistics of different tree characteristics

Tree Characters	Age (Years)	Girth (m)	Height (m)	Spread (m)	Volume (m^3)	Number of main branches	Number of secondary branches	Number of spurs per tertiary branch	Length of spurs (cm)	Number of flowers per tertiary branch	Number of fruits per tertiary branch	Fruit weight (kg)	Yield (kg)
Mean	20.350	0.560	10.163	7.932	1291.810	4.01	14.997	6.742	1.443	27.313	12.118	149.170	116.510
SD	9.487	0.264	8.361	7.298	2448.410	1.41	6.128	5.212	0.141	18.018	9.673	28.068	103.410
CV	0.466	0.473	0.822	0.920	1.895	0.351	0.409	0.773	0.098	0.660	0.798	0.188	0.888
SE	0.548	0.015	0.483	0.421	141.360	0.081	0.354	0.301	0.008	1.040	0.559	1.620	5.970
Skewness	0.360	0.585	1.143	1.184	2.137	0.553	0.578	1.515	0.719	1.025	1.354	-0.607	1.059
Minimum value	4	0.13	1.9	0.65	0.49	1	4	1	1.18	2.75	1.5	48.33	1.5
Maximum value	50	1.22	35	30.95	13257	10	30	30	1.825	92.5	42	290	500

Table 3. Correlation between different tree characters

Tree Characters	Age (Years)	Girth (m)	Height (m)	Spread (m)	Volume (m^3)	Number of main branches	Number of secondary branches	Number of spurs per tertiary branch	Length of spurs (cm)	Number of flowers per tertiary branch	Number of fruits per tertiary branch	Fruit weight (kg)	Yield (kg)
Age (Years)	1.000												
Girth (m)	0.727**	1.000											
Height (m)	0.487**	0.463**	1.000										
Spread (m)	0.459**	0.507**	0.931**	1.000									
Volume (m^3)	0.406**	0.407**	0.888**	0.924**	1.000								
Number of main branches	0.261**	0.333**	0.000	0.078	0.059	1.000							
Number of secondary branches	0.407**	0.752**	0.154**	0.211**	0.142*	0.328**	1.000						
Number of spurs per tertiary branch	0.184**	0.092	-0.357**	-0.293**	-0.249**	0.173**	0.085	1.000					
Length of spurs (cm)	0.308**	0.507**	0.550**	0.595**	0.587**	0.231**	0.331**	-0.246**	1.000				
Number of flowers per tertiary branch	0.239**	0.304**	-0.247**	-0.161**	-0.175**	0.234**	0.315**	0.924**	-0.063	1.000			
Number of fruits per tertiary branch	0.312**	0.578**	0.055	0.151**	0.043	0.256**	0.569**	0.505**	0.238**	0.764**	1.000		
Fruit weight (kg)	0.244**	0.236**	0.423**	0.418**	0.337**	0.100	0.151**	-0.158**	0.339**	-0.069	0.129*	1.000	
Yield (kg)	0.641**	0.651**	0.604**	0.663**	0.662**	0.285**	0.375**	0.115*	0.597**	0.207**	0.336**	0.291**	1.000

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 4. Eigen values, variance and communalities of different tree characteristics considered for PCA

Tree Characters	Communalities: Extraction	Initial Eigen values			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
		Total (λ)	% of Variance	Cumulative %	Total (λ)	% of Variance	Cumulative %	Total (λ)	% of Variance	Cumulative %
Age (Years)	0.554	4.591	38.257	38.257	4.591	38.257	38.257	4.020	33.503	33.503
Girth (m)	0.842	3.177	26.477	64.734	3.177	26.477	64.734	2.555	21.292	54.795
Height (m)	0.929	1.065	8.879	73.613	1.065	8.879	73.613	2.258	18.817	73.613
Spread (m)	0.926	0.823	6.855	80.467						
Volume (m3)	0.872	0.775	6.456	86.923						
Number of main branches	0.474	0.684	5.701	92.624						
Number of secondary branches	0.749	0.402	3.346	95.971						
Number of spurs per tertiary branch	0.914	0.241	2.011	97.981						
Length of spurs(cm)	0.584	0.108	0.899	98.880						
Number of flowers per tertiary branch	0.967	0.075	0.622	99.502						
Number of fruits per tertiary branch	0.752	0.046	0.382	99.884						
Fruit weight (kg)	0.270	0.014	0.116	100.000						

Table 5. Loadings of correlation matrix and rotated component matrix

Tree Characters / Variables	Eigen vectors			Rotated component matrix (Varimax with Kaiser Normalization)		
	PC1	PC2	PC3	PC1	PC2	PC3
Age	0.325	0.142	0.075	**0.545**	0.334	0.382
Girth	0.380	0.207	-0.203	0.511	0.285	**0.707**
Height	0.386	-0.242	0.237	**0.956**	-0.123	0.013
Spread	**0.404**	-0.197	0.227	**0.958**	-0.058	0.074
Volume	0.373	-0.218	0.279	**0.931**	-0.066	-0.015
No. of main branches	0.140	0.205	-0.485	-0.034	0.061	**0.685**
No. of secondary branches	0.259	0.263	-0.455	0.163	0.194	**0.828**
No. of spurs	-0.052	0.471	**0.429**	-0.244	**0.924**	-0.020
Length of spurs	0.339	-0.080	-0.179	**0.628**	-0.134	0.415
No. of flowers	0.041	**0.517**	0.321	-0.134	**0.955**	0.195
No. of fruits	0.200	0.421	0.078	0.139	**0.715**	0.471
Fruit weight	0.230	-0.091	0.027	**0.497**	-0.059	0.140

Table 6. Parameters of probability distributions at different locations

Distribution	Spread	Height	Volume	No. of flowers	No. of spurs	Yield
Dagum 3P	k=22.135 α=1.290 β=0.313	k=176.570 α=1.661 β=0.139	k=54.733 α=0.471 β=0.007	k=0.557 α=3.094 β=31.339	k=0.856 α=2.305 β=5.724	k=0.126 α=5.202 β=267.970
Fatigue Life 3P	α=1.103 β=4.704 γ=0.384	α=1.054 β=5.660 γ=1.393	α=3.983 β=166.6 γ=-0.472	α=0.614 β=26.032 γ=-3.638	α=0.805 β=5.190 γ=-0.139	α=1.360 β=60.164 γ=-3.991
Gamma 3P	α=1.103 β=4.704 γ=0.384	α=0.961 β=8.780 γ=1.900	α=0.272 β=4395.0 γ=0.490	α=1.744 β=14.341 γ=2.294	α=0.705 β=7.325 γ=1.000	α=0.806 β=138.810 γ=1.500

Gen. Gamma 4P	k=0.374 α=6.206 β=0.0405 γ=0.611	k=0.402 α=5.352 β=0.095 γ=1.809	k=0.665 α=0.425 β=2611.9 γ=0.490	k=1.547 α=0.821 β=31.761 γ=2.673	k=1.253 α=0.678 β=7.846 γ=1.000	k=2.667 α=0.216 β=307.230 γ=1.500
Inv. Gaussian 3P	λ=5.279 μ=7.648 γ=0.284	λ=6.961 μ=8.913 γ=1.250	λ=23.015 μ=1294.000 γ=-2.228	λ=88.971 μ=32.227 γ=-4.914	λ=11.379 μ=7.178 γ=-0.436	λ=129.720 μ=136.400 γ=-19.892
Log-Logistic 3P	α=1.522 β=4.241 γ=0.627	α=1.553 β=4.807 γ=1.822	α=0.637 β=86.951 γ=0.490	α=2.530 β=23.427 γ=-0.976	α=1.948 β=4.663 γ=0.393	α=1.132 β=68.063 γ=1.374
Lognormal 3P	σ=1.062 μ=1.487 γ=0.533	σ=1.030 μ=1.652 γ=1.621	σ=2.619 μ=4.626 γ=0.470	σ=0.592 μ=3.265 γ=-3.656	σ=0.778 μ=1.632 γ=-0.045	σ=1.188 μ=4.268 γ=-4.456
Log- Pearson 3	α=84.446 β=0.100 γ=-6.771	α=19.078 β=0.174 γ=-1.313	α=40.860 β=0.399 γ=-11.644	α=23.338 β=-0.150 γ=6.573	α=117.480 β=0.07276 γ=10.168	α=6.136 β=0.572 γ=7.612

Table 7. Results of Goodness of fit tests

Distribution	Spread			Height			Volume			No. of flowers			No. of spurs			Yield		
	KS	AD	CS	KS	AD	CS	KS	AD	CS	KS	AD	CS	KS	AD	CS	KS	AD	CS
Dagum 3P	0.085 (1)	3.798 (3)	51.820 (3)	0.361 (8)	77.215 (8)	225.240 (8)	0.094 (1)	4.211 (1)	41.949 (1)	0.051 (6)	1.136 (7)	8.448 (5)	0.056 (6)	1.329 (5)	28.276 (6)	0.083 (1)	2.313 (1)	32.357 (1)
Fatigue Life 3P	0.094 (3)	3.274 (1)	56.770 (5)	0.117 (4)	6.037 (4)	95.562 (4)	0.135 (6)	6.466 (5)	60.192 (2)	0.048 (3)	0.825 (4)	8.168 (3)	0.045 (1)	0.929 (2)	13.009 (4)	0.211 (8)	12.157 (8)	101.880 (6)
Gamma 3P	0.112 (8)	6.033 (8)	78.301 (8)	0.141 (7)	9.241 (7)	129.840 (7)	0.173 (8)	19.039 (8)	**N/A**	0.046 (2)	0.631 (2)	17.327 (8)	0.175 (8)	75.975 (8)	**N/A**	0.159 (5)	10.272 (6)	**N/A**
Gen. Gamma 4P	0.104 (6)	4.314 (6)	73.252 (7)	0.121 (5)	6.745 (5)	107.470 (6)	0.133 (5)	15.282 (6)	**N/A**	0.048 (4)	0.598 (1)	17.039 (7)	0.103 (7)	66.700 (7)	**N/A**	0.116 (2)	6.461 (3)	**N/A**
Inv. Gaussian 3P	0.090 (2)	3.480 (2)	50.209 (2)	0.105 (2)	5.646 (1)	75.976 (1)	0.153 (7)	15.817 (7)	81.968 (5)	0.050 (5)	0.861 (5)	8.580 (6)	0.050 (3)	0.972 (3)	8.650 (1)	0.175 (6)	9.204 (4)	82.335 (3)
Log-Logistic 3P	0.098 (4)	4.045 (5)	56.580 (4)	0.100 (1)	5.729 (2)	81.198 (3)	0.107 (2)	6.107 (4)	78.207 (4)	0.054 (8)	1.357 (8)	8.243 (4)	0.051 (4)	1.543 (6)	13.575 (5)	0.157 (4)	10.783 (7)	96.736 (5)
Lognormal 3P	0.099 (5)	3.813 (4)	46.409 (1)	0.110 (3)	5.986 (3)	76.308 (2)	0.122 (4)	5.986 (3)	82.156 (6)	0.051 (7)	0.934 (6)	7.384 (2)	0.051 (5)	1.068 (4)	8.686 (2)	0.179 (7)	9.865 (5)	87.855 (4)
Log- Pearson 3	0.111 (7)	4.718 (7)	60.590 (6)	0.131 (6)	7.786 (6)	96.491 (5)	0.120 (3)	5.574 (2)	74.114 (3)	0.043 (1)	0.669 (3)	3.889 (1)	0.047 (2)	0.876 (1)	11.217 (3)	0.148 (3)	5.942 (2)	61.359 (2)

Table 8. Identification of best –fit distribution for frequency analysis of different tree characters

Distribution \ Tree characters	Spread	Height	Volume	No. of flowers	No. of spurs	Yield
Dagum 3P	7	24	3	18	17	3
Fatigue Life 3P	9	12	13	10	7	22
Gamma 3P	24	21	16	12	16	11
Gen. Gamma 4P	19	16	11	12	14	5
Inv. Gaussian 3P	6	4	19	16	7	13
Log-Logistic 3P	13	6	10	20	15	16
Lognormal 3P	10	8	13	15	11	16
Log- Pearson 3	20	17	8	5	6	7
Best Fit(s)	Inv. Gaussian 3P Dagum 3P	Inv. Gaussian 3P Log-Logistic 3P	Dagum 3P Log-Pearson 3	Log- Pearson 3 Fatigue Life 3P	Log- Pearson 3 Fatigue Life 3P Inv. Gaussian 3P	Dagum 3P Gen. Gamma 4P

Table 9. Goodness of fit detail of best-fit distributions

Tree characters	Best fit distribution		Kolmogorov-Smirnov (D)					Anderson-Darling (A^2)					Chi-Squared (χ^2 , d.f.=8)				
		α	0.2	0.1	0.05	0.02	0.01	0.2	0.1	0.05	0.02	0.01	0.2	0.1	0.05	0.02	0.01
		Critical Value	0.061	0.071	0.078	0.087	0.094	1.375	1.929	2.502	3.289	3.907	11.030	13.362	15.507	18.168	20.090
Spread4	Dagum 3P	Statistic	0.085					3.798					51.82				
		P-Value	0.024					-					1.8234E-8				
		Reject?	Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Number of flowers10	Log-Pearson 3	Statistic	0.043					0.669					3.890				
		P-Value	0.616					-					0.867				
		Reject?	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Number of spurs8	Inverse Gaussian 3P	Statistic	0.049					0.972					8.650				
		P-Value	0.432					-					0.373				
		Reject?	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
Yield13	Dagum 3P	Statistic	0.083					2.313					32.357				
		P-Value	0.029					-					0.00008				
		Reject?	Yes	Yes	Yes	No	No	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes

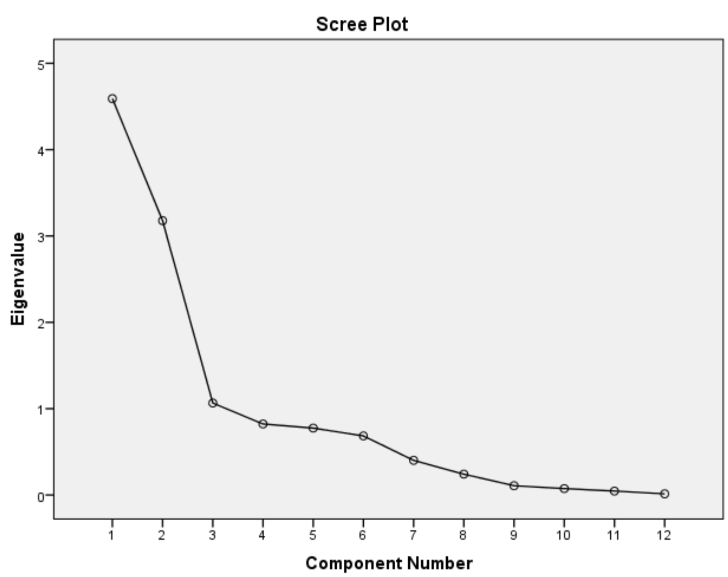
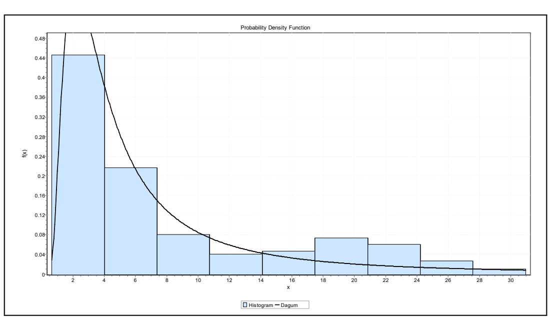
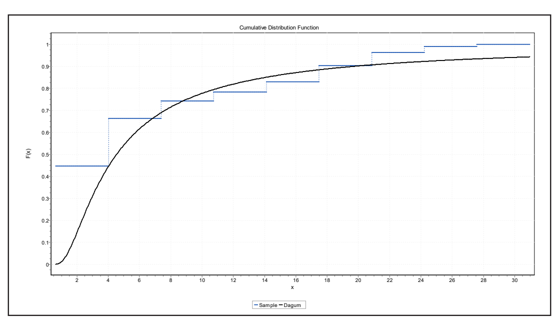


Fig. 1. Scree plot

PDF **CDF**

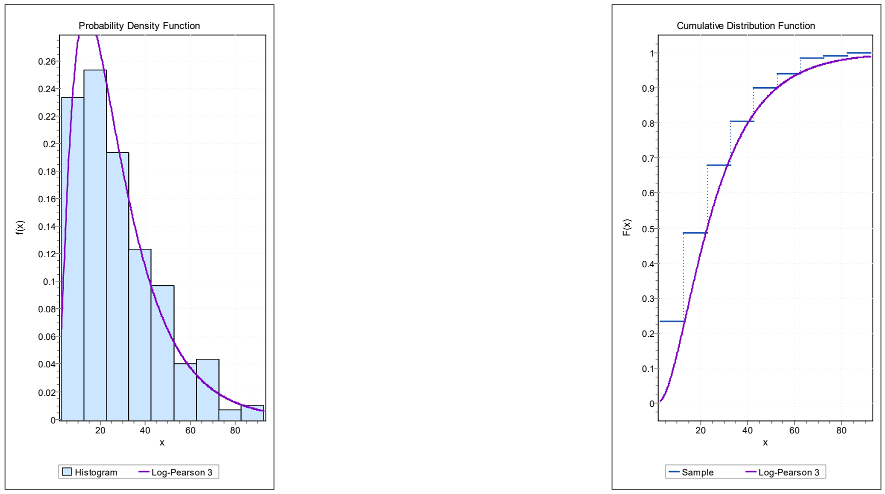
Spread





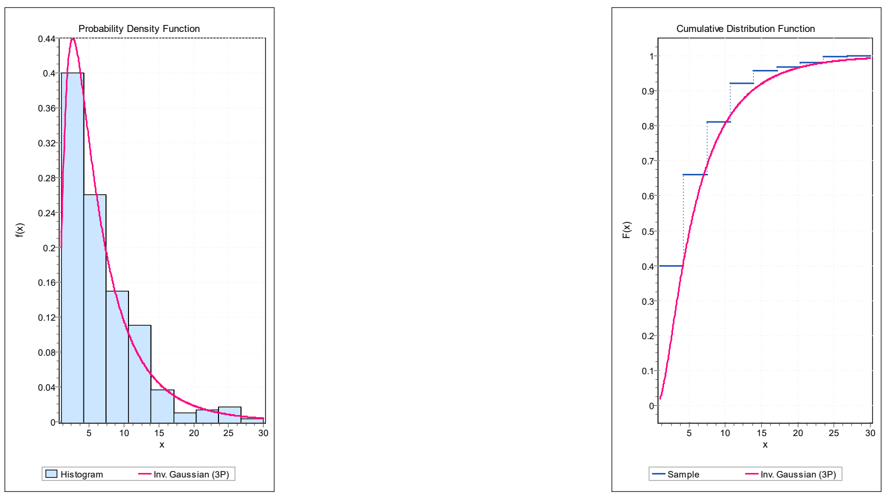
Dagum 3P

Number of flowers



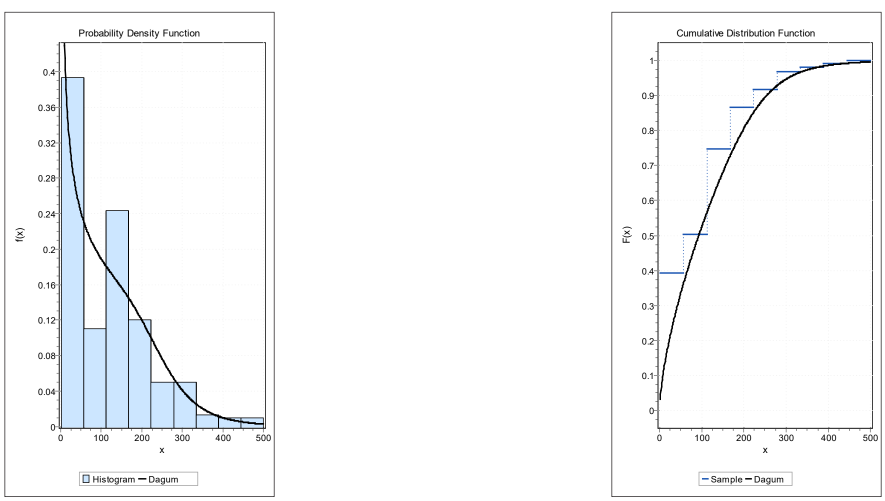
Log-Pearson 3

Number of spurs



Inverse Gaussian 3P

Yield



Dagum 3P

Fig. 2. Graphs of probability density function and cumulative distribution function of best fit