An Analysis of the Distribution of Seed Size: A Case Study of the Gymnosperms

Lorentz JÄNTSCHI1, Rodica C. SOBOLU2, Sorana D. BOLBOACĂ3*

1Horticulture and Landscaping Chair, University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, 3-5 Măneștir, 400372 Cluj-Napoca, Romania; lorentz.jantschi@gmail.com
2Department of Chemistry, Technical University of Cluj-Napoca, 103-105 Muncii Bld., 400641 Cluj-Napoca, Romania
3Department of Medical Informatics and Biostatistics, "Iuliu Hațieganu" University of Medicine and Pharmacy Cluj-Napoca, 6 Louis Pasteur, 400349 Cluj-Napoca, Romania; sbolboaca@unfcluj.ro (corresponding author)

Abstract
Seed morphology is one of the most addressed issues in seeding plants studies due to its importance in the propagation of seeding plants, which can be related to the influence of the environment of the genetic structure in plant populations. A distribution analysis was conducted on extreme values (minimum and maximum) of seed width and length for gymnosperms spread within the Carpathian Mountains region. Combining the probabilities from independent tests successfully limited the best-fit distribution to a small number of distribution laws. Analyses revealed that the extreme values of investigated seed width and length best fit a log-logistic distribution or one of its generalised forms. The left-weighting of the distribution (to small sizes) revealed a better adaptation of small-sized seeding plants, which can be related to the influence of the environment of genetic structure in plant populations. A distribution analysis was conducted to solve Darwin’s mystery regarding the origin of angiosperms led to the conclusion that Gnetales and various fossil groups are related to angiosperms, forming the anthophytes and sustaining the idea that angiosperm origins and homologies should be sought among extinct seed plant groups (Bowe et al., 2000; Chaw et al., 2011).

Introduction
Several studies have addressed seed morphology as a relevant trait for the prediction of successional trajectories (Grime et al., 1997), the role of herbivores in vegetation succession (Aarrestad et al., 2011), the effect of eutrophication on ecosystems (Kadoya et al., 2011) and for the effect of climate change on species diversity (Floran et al., 2011; Sieck et al., 2011; Walck et al., 2011). Grime and co-authors suggested that investigation of seed morphology might provide a Darwinian underpinning (Grime et al., 1997) for Odum’s theory of ecosystem maturity (Odum, 1969). Seed morphology serves as a characteristic for making classifications within sub-generic groups and deducing phylogenetic relationships (Coulter and Chamberlain, 1910). The documented molecular evidence reliably shows that similarities in plants and their seed morphology can be derived independently without strong phylogeny support (Bowe et al., 2000; Chaw et al., 2011). For example, studies conducted to solve Darwin’s mystery regarding the origin of angiosperms led to the conclusion that Gnetales and various fossil groups are related to angiosperms, forming the anthophytes and sustaining the idea that angiosperm origins and homologies should be sought among extinct seed plant groups (Bowe et al., 2000; Chaw et al., 2011).

One important issue often addressed in relation to plant morphology is how the environment can drive and explain the genetic structure in plant populations (Givnish, 2010; Loveless and Hamrick, 1984). For instance, the distribution of seeds can be considered from various points of view: phenology and geographical distribution in relation to seed morphology (Norman, 1994), dispersal of seeds by animals (Nathan et al., 2008; Schupp, 1993), and plant-animal interactions within morphological parameters (Szentesi and Jermy, 1995). The overall conclusion of these studies is that bruchids or other seed predators do not likely drive the evolution of the seed size of the plant species (Nathan et al., 2008; Norman, 1994; Schupp, 1993; Szentesi and Jermy, 1995).

Seed size is a central trait of plant ecology and evolution (Moles et al., 2005a; 2005b), conditioning the probability of seed abundance and dispersal (Guo et al., 2000), predation, germination (Pearson et al., 2002), and seedling survival, even within a single species (Obeso et al., 2011). Additionally, evidence of early plant performance can be found by examining the distribution of seed size (Rodríguez-Calcerrada et al., 2011).

Studies of the link between seed size and growth have been reported since 1908 (Zavit, 1908). For cereals, origination from large seeds led to higher productivity, competitive abilities against weeds, and pests compared to those grown from small seeds (Baalbaki et al., 1997).
The question of whether the living gymnosperms form a clade still remains (Burleigh et al., 2004; Jiao et al., 2011; Rydin et al., 2002), especially as to whether they represent a "natural" group in early classification systems (Schmidt and Schneider-Poetsch, 2002). This question inspired the present research. In this study, extreme values of seed size (minimum and maximum values of both width and length) and the overall distribution from 79 species assigned to the gymnosperms group were analysed. The distribution of seed size among species of gymnosperms is expected to provide further information regarding the origins of the group.

Materials and methods

The measurements of seeds from the Carpathian Mountains region (79 species) were included in the present study. The observed data were obtained from the study by Bojňanský and Fargašová (2007) and are presented in Tab. 1.

All continuous probability density functions available in EasyFit Professional (v 5.2, MathWave Technologies, USA) were used in the analysis. Kolmogorov-Smirnov (Kolmogoroff, 1941; Smirnov, 1948), Anderson-Darling (Anderson and Darling, 1952), and Chi-Squared (Fisher, 1922a, 1924; Pearson, 1900) statistics were applied to measure the agreement between the observations and the model. A global measure of agreement between the observation and the model was calculated for each given probability density function (PDF) using the Fisher’s Chi-Squared (abbreviated as F-C-S) formula (Fisher 1948). The global probability of observation of a specific value based on Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D), and Chi-Squared (C-S) used is presented in Eq (1):

\[ X^2 = \sum_i \ln(1/p_i) \sim \chi^2(n, \overline{p}) \]

where \( X^2 \) = value of the Chi-Square statistic; \( p_i \) = probability of the \( i \)-th test; \( \chi^2 \) = value of the Chi-Squared parameter from the Chi-Square distribution; \( n \) = number of tests; and \( \overline{p} \) = global probability.

A step-based procedure was applied to reduce the number of likely PDF alternatives:

Step 1: For every list of observations for the 79 species, including minimum and maximum length and width, the series of alternatives were constructed using a maximum likelihood estimate (Fisher, 1922b).

Step 2: The alternative distribution laws that failed to meet the criterion to not be rejected at 20% risk of error by either the Kolmogorov-Smirnov or Anderson-Darling statistics were removed. It is well known that both the Kolmogorov-Smirnov and Anderson-Darling statistics are more frequently susceptible to type I than to type II errors (Jäntschi and Bolboacă, 2009).

Step 3: All alternative distributions that failed to meet the criterion to not be rejected at 5% risk of error by the Chi-Squared statistic were removed from further analysis. This criterion was imposed knowing that the Chi-Squared statistic is more susceptible to type II than to type I errors (Bolboacă, 2011; Neyman and Pearson, 1967).

Step 4: The global probabilities of observations were computed using the probabilities given by the Kolmogorov-Smirnov and Chi-Square statistics. As a combined statistic, we assumed that the associated Chi-Squared statistic would not be further exposed to type II errors, and consequently, a filter with 20% risk of error was applied to further reduce the list of alternatives.

Step 5: The four obtained lists were cross-referenced to obtain the distribution law that fit best for all the lists.

Step 6: The procedure described by Fisher (1948) was applied to the revised list, but for eight probabilities in this instance instead of two (to verify the assumption that a given distribution law fits for every one of the four independent sets of observations - maximum and minimum length and width).

Step 7: We next removed from the intersected list all alternatives with negative values (negative-domain). This step was implemented because the investigated character (the seed sizes) could not take on any negative values. It would be inappropriate to apply this step at the beginning of the analysis because in the general case, it is possible to accept a distribution with a negative domain if the probability of its negative values falls within the range of sampling or in the range of computational error.

Results

The summary of the results obtained after applying the first four steps of the analysis approach is presented in Fig. 1.

![Fig. 1. Reduction of PDF alternatives from steps 1-4 (A-D = Anderson-Darling statistic; K-S = Kolmogorov-Smirnov statistic)](image)
Tab. 1. Seed width and length for gymnosperms observed in the Carpathian Mountains region - from Bojňanský and Fargášová (2007)

<table>
<thead>
<tr>
<th>Species</th>
<th>Length (mm)</th>
<th>Width (mm)</th>
<th>Species</th>
<th>Length (mm)</th>
<th>Width (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>min.</td>
<td>max.</td>
<td>min.</td>
<td>max.</td>
<td>min.</td>
</tr>
<tr>
<td>C. thumbergii</td>
<td>30.0</td>
<td>45.0</td>
<td>25.0</td>
<td>35.0</td>
<td>3.8</td>
</tr>
<tr>
<td>C. engelmannii</td>
<td>2.0</td>
<td>3.0</td>
<td>1.5</td>
<td>2.0</td>
<td>10.0</td>
</tr>
<tr>
<td>C. glauca</td>
<td>2.6</td>
<td>3.1</td>
<td>1.4</td>
<td>1.7</td>
<td>5.0</td>
</tr>
<tr>
<td>C. reticulata</td>
<td>1.5</td>
<td>2.3</td>
<td>0.8</td>
<td>1.2</td>
<td>4.4</td>
</tr>
<tr>
<td>C. obtusa</td>
<td>3.5</td>
<td>4.0</td>
<td>1.8</td>
<td>2.1</td>
<td>5.0</td>
</tr>
<tr>
<td>C. sempervirens</td>
<td>3.0</td>
<td>3.6</td>
<td>1.8</td>
<td>2.2</td>
<td>4.0</td>
</tr>
<tr>
<td>C. orientalis</td>
<td>2.5</td>
<td>3.5</td>
<td>1.8</td>
<td>2.2</td>
<td>4.0</td>
</tr>
<tr>
<td>P. abies</td>
<td>4.0</td>
<td>5.0</td>
<td>2.0</td>
<td>3.0</td>
<td>14.0</td>
</tr>
<tr>
<td>P. engelmannii</td>
<td>2.0</td>
<td>3.0</td>
<td>1.5</td>
<td>2.0</td>
<td>10.0</td>
</tr>
<tr>
<td>P. glauca</td>
<td>2.6</td>
<td>3.1</td>
<td>1.4</td>
<td>1.7</td>
<td>5.0</td>
</tr>
<tr>
<td>P. marrana</td>
<td>1.5</td>
<td>2.3</td>
<td>0.8</td>
<td>1.2</td>
<td>4.4</td>
</tr>
<tr>
<td>P. obvata</td>
<td>3.5</td>
<td>4.0</td>
<td>1.8</td>
<td>2.1</td>
<td>5.0</td>
</tr>
<tr>
<td>P. sitchensis</td>
<td>3.3</td>
<td>3.8</td>
<td>1.5</td>
<td>2.0</td>
<td>5.0</td>
</tr>
<tr>
<td>P. halepensis</td>
<td>5.5</td>
<td>6.5</td>
<td>3.0</td>
<td>3.5</td>
<td>4.2</td>
</tr>
<tr>
<td>P. jeffreyi</td>
<td>10.0</td>
<td>12.0</td>
<td>6.0</td>
<td>7.0</td>
<td>5.2</td>
</tr>
<tr>
<td>P. montezumae</td>
<td>5.0</td>
<td>5.4</td>
<td>3.8</td>
<td>4.2</td>
<td>5.0</td>
</tr>
<tr>
<td>P. mugo</td>
<td>4.0</td>
<td>4.5</td>
<td>2.5</td>
<td>3.0</td>
<td>4.0</td>
</tr>
<tr>
<td>P. nigra</td>
<td>6.8</td>
<td>7.5</td>
<td>3.8</td>
<td>4.5</td>
<td>4.4</td>
</tr>
<tr>
<td>P. pallidissima</td>
<td>7.0</td>
<td>9.0</td>
<td>4.2</td>
<td>4.6</td>
<td>6.0</td>
</tr>
<tr>
<td>P. picea</td>
<td>7.5</td>
<td>8.5</td>
<td>5.0</td>
<td>6.0</td>
<td>3.7</td>
</tr>
<tr>
<td>P. pinaster</td>
<td>7.6</td>
<td>8.0</td>
<td>4.2</td>
<td>4.6</td>
<td>4.5</td>
</tr>
<tr>
<td>P. pini</td>
<td>15.0</td>
<td>20.0</td>
<td>7.0</td>
<td>11.0</td>
<td>3.5</td>
</tr>
<tr>
<td>P. ponderosa</td>
<td>7.0</td>
<td>9.0</td>
<td>5.0</td>
<td>6.0</td>
<td>4.8</td>
</tr>
<tr>
<td>P. radiata</td>
<td>6.0</td>
<td>8.0</td>
<td>3.5</td>
<td>4.5</td>
<td></td>
</tr>
</tbody>
</table>

The identified alternatives for min-length, max-length, min-width and max-width were intersected (Step 5), and the alternative list was reduced to the following PDFs: Burr (3P), Burr (4P), Dagum (3P), Dagum (4P), Generalised Extreme Value (3P), log-Logistic (2P), log-Logistic (3P), lognormal (3P), Pearson 5 (2P), Pearson 5 (3P), Pearson 6 (3P), and Pearson 6 (4P). No alternatives were removed from the list in the sixth step.

Seven PDFs were removed after encountering negative domains for the sizes of Gymnosperm seeds. The excluded probability distribution functions from Step 7 of the analysis are presented in Tab. 2.

The final list of distribution laws obtained by applying the proposed approach is presented in Tab. 3.

As the results presented in Tab. 3 show, the investigated seed sizes of the gymnosperm group best fit a generalised log-Logistic (3P) distribution. This distribution was plotted for both maximum and minimum values in Fig. 2.

The small dispersal of seed dimensions for both width and length is observed on the plotted distribution (Fig. 2).
The extreme values of seed size (width and length) from 79 species assigned to the gymnosperm group were successfully analysed with reference to probability distribution laws. A seven-step approach was developed and applied to identify the distribution law that best fit the investigated characteristics. In the first four steps, the reduction of alternatives varied from 0% (min-length, 3rd step; max-length, 3rd step; max-width, 3rd step; and max-width, 4th step) to 67% (max-width, 2nd step). The most reductions were observed in the second step relative to the first step, followed by the 4th step. A list of twelve alternatives was ultimately obtained after the four lists were intersected. The negative domains for size of gymnosperm seeds further led to a narrower list of probable distributions (see Tab. 3).

The analysis identified five probable distributions, and four of them were generalizations of the log-logistic distribution. According to the ΣMLE scores (Dey and Kundu, 1988; Holcomb et al., 1999; Nixon and Thompson, 2004), the descending classification of distributions is as follows: log-logistic (3P) - log-normal - log-logistic (2P) - Burn (4P) - Burn (3P). The log-normal distribution was ranked somewhere between two log-logistic distributions, in agreement with the literature Dey and Kundu (2010). The log-logistic (3P) distribution proved to be able to characterise the extreme values of length and width of investigated seeds according to the ΣMLE score criterion. The three-parameter log-logistic distribution is frequently used in models of flood frequency (Ahmad et al., 1988; Hosking and Wallis, 1997; Robson and Reed, 1999) and is related to the modelling of environmental conditions. The relationship between seed size and environmental conditions may facilitate adaptation. Foster and Janson (1985), for example, demonstrated a relationship between large seed size and establishment in shady, stable plant associations. Moreover, Eriksson and Kainulainen showed that the selection for increasing seed size associated with the expansion of modern type tropical forests spurred a competition/colonization trade-off, initiating a reversed evolutionary trajectory towards smaller seeds (Eriksson and Kainulainen, 2011).

The distribution of extreme values may serve for breeding, increasing probabilities of finding seeds of specific size across the species of the genus (Bettge et al., 2002; Cheng et al., 2010; Cilas et al., 2010). Some studies have shown that adaptation is directly linked to seed size, that is, the smaller the better (Howe and Smallwood, 1982). However, in

---

**Tab. 2. Rejected distribution laws based on negative domains**

<table>
<thead>
<tr>
<th>Distribution</th>
<th>EstMin</th>
<th>Observable - Discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dagum (3.8847; 18.3891)</td>
<td>0.0000</td>
<td>-2.8574 min-length - Negative part in the domain of Dagum (4P); Dagum (4P) is rejected, and this attracted the rejection of the Dagum (3P) as well;</td>
</tr>
<tr>
<td>Pearson5 (4.5623; 23.984; -0.42052)</td>
<td>0.0000</td>
<td>-0.4205 min-width - Negative part in the domain of Pearson5 (3P); Pearson5 (3P) is rejected and also attracted the rejection of the Pearson5 (2P);</td>
</tr>
<tr>
<td>Pearson6 (38.353; 5.25; 0.45002; -0.1967)</td>
<td>0.0000</td>
<td>-0.1967 min-width - Negative part in the domain of Pearson6 (4P); Pearson6 (4P) is rejected and also attracted the rejection of the Pearson6 (3P);</td>
</tr>
</tbody>
</table>

EstMin = estimated minimum value; Dagum (4P: k, α, β, γ): three continuous shape parameters (k > 0, α > 0, and β > 0) and one continuous location parameter (γ); Dagum (3P: k, α, β); GenExtreme (3P: k, σ, μ): Generalised Extreme Value Distribution with shape parameter (k), scale parameter (σ > 0), and location parameter (μ); Pearson5 (3P: α, β, γ): shape parameter (α > 0), scale parameter (β > 0), location parameter (γ, for γ = 0 → Pearson5 (2P)); Pearson6 (4P: α1, α2, β, γ): shape parameters (α1 > 0, α2 > 0), scale parameter (β > 0), location parameter (γ, for γ = 0 → Pearson6 (3P))
other cases, seed size maximization is a breeding objective (Damayanti et al., 2010; Dansi et al., 2010; Saxena, 2008).

The present study identified the most likely probability distribution function for the extreme values of seed size across Gymnosperms. The probability of finding seeds of a given size could thus be obtained from the probability distribution function of extreme seed size values. Our study showed that the investigated seed sizes of the gymnosperm group most likely fit a generalised log-logistic distribution. It is well known that maximum and minimum values are order statistics and depend on the sample size and study design. The main limit of our study is that the analysis was performed based on the assumption that the sample size was sufficient and that the design utilised to measure seed width and length was reliable and valid (Bojňanský and Fargašová, 2007). The sample’s extreme values could be viewed as measures of dispersion when the range is of interest as well as a measure of location when the mid-range is of interest. However, the sample’s minimum and maximum values could be used to obtain the prediction interval as estimators for values outside the sample (e.g., for n = 15, a prediction interval of 93% is obtained for the next random observation; the 16th observation falls between the smallest and the largest observation with a chance of 93%) (Whitmore, 1986). However, why are the dimensions of seeds of interest? Seed gymnosperms produce proteins (Jensen and Berthold, 1989; Konarev et al., 2008) that are important characters used in phylogenetic analysis (Braun et al., 1996; Jensen, 1994). Moreover, the proteins could later be used as a source of food plant proteins, and given a relationship between seed dimensions and composition, it is therefore plausible the results of our study could be relevant.

Conclusions

The minimum and maximum values for seed width and length of investigated gymnosperms best fit a generalised log-logistic distribution. This information was obtained by combining the probabilities from independent tests. The extreme values of seed dimensions could be used to predict the dimension of a random observation.

Acknowledgments

The study was supported by the European Social Fund, Human Resources Development Operational Program, project number 89/1.5/62371 through a fellowship for L. Jäntschi. The funding source had no role in the study design, data collection and analysis, decision to publish, or in preparation of the manuscript.

References

Aarrestad PA, Masunga GS, Hytteborn H, Pitlagano ML, Marokane W, Skarpe C (2011). Influence of soil, tree cover and large herbivores on field layer vegetation along a savanna


Jensen U, Berthold H (1989). Legumin-like proteins in

52


