Supplementary Statistical notes

Article title: **Anomalous germination of dormant dehulled red rice seeds provides a new perspective to study the transition from dormancy to germination and to unravel the role of the caryopsis coat in seed dormancy**

Journal: Seed Science Research

Author: **Alberto Gianinetti**

Dr. Alberto Gianinetti

Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria,

Genomics Research Centre,

via S. Protaso 302,

29017 Fiorenzuola D'Arda (PC),

Italy

e-mail: alberto.gianinetti@crea.gov.it

Statistical notes on fitting data to a Gompertz distribution

For curve-fitting three datasets were considered: coat tearing, ct, the sum of pericarp splitting and coat tearing, ps+ct, and growth stage S1. To avoid any artificial overfitting, every dataset for distribution-fitting was stopped once its maximum value was reached (whereas in Fig. 3A all the datasets are prolonged up to the last overall observed event, namely the last seed attaining stage S1, in order to facilitate comparison). For distribution-fitting, the percentages were corrected for both the (very small) number of seeds that showed normal germination (assumed as seeds germinating by pericarp splitting within the first two weeks of incubation, i.e. 0.36% (of the initial number of seeds, that is, one seed), which was excluded by data for fitting) and for the viability observed at the end of the experiment. In other words, for this analysis, data (diminished by 0.36%) were expressed as a fraction of their end value (also diminished by 0.36%) because distributions assume maximum values of 1, and once all the seeds have either germinated or rotted, end values represent maximum values.

To have a preliminary guess of the distribution parameters, a cumulative distribution function (cdf) was built by manually varying the parameters to obtain a distribution similar to that of S1. Then, from these preliminary guesses ($k =$ 0.001 and $\alpha =$ 0.01), the nonlinear regression procedure of Systat employing least-squares estimation (by the Gauss-Newton iterative method) was used to fit the cumulative Gompertz distribution to the observed data (expressed as fractions of 1), because such procedure minimizes the squared deviations of model's predictions from observed data, thereby optimizing the fitting of the curve. In fact, the main target of this approach was to show that the timecourse of the anomalous germination indeed fits a Gompertz distribution. By this way, the three cumulative datasets for ct, ps+ct and S1 produced the following results:

	ct	**ps+ct**	**S1**
Raw R^2 (1-Residual/Total)	1.000	0.999	0.999
Mean Corrected R^2 (1-Residual/Corrected)	0.999	0.998	0.997
R^2 (observed vs predicted)	0.999	0.999	0.997
k	0.000222	0.000416	0.000430
Wald 95% Confidence Interval	0.000212-0.000233	0.000394-0.000438	0.000400-0.000461
α	0.01720	0.014585	0.01429
Wald 95% Confidence Interval	0.01695-0.01745	0.014289-0.014882	0.01389-0.01469

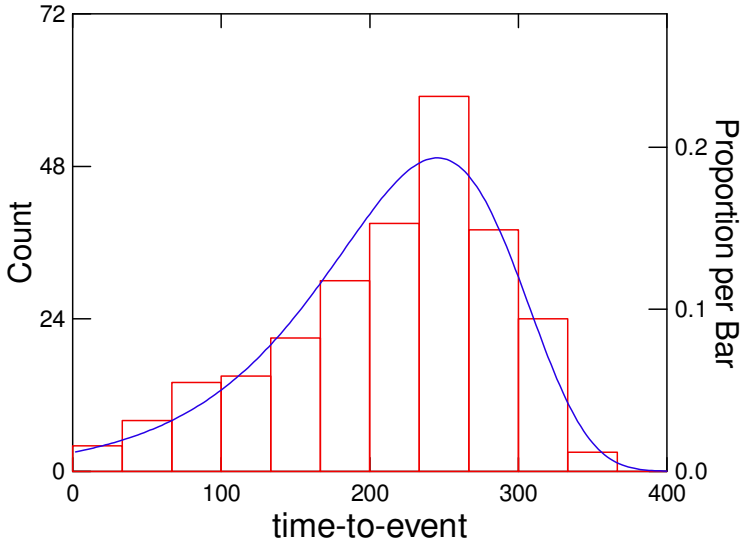
Clearly, all three datasets fit very well a cumulative Gompertz distribution. The overall percentage of seeds showing a rupture of the caryopsis coat (ps+ct in Fig. 3A) was, however, deemed to be more interesting than S1 because the former shows the direct effect of the failure of the caryopsis coat (the failure of some function is the typical reason for observing a Gompertz distribution; Kirkwood, 2015), and it is also subject to a reduced interference by seed mortality (that is, only the percentage of seeds rotted before the rupture of the caryopsis coat has to be considered, which is a useful feature because germination and rotting after the rupture of the caryopsis coat are not independent from each other). Despite a slightly lower fitting, which however is still extremely good, the overall percentage of seeds showing a rupture of the caryopsis coat (ps+ct) was also preferred to the percentage of seeds showing only tearing of the caryopsis coat (ct) because, apart from the seeds germinating (all by pericarp splitting) within the first two weeks of incubation (0.36%) that are typically considered to

undergo normal germination and were therefore subtracted from the cumulative data for this analysis, the pericarp splitting observed later can be interpreted as due to a failure of the ventral junction consequent to the embryo thrust, rather than to a programmed weakening of this specialized tissue. Indeed, if failure can occur at the caryopsis coat around the embryo, it can occur at the ventral junction as well, since the latter is just the predetermined breaking site for germination, and can therefore be expected to fail first.

In spite of the good fitting, the nonlinear least-squares procedure is sometimes deemed to provide an estimation of the curve parameters that could be biased, whereas the maximum likelihood procedure is superior to this aim (Garg *et al*., 1970; O'Neill *et al*., 2004). In fact, variation in the proportions of germinated seeds will vary with time, being largest at intermediate monitoring intervals and smallest at the initial and final intervals where germination activity is low (Ritz *et al*., 2013). This means that a fundamental assumptions of variance homogeneity implicitly underlying nonlinear regression by LSe is not satisfied (O'Neill *et al*., 2004; Ritz *et al*., 2013). In addition, a relevant problem involved in regressing a time series is the autocorrelation of the errors that occurs when considering cumulative data (Mandel, 1957; O'Neill *et al*., 2004; Appendix S3 in Mesgaran *et al*., 2013): since subsequent cumulative recordings are made on the same sample(s), the errors are not independent, as required by regression analysis. In fact, each successive datum has an additional error term accruing from the proceeding observations, thus that cumulative counts have cumulative, heteroscedastic errors (Mandel, 1957; O'Neill *et al*., 2004; Appendix S3 in Mesgaran *et al*., 2013). Anyway, a solution to these problems is available (Mandel, 1957; O'Neill *et al*., 2004): in the probability density function (pdf) the error of a probability datum observed at a given time is independent of the errors observed at any other time. Hence, a function based on single-time events, rather than the cumulative distribution function, should be used for parametric estimations in curve-fitting of time series by MLe, just because in the pdf there is no autocorrelation of the errors and, since data are not cumulated, it also alleviates problems related to non-homogenous errors, as neither the errors are therefore cumulated (Mandel, 1957; O'Neill *et al*., 2004).

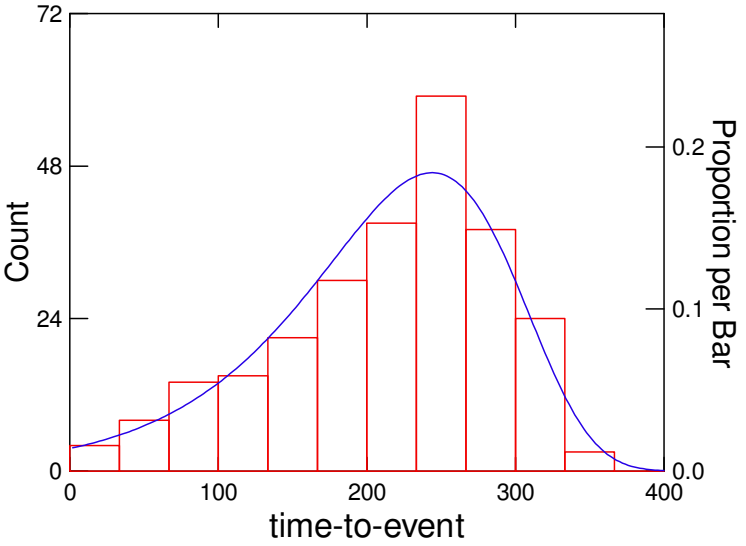
Cumulative data for ps+ct were then used to generate a set of time-to-event data (where an event is the attainment of the ps+ct stage). In other words, the time required by every seed to attain the ps+ct stage (actually, the time of recording) was used, and when more seeds (n seeds, with $n > 1$) were recorded to attain the ps+ct stage at a given observation time, their times to-event were considered separately, that is, the recording time was repeated n times (as, according to the traditional custom of Survival Analysis, the Systat 12 program does not allow to indicate frequencies for times to-event). Thereafter, the subroutine for Fitting Distributions of Systat 12 was used to fit a continuous Gompertz probability density distribution, $g(\mathrm{t}, k, \alpha) = k\mathrm{e}^{\alpha \mathrm{t}}\exp(-k(\mathrm{e}^{\alpha \mathrm{t}}-1)/\alpha)$, to the time-to-event data (see the graph below, wherein the curve is the fitted pdf and the actual data are grouped in histograms), by estimating parameters with a Maximum Likelihood estimation (MLe) procedure. Estimated parameters were: $k =$ 0.000350 and $\alpha =$ 0.015421, with a Chi-square goodness-of-fit test statistic = 6.013047, degrees of freedom = 7, and p-value = 0.538227 (note that for a Chi-square goodness-of-fit test the null hypothesis is that the data are consistent with the specified distribution). This fitting values were used for Fig. 3B. It can be noted that the value of k estimated by this procedure (MLe) is outside the confidence interval provided by the estimation according to the least-squares estimation (LSe, see the table above).

Fitted Distribution

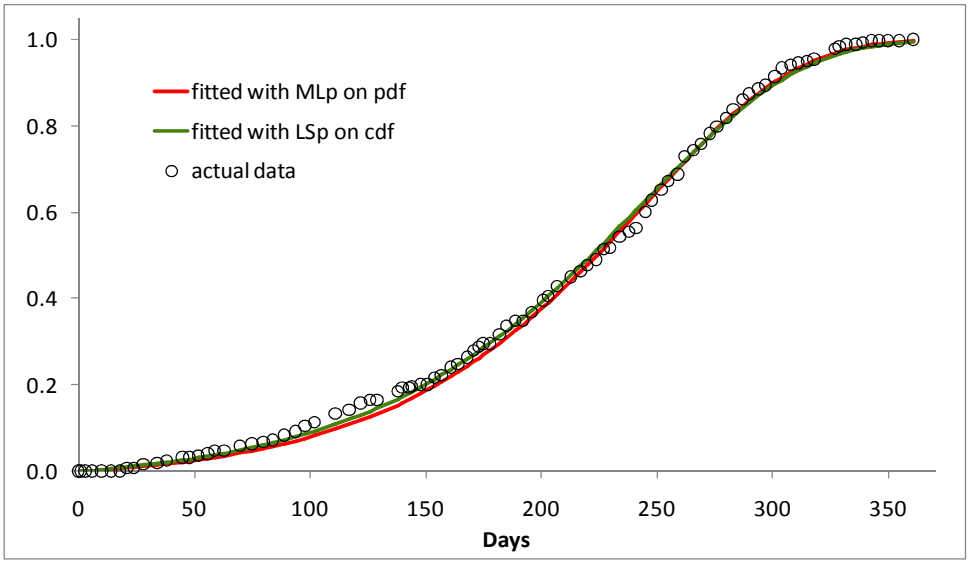


The result obtained by estimating the parameters on the pdf by MLe can be compared with the fitting to the data of the pdf built on the parameters previously estimated on the cdf by LSe (k = 0.000416 and α = 0.01459), whose graph is reported below: Chi-square goodness-of-fit test statistic = 6.194440, degrees of freedom = 9, and p-value = 0.720311. The higher p-value is evidently due to the higher degrees of freedom (as the parameters were given), nevertheless it confirms the good fitting (see the graph below), even though the slightly higher value of the Chi-square goodness-of-fit test statistic indicates the fitting is a little bit less good than with parameters optimized on the pdf by MLe, as expected. Anyway, these findings suggest the statistical difference between results obtained by the two fitting procedures is small in this case.

Fitted Distribution



The two cumulative Gompertz distributions based on fitting the data on the cdf by LSe and on the pdf by MLe are shown below together with actual data.



The two curves are very close and their R^2 (determined by the Excell function RSQ, which returns the square of the Pearson product moment correlation coefficient between the observed and predicted datapoints) is 0.998601 for LSe and 0.998599 for MLe. This confirms that the two fitting procedures provide very close fits in this case. The curve for MLe was chosen for Fig. 3B because this procedure is recommended for providing a better estimate of the parameters (Garg *et al.*, 1970; O'Neill *et al.*, 2004). To provide a better estimation of the goodness-of-fit for the curve in Fig. 3B, an ANOVA was performed by using the GLM procedure of Systat 12 to test the capability of the curve obtained by MLe to predict actual data, and an Adjusted Squared Multiple R (R^2_{adj}) of 0.9985842 was thereby obtained. The R^2_{adj} is deemed a better indicator of goodness-of-fit in nonlinear models (Zar, 1999).

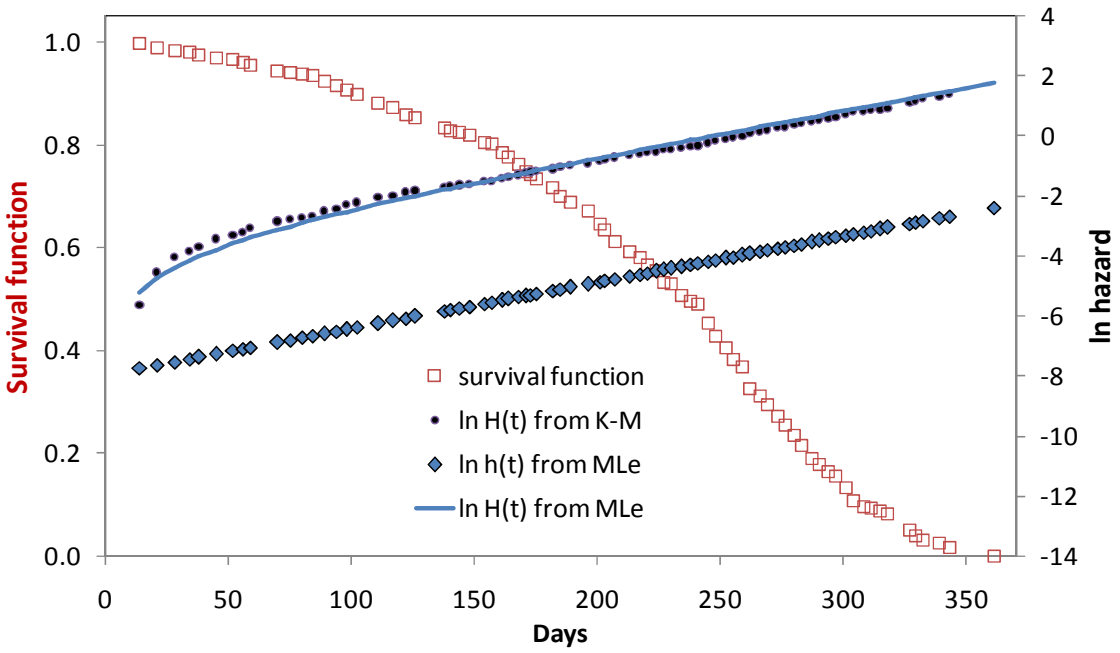
There are at least two censoring problems with fitting germination data, namely, initial lag and the fact that data are recorded only at intervals. In germination studies, it can be useful to subtract the physiological lag time that seeds must undergo before they can germinate (Scott *et al.*, 1984). Since at 30℃ it is less than one day, it was ignored in the present work, as the overall duration of the experiment is much longer than that. Anyway, a preliminary test confirmed that no noticeable change would be obtained if time data were elaborated after they were diminished by one day.

A second problem involved in regressing a time series for germination is that the exact germination time is never known precisely, but it is somewhere between two successive monitoring dates, a feature defined "interval censoring" (Onofri *et al.*, 2011; Ritz *et al.*, 2013). Thus, in interval censoring a given germination percentage is considered implicitly in most analyses to be attained at a precise moment, when in fact it will usually have occurred earlier (but never later) than the time that counts are made (Onofri *et al.*, 2011). This can obviously be a problem whenever the intervals between two successive monitoring times are relevant with respect to the overall timecourse, as the values of the independent variable (time) would then be known only with a relevant uncertainty, whereas regression analysis assumes they are known without relevant errors. Again, this is not considered an issue in the

present study, wherein the long timecourse makes negligible every uncertainty in the precise moment the observed germination percentage is attained between two successive monitoring dates.

Another statistical problem, which can be very important, is that successive observations on the germination curve are highly correlated: the total number of seeds that have germinated at a particular time is highly dependent on the number of seeds that germinated previously (Ritz *et al.*, 2013). This means that not only the errors, but the data themselves are not independent. Thus, even another fundamental assumption underlying nonlinear regression, i.e. independence between proportions, is not satisfied (Ritz *et al.*, 2013). This trouble can affect the whole shape of the timecourse curve, and therefore bias any modeling. A solution to this problem would still be possible if there is some variable that, for each observation, is not dependent on what has previously happened: such a variable could be used to assess whether the modeled data show deviations from the observed data by comparing predicted and actual values of this variable. In Survival Analysis, of which the Gompertz distribution represents a parametric model, this assessment of the parametric model's appropriateness may be made by comparing predicted and actual values of the cumulative hazard (Bradburn *et al.*, 2003). The cumulative hazard is therefore introduced below.

Although the Gompertz distribution is a parametric approach to failure time series, other approaches of Survival Analysis, which provides the general statistical background to study this kind of time-to-event processes, can be applied too. A common one is the Kaplan-Meier product-limit estimator (Hill and Lewicki, 2006). Thus, the nonparametric procedure for Survival Analysis of Systat 12 was used to estimate (by MLe) the Kaplan-Meier (K-M) probabilities that define the survival function of the seeds (the decreasing proportion of initial seeds that survive at the end of each interval). In order to elaborate these data, Systat 12 employs the Turnbull's generalization of the Kaplan-Meier estimator for interval-censored data to extended the application of this estimator. Quite interestingly, this analysis allows to consider both data that are interval censored, just like the ps+ct data, and data that are right censored, like data of seeds rotted before attaining ps+ct, which have not attained such stage prior to the end of their last interval and therefore provide the analysis with more power, that is more seeds, to calculate the frequencies of seeds not attaining ps+ct for the time prior to their rotting. In this analysis, also the seed germinated within 14 days was included, to see if it was effectively associated to a different statistical behavior, since its presence does not affect the overall output of non-parametric analysis. The graph below shows the survival function (red squares), $S(\mathrm{t})$, established according to the Turnbull's generalization of K-M probabilities. The survival probability function gives the probability that an event will not happen until time t. From this analysis it is also possible to obtain values of $\ln H(\mathrm{t})$ (black points in the graph below), that is, of the logarithm of the cumulative hazard, as $H(\mathrm{t})$ is the negative logarithm of the survival function: $H(\mathrm{t}) = -\ln S(\mathrm{t})$. It turns out, therefore, that $\ln H(\mathrm{t}) = \ln(-\ln S(\mathrm{t}))$. The logarithm of the cumulative hazard was plotted because it gives a more linear curve than the cumulative hazard, in the case of a Gompertz distribution, thereby facilitating the comparison. The cumulative hazard measures the total amount of risk of failure that has been accumulated up to time t (Cleves *et al.*, 2008), and as such it has no upper bound: a cumulative probability of 100% is never reached in a finite time, but the cumulative risk that every initial seed will undergo ps+ct within a given finite time is always increasing. As seen, $H(\mathrm{t})$ is directly computable from the Kaplan-Meier estimator of the survival function, whereas the hazard rate, which is the derivative of $H(\mathrm{t})$, needs the step up function of the cumulative hazard is smoothed before it can be differentiated (Cleves *et al.*, 2008).



The hazard rate, $h(\mathrm{t})$, aka the age-specific failure rate, is defined as the probability per time unit that a case that has survived to the beginning of the respective interval will fail in that interval (Hill and Lewicki, 2006). It is an unobserved yet fundamental variable that determines the timing of events in a given process (Allison, 2014), and in nonparametric models it is computed as the number of failures per time units in the respective interval, divided by the average number of surviving cases at the mid-point of the interval (Hill and Lewicki, 2006). Specifically to the case of germination, the hazard is the probability that a seed will germinate in a particular time interval, given that it has not already germinated (Scott *et al.*, 1984). This latent variable is a key feature distinguishing among different models for continuous-time data (Allison, 2014) and is not dependent on what happened previously as it depends only on the underlying process that affects the survived individuals, whatever their number is.

The hazard rate can be computed by both nonparametric (with some further elaboration of $H(\mathrm{t})$, as said above) and parametric models. In the latter case, it is assumed to follow some specific function that determines a parametric distribution (Allison, 2014). In the case of a Gompertz parametric model for survival probability the basic assumption is that $h(\mathrm{t}, k, \alpha) = k\mathrm{e}^{\alpha \mathrm{t}}$ (Garg *et al.*, 1970; Johnson *et al.*, 1995; Kirkwood, 2015). A direct consequence of this assumption is that by plotting the hazard rate on a logarithmic scale against time, a linear increase can be observed for the Gompertz parametric model for survival probability, in accordance with the logarithmic version of the hazard rate equation, i.e. $\ln h(\mathrm{t}, k, \alpha) = \ln k + \alpha \mathrm{t}$ (Kirkwood, 2015). Accordingly, a linear plot of $\ln h(\mathrm{t}, k, \alpha)$, with the values of parameters estimated by the MLe, is shown in the graph above (blue diamonds in the graph above).

In parametric models, the cumulative hazard, $H(\mathrm{t})$, is obtained for every time (upper end of time interval, $\mathrm{t_u}$) as the integral between $\mathrm{t} = 0$ and $\mathrm{t} = \mathrm{t_u}$ of the hazard ratio $h(\mathrm{t})$ (Cleves *et al.*, 2008). Thus, in a parametric Gompertz model $H(\mathrm{t}, k, \alpha) = k(\mathrm{e}^{\alpha \mathrm{t}} - 1)/\alpha$ (Garg *et al.*, 1970). On the other hand, in Survival Analysis it has been established that $H(\mathrm{t}) = -\ln S(\mathrm{t})$, and then $S(\mathrm{t}) = \exp(-H(\mathrm{t}))$, which holds for every function, even parametric (Cleves *et al.*, 2008). Hence, the survival function of the

Gompertz parametric model for survival probability is (Garg *et al*., 1970; Johnson *et al*., 1995): $S(\mathrm{t}, k, \alpha) = \exp\left(\frac{-k\left(\mathrm{e}^{\alpha \mathrm{t}} - 1\right)}{\alpha}\right)$. At each survival function is associated a complementary failure function, $F(\mathrm{t})$, such that $F(\mathrm{t}) = 1 - S(\mathrm{t})$. Therefore, the cumulative Gompertz distribution, which is the failure function for a model assuming that $h(\mathrm{t}, k, \alpha) = k\mathrm{e}^{\alpha \mathrm{t}}$, is $G(\mathrm{t}, k, \alpha) = 1 - \exp\left(\frac{-k\left(\mathrm{e}^{\alpha \mathrm{t}} - 1\right)}{\alpha}\right)$. This shows that the parametric Gompertz distribution is based on the assumption of a specific hazard model, and a cumulative hazard can be promptly computed from the parameters estimated for this distribution. It is therefore possible to compare the plot of $\ln H(\mathrm{t})$ computed from the Kaplan-Meier nonparametric Survival Analysis (black points in the graph above) with the plot of $\ln H(\mathrm{t}, k, \alpha)$ estimated by the parametric Gompertz model (blue line in the graph above). In fact, an informal assessment of a parametric model's appropriateness may be made via plotting the cumulative hazard against values estimated by the model (Bradburn *et al*., 2003). It is evident that the two plots match very well (R^2_{adj} = 0.994).

It is therefore concluded that: (i)- the assumption of a log-linear increase of the hazard rate is supported by the good match between the logarithmic plots of cumulative hazard values estimated according to the parametric model and the values computed by the non-parametric analysis, thus that the adoption of the Gompertz parametric model for failure probability is fully justified; (ii)- the inclusion of right censored data, that is, times to rotting for seeds that had not yet attained ps+ct, does not change the resulting hazard, thus it is confirmed that normalizing the ps+ct data to the end viability to perform the parametric analysis does not alter the outcome distribution; (iii)- the germinative event (by pericarp splitting) that occurred in the first two weeks of incubation in water is compatible with the frequencies expected by the Gompertz model of caryopsis coat failure (i.e., the first black point on the left in the graph above is quite close to the blue line), that is, this event (representing about 0.36% of initial number of seeds) could be due to a failure of the coat at the ventral junction rather than to programmed physiological germination. Clearly, there is no statistical way to say which is the case, this only highlights that the anomalous germination can include a very few early events that are however indistinguishable from normal germination in the absence of a physiological marker. Nonetheless, these events can only represent a tiny portion of seeds, which is therefore absolutely negligible when considering data of normal germination. In fact, based on the parametric Gompertz model adopted in this study, the cumulative hazard for dehulled red rice caryopses is 0.0055 at 14d and 0.0134 at 30d of incubation, which means the total risk of failure up to this times is minimal. Actually, it should be noted that the interpretation of these figures is not straightforward, as they are risks, not probabilities: $H(\mathrm{t})$ reaches the value of 1 at 247d (this means seeds are expected to undergo ps+ct within 247d, on average), 2 at 291d (which means that seeds whose coat has not failed at 247d are then expected to undergo ps+ct within an additional 291-247= 44 days, on average) and 6.29 at 365d (which is the number of times the eventually surviving seeds have exceeded a time threshold at which they were expected, on average for the seeds survived each time at the previous threshold, to undergo ps+ct, and have thus endured without attaining the ps+ct stage up to this time of incubation). The probabilities are given by the cumulative Gompertz distribution, which foretells probabilities of anomalous germination of 0.55% and 1.33% within 14d and 30d, respectively (and 99.82% at 365d). Hence, the timeframes of normal germination (usually 1-2 weeks) and of the anomalous germination (one year and more) are so different that any overlapping can be safely ignored.

LITERATURE CITED

Allison PD. 2014. *Event History and Survival Analysis*, 2nd edn. Los Angeles, CA: Sage Publications, Inc.

Bradburn MJ, Clark TG, Love SB, Altman DG. 2003. Survival Analysis Part III: Multivariate data analysis – Choosing a model and assessing its adequacy and fit. *British Journal of Cancer* **89**:605-611.

Cleves M, Gould W, Gutierrez RG, Marchenko YV. 2008. *An introduction to survival analysis using Stata*, 2nd edn. College Station, TX: Stata Press.

Garg ML, Rao BR, Redmond CK. 1970. Maximum-likelihood estimation of the parameters of the Gompertz survival function. *Applied Statistics* **19**:152-160.

Hill T, Lewicki P. 2006. *STATISTICS methods and applications - A comprehensive reference for science, industry and data mining*. Tulsa, OK: StatSoft, Inc.

Johnson NL, Kotz S, Balakrishnan N. 1995. *Continuous Univariate Distributions*, 2nd edn. Vol. II. New York: John Wiley & Sons.

Kirkwood TBL. 2015. Deciphering death: a commentary on Gompertz (1825) 'On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies'. *Philosophical Transactions of the Royal Society B - Biological Sciences* **370**: 20140379.

Mandel J. 1957. Fitting a straight line to certain types of cumulative data. *Journal of the American Statistical Association* **52**:552-556.

Mesgaran MB, Mashhadi HR, Alizadeh H, Hunt J, Young KR, Cousens RD. 2013. Importance of distribution function selection for hydrothermal time models of seed germination. *Weed Research* **53**:89-101.

O'Neill ME, Thomson P, Jacobs BC, Brain P, Butler RC, Turner H, Mitakda B. 2004. Fitting and comparing seed germination models with a focus on the inverse normal distribution. *Australian and New Zealand Journal of Statistics* **46**:349-366.

Onofri A, Mesgaran MB, Tei F, Cousens RD. 2011. The cure model: an improved way to describe seed germination? *Weed Research* **51**:516-524.

Ritz C, Pipper CB, Streibig JC. 2013. Analysis of germination data from agricultural experiments. *European Journal of Agronomy* **45**:1-6.

Scott S, Jones R, Williams W. 1984. Review of data analysis methods for seed germination. *Crop Science* **24**:1192-1199.

Zar JH. 1999. *Biostatistical analysis*, 4th edn. Upper Saddle River, NJ: Prentice Hall.