Asian Fisheries Science 4(1991): 123-135.
Asian Fisheries Society, Manila, Philippines

https://doi.org/10.33997/j.afs.1991.4.2.001

Extreme Value Theory Applied to the Statistical Distribution of the Largest Lengths of Fish*

SONIA P. FORMACION
JOANNA M. RONGO

College of Fisheries
University of the Philippines in the Visayas
Iloilo, Philippines

VICTOR C. SAMBILAY, JR.

International Center for Living
Aquatic Resources Management
MC P.O. Box 1501
Makati, Metro Manila
Philippines

Abstract

A method is presented which allows one to estimate, based on the maximum lengths of a series of length-frequency samples, and on the theory of extreme values, a single, expected largest value (L_{max}) and its confidence interval. The method is applied to two sets of samples of chub mackerel *Rastrelliger brachysoma* from the Central Philippines.

Introduction

Lai and Gallucci (1987) have clearly demonstrated that misestimation of the growth parameter, L_∞, of the von Bertalanffy growth equation, may be a source of significant error when applying length-based cohort analysis techniques. In addition, Castro and Erzini (1988) have indicated that unreasonable estimates of L_∞ may be a source of bias when applying length-frequency-based analysis techniques such as ELEFAN (Pauly 1987). It therefore seems desirable to investigate methods which are designed to minimize this potential source of bias.

---

*ICLARM Contribution No. 705.

The objective of this report is to demonstrate the utility of the extreme value theory in estimating L_∞ from the largest lengths in length-frequency distributions.

It is confusing to some biologists that a deterministic growth equation such as the von Bertalanffy growth equation is generally obeyed when a relatively large number of fish are studied and yet the same equation may fail miserably when applied to an individual fish growing over time. To clarify this matter, it must be recalled that the von Bertalanffy equation:

$$L_t = L_\infty (1 - e^{-K(t - t_o)}) \qquad \ldots 1)$$

(where L_t is the length at age t; L_∞ is the mean length the fish would have reached if they were to grow to a very old age; K is a constant growth coefficient; and t_o is the hypothetical "age" of the fish at length zero) is a deterministic equation. As such, it indicates that if the parameters L_∞, K and t_o are fixed, or are functionally determined, then for a particular t value, there is only one value of L_t. Thus we can imagine that if we have two identical fish (i.e., of the same species, population and age and initially of the same length) and we let these two fish grow in exactly the same environment, then after a certain period of time, the two fish would have attained exactly the same length. Yet, observations of growth of individual fish growing under controlled conditions do not bear this out. Rather, the most likely event is to have the fish grow to different lengths during the same length of time.

Clearly, fish growth is not deterministic in nature. Rather, fish growth is better represented as a process whose development over time is governed by probabilistic laws, i.e., as a stochastic process.

A stochastic model for growth may be taken as:

$$Y(t) = L(t) + \xi(t) \qquad \ldots 2)$$

where Y (t) is the actual length of the fish at time t,

L (t) is the length that the fish would have attained if it grew according to a deterministic equation such as the von Bertalanffy equation, and

ξ (t) is a random variable, sometimes referred to as a random noise or a perturbation, which represents all other influences affecting growth that cannot be exactly accounted for nor determined as in L (t).

The usual assumptions imposed on the random variable ξ (t) are that:

$$E[\xi(t)] = 0 \text{ and } Var[\xi(t)] = \sigma^2 \text{ (a constant).}$$

Thus, under this stochastic model for growth, the actual length of the fish, Y (t), is a random variable whose distribution is determined by the distribution of ξ (t) and whose values for fixed t values will vary according to this probability distribution.

An important consideration is the fact that

$$E[Y(t)] = L(t)$$

that is, the average in the long run of the length values of the fish of the same species and age is provided by the length value obtained from the deterministic length equation. Thus, if a large number of fish of the same species and age are studied, it is the average characteristic of this large group of fish that is predicted by the von Bertalanffy equation.

One of the parameters affecting the value of L (t), the length predicted by the von Bertalanffy equation, is L_∞, the mean length the fish would have attained if the fish were to grow to a very old age. Various techniques of estimating L_∞ are found in the literature, the majority of which are based on a linear transformation of the von Bertalanffy equation. Attempts to estimate L_∞ independently of the von Bertalanffy equation were provided by such rules of thumb as: take L_∞ to be the biggest length measurement recorded for the population in question; or take L_∞ to be the average length of a number of very old fish. These rules imply a relationship of the type $L_{max} \approx L_\infty$; this contribution is devoted to presenting a method for estimating L_{max} from the maximum lengths (L_m) of a series of length-frequency samples.

This method is independent of the assumed underlying deterministic equation governing the growth of fish. Rather, it is based on the observation that various samples have different values of L_m, indicating that the longest length of fish of a given population is not a fixed quantity but a random variable which takes on different values according to some probabilistic law. Thus, in order to estimate L_{max}, the statistical distribution of the L_m values must first be established.

The distribution of the longest lengths of the chub mackerel *Rastrelliger brachysoma* (Scombridae; local name: hasa-hasa) caught in the Visayan Sea, Central Philippines, from January to December 1984 was studied via the theory of extreme values developed by Gumbel (1954). This theory attempts to explain observed extremes arising in samples of given sizes, and valid for a given period, area or volume, and to forecast extremes that may be expected to occur within a certain sample, period, area or volume.

The application of the theory of extreme values assumes that the following conditions are met:

1) The variables (length, in our case) are continuous;
2) The samples from which the extreme lengths are drawn have a constant distribution with fixed parameters;
3) The extreme lengths are taken from independent samples.

Note that length is a continuous measure. The distribution of fish lengths is assumed normal with fixed parameters for a particular population in a particular area, and the various samples from which the extreme lengths are obtained may be treated as independent. Thus, the above stated assumptions of extreme value theory are met in the case of L_m values.

Theoretical Considerations

Exact Distribution of Extreme Values

Consider the original set of length measurements. F (x) is the probability that any observed length is less than a specified value, x.

Consider also the set of L_m values drawn from the original observations. Let Φ_n (x) be the probability that the largest value is less than a given length x. Therefore

$$\Phi_n (x_n) = F^n (x_n)$$

whose derivative

$$\Phi_n (x_n) = nF^{n-1} (x_n) f (x_n)$$

is the distribution of the largest value among n independent samples.

Similarly, the distribution of the smallest values among n independent samples is:

$$_1\Phi_n(x_1) = 1 - [1 - F(x_1)]^n$$

and

$$_1\Phi_n(x_1) = n[1 - F(x_1)]^{n-1} f(x_1)$$

Here x_1 is the smallest value and x_n is the largest value.

Asymptotic Distribution of Extreme Values

Even if the initial distribution of the sample is unknown, knowledge of the *type* of distribution is sufficient to determine the distribution of the extreme values, by deriving its asymptotic distribution. An asymptotic distribution of a random variable (maximum length, in our case) is any distribution that is approximately equal to the actual distribution of the extreme lengths for a large sample size.

If the variable is infinite to the right, then its cumulative distribution function F (x) approaches 1 as quickly as the exponential function. Variables with this characteristic have asymptotic distributions which belong to the *exponential type*. Variables which are initially distributed as exponential, normal, chi square, logistic and log-normal belong to the exponential group. Under this type of asymptotic distribution, all moments exist but not all distributions with existing moments belong to this class. The distribution of extreme values belonging to this type is:

$$\Phi(y) = e^{-e^{-y}},\ \phi(y) = \alpha\, e^{-y-e^{-y}},\ -\infty < y < \infty \qquad \ldots 3)$$

with the reduced variate $y = \alpha (x - u)$, ... 4)

where x is the variable belonging to the exponential type (and is continuous to the right),

$1/\alpha$ is a measure of dispersion which gives the scale of measure applicable to the observed value of x to that of the reduced variate y, and

u is an average (specifically, the mode for the exponential type) of extreme value distribution.

Extreme-Value Probability Paper

A simple tool for the study of extreme values is the probability paper, which gives a simple graphical method of testing the fit between theory and observations.

Let x be a continuous variate, unlimited in both directions, and for which a linear reduction exists:

$$x = u + y/\alpha \Longrightarrow y = \alpha (x - u)$$

where u is a certain average and $1/\alpha$ a certain measure of dispersion.

The probability paper is a rectangular grid where the observed variate x is plotted on one axis and the reduced variate y is plotted on the other axis. Note that if F (x) is the probability distribution of the variate x and Φ (y) is the probability distribution of the reduced variate y, then

$$\Phi (y) = F (x).$$

Thus, the probability paper also includes the probability Φ (y) = F (x) plotted on a scale parallel to the scale of y.

If the theory holds (that the observations x are distributed according to F (x)), then the observations plotted on the probability paper should fit the straight line given by:

$$x = u + y/\alpha \qquad \text{... 5)}$$

An extreme-value probability paper may be constructed using ordinary graph paper, based on the fact that its horizontal and vertical axes have linear scales. The length units are arranged along the vertical axis, while the reduced variate y and the probabilities are plotted independently along the horizontal axis.

Since the scale of the probabilities is nonlinear, this axis is constructed based on the linear scale of y. Values for Φ (y) are purposely selected for quick interpretation and are computed and positioned using the formula for Φ (y) in equation 3. (Recall that Φ (y) = F (x)). In the graph, y values range from -2 to 7 since Φ (y) of points outside this interval converge toward 0 and 1, respectively.

If a normal probability paper is used instead, the most obvious difference is that the expected extremes will form a *curved* scatterplot, while an extreme-value probability paper clearly shows a straight line for such values. However, the scatter of the observations around the theoretical curve or line seem to be the same in both cases.

The line of expected extremes on the extreme-value probability paper is a straight line because of the linear scale of the reduced variate y (along the horizontal axis), from where the location of the probability values are based, and the assumption of the existence of equation 5, which is another way of writing equation 4.

Methodology

Plotting the Observations

After the values of L_m are identified from their respective length-frequency samples, they are arranged in ascending order.

They are then plotted on an extreme-value probability paper using plotting positions computed from the order of the observations. A plotting position may be interpreted as the cumulative probability assigned to the m^{th} observation.

Gumbel (1954) prefers to use the plotting position, $m/(n+1)$, for the m^{th} observation (n is the total number of extreme values). This choice of plotting position enables the plotting of both the smallest and the largest of L_m values. Alternative positions lose either of the two mentioned extremes.

An observation is plotted using its length as the ordinate and its plotting position as the abscissa.

Estimation of Parameters

The parameters u (intercept) and $1/\alpha$ (slope) of equation 5 are estimated by ordinary least square regression.

Generally, a Type I (or AM) regression, as preprogrammed on most scientific calculators and computer software, will suffice. However, for cases in which the data points are widely dispersed about the regression line, improved estimates of u and $1/\alpha$ can be obtained by using a Type II (or GM) regression. In the latter case, the slope of the GM regression ($1/\alpha'$) is obtained from that of the AM regression ($1/\alpha$) using

$$1/\alpha' = (1/\alpha)/r$$

where r is the coefficient of correlation between the x and the y_n values, while, the corresponding intercept, u', is obtained from

$$u' = \bar{x} - (1/\alpha')\,\bar{y}_n$$

(Ricker 1973).

Here, $\bar{y}_n$ and σ_n are the mean and standard deviation, respectively of the plotting positions, $m/(n+1)$. $\bar{y}_n$ and σ_n are fixed for a specific n and

a table of such values is given in Gumbel's monograph. Likewise, a computer program which includes the computation of y_n and σ_n values is available (see below).

Having computed the parameters u and $1/\alpha$ (or u' and $1/\alpha'$), the theoretical straight line (equation 5), can now be fitted to the observations. With length as the ordinate and the reduced variate as the abscissa, plot several points over the observations on extreme-value probability paper, then connect these theoretical points to draw the line of expected extremes.

A good fit between the observations and the theoretical line implies that the statistical theory of extreme values holds true.

Control Curves

The control curves provide a method for illustrating the goodness of fit of the theoretical straight line to the actual observations. To construct control curves, first compute the standard error of the m^{th} reduced variate using:

$$\sqrt{n}\,\sigma(y_m) = \sqrt{\Phi(y_m)\,[1 - \Phi(y_m)]} \,/\, \phi(y_m) \qquad \ldots 6)$$

where $\Phi(y_m)$ is the frequency of the m^{th} extreme length computed from equation 3, and $\phi(y_m)$ is the first derivative of $\Phi(y_m)$; $\phi(y_m)$ is computed from equation 3. With the value obtained in equation 6 as the numerator, solve for the standard error of the m^{th} observation, x_m, with:

$$\sigma(x_m) = \sqrt{n}\,\sigma(y_m) \,/\, (\sqrt{n}\alpha) \qquad \ldots 7)$$

where $1/\alpha$ and n are parameters previously defined.

The standard error of x_m computed from equation 7 is added to and subtracted from the length x_m found along the theoretical line to obtain the upper point and lower point of the control curves. Plot these two points parallel to the length axis since these are length values. If only one σ unit is used to construct the control curves, then there is a probability of 0.68 that each point is contained in the area enclosed by the two curves. If two σ units are used, the interval between the control curves expands and the probability increases to 0.95.

The control curves are used as a check on the amount of scatter of the extreme length values about the fitted line. In other words, they may be considered as confidence bands for the dispersion of observations about their theoretical values.

Analysis of the data related to control curves may be safely made for $\Phi(y)$ values between .15 and .85. At extremes, errors may be encountered in interpretation.

Expected Extremes

An expected largest value, u_n, (here also: L_{max}) is defined as an extreme length that is expected to occur in a sample of size n with a probability given by:

$$F(u_n) = 1 - 1/n \qquad \ldots 8)$$

Of course, the expected largest value is not the mean largest value. Rather, the expected extreme $x = u_n$ is obtained by first getting the corresponding reduced variate y for the probability $F(u_n)$ using equation 3. (Recall that $\Phi(y) = F(x)$). Then solve for u_n from equation 5 using the parameters u and $1/\alpha$ computed from the observations.

To determine the relationship of the expected largest u_n and the sample size n, the quantity α_n is introduced and is defined as:

$$\alpha_n = nf(u_n)$$

where $f(u_n) = F'(u_n)$ or may be computed using equation 3.

Taking the derivative of equation 8 with respect to n, the following is obtained:

$$du_n/d \log n = 1/\alpha_n \qquad \ldots 9)$$

Therefore, $1/\alpha_n$ measures the increase of u_n with the logarithm of n.

Equation 9 is called the *trend of logarithmic increase of the extremes* and is further stated as follows: If α_n is independent of n, u_n increases with log n. If α_n increases with n, u_n increases more slowly than log n. If α_n decreases with n, u_n increases more quickly than log n.

Therefore, the trend of logarithmic increase of the extremes determines whether expected extremes vary greatly with varying sample sizes.

Results and Discussion

Length data from fishing vessels using purse seine and trawl nets are considered separately. (The data used were provided by the Department of Agriculture and are part of an on-going fish stock assessment project).

Purse Seine

For the purse seine samples, n=31; the smallest and largest L_m values were 18.5 and 28 cm, respectively (Fig. 1).

There seems to be a good fit between the observed lengths and the theoretical line, except for L_m = 18.5 and 24 cm. The fitted line is defined by:

$$y = 0.6124 (x - 22.38)$$

The computed expected extreme u_n for this data set is L_{max}=28.0 cm. Taking the value of the quantity α_n for n = 31 and for other sample sizes, α_n increases with increasing n. Therefore, based on the trend of logarithmic increase of the extremes, the expected extreme u_n increases more slowly than log n. This means that increasing n will also change L_{max}, however, the rate of increase will be rather low.

With the value of u_n given above and taking its reduced variate y, the length interval covered by the control curves is (26.3, 29.6 cm). There is a probability of around 68% that the true maximum length of *R.*

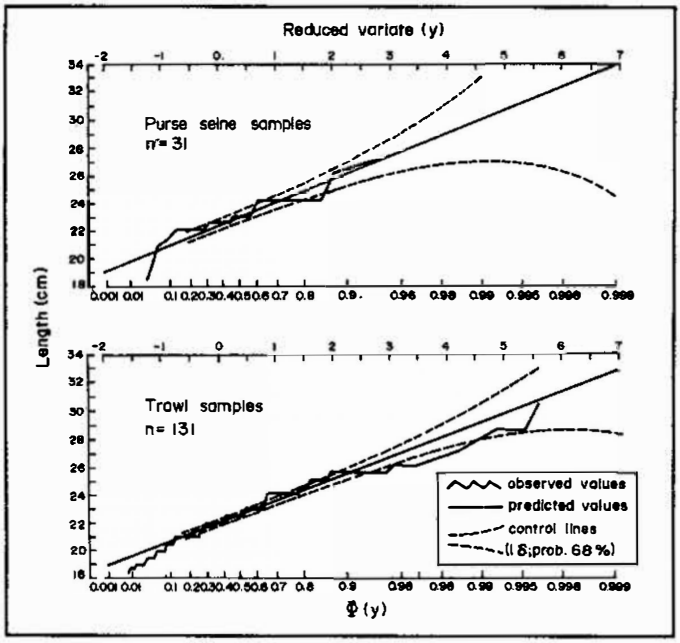

Fig. 1. Observed and predicted (theoretical) extreme lengths of chub mackerel *Rastrelliger brachysoma* from the Central Philippines (see text for interpretation).

brachysoma in the population sampled by the purse seiners lies in the interval (26.3, 29.6 cm).

Trawl Net

This data set (Fig. 1) contains maximum lengths from 131 samples, based on a survey of fishing vessels using trawl nets. The figure shows that these maximum lengths do not greatly depart from the theoretical line:

$$y = 0.6476\ (x - 21.97)$$

when the analysis is based on $\Phi(y)$ values from .15 to .85; the control curves indicate that the above fitted line provides a good fit for the survey data. However, for larger lengths, the observed points lie beyond the interval of the control curves and the fitted line.

The expected extreme u_n is computed as 29.5 cm and validation of maximum length values is based on the interval 28.0 to 31.0 cm.

Interpretation of Results

Combining the results from trawls and purse seines, it can be stated that the value of L_{max} for *R. brachysoma* from the Visayan Sea has a 68% probability of being between 26.3 and 31.0 cm.

Values of L_∞ for various Philippine populations of *R. brachysoma* range between 24.5 and 34.0 on (see Corpuz et al. 1985, Ingles and Pauly 1984, respectively), which bracket the range of L_{max} computed here. Thus, L_{max} may indeed serve for at least preliminary estimation of L_∞ values, independently of growth data.

A computer program written in MSDOS BASICA is available from ICLARM (MC P.O. Box 1501, Makati, Metro Manila, Philippines) for implementation of the method presented here.

Discussion

The statistical theory of extreme values attempts to explain the occurrence of far-removed observations and to predict extreme points that may occur. There is a wide area of interest over which this theory

may be applied. In this case, the statistical theory of extreme values was applied to the maximum lengths of fish (obtained from catches of commercial fishing vessels) and with the objective of validating L_{max} values obtained by other authors.

If the initial sample distribution is known, the exact distribution of extremes may be easily obtained. If it is not known, but the type of distribution is known, the asymptotic distribution of extremes may then be obtained. There are three types of asymptotic distribution available: the exponential type, the Cauchy type and the limited distribution. The exponential type is emphasized since the two other asymptotic distributions may be transformed into this type.

The expected extreme value (L_{max}) for a given number of extreme observations (n) and its corresponding confidence interval may be used to validate extreme length values obtained using other methods. The control curves can be constructed with one σ unit to establish their distances from the theoretical line, which provide a 68% probability that the true maximum length lies in the length interval enclosed by the control curves for the computed expected extreme value. Control curves constructed using two σ units provide the corresponding 95% probability, and thus express a conventional confidence interval.

Acknowledgements

The research leading to this contribution was supported by the Fisheries Stock Assessment-Collaborative Research Support Program (sponsored in part by USAID Grant No. DAN-4146-G-SS-5071-00), conducted in the Philippines by the University of Rhode Island in collaboration with the University of the Philippines and the International Center for Living Aquatic Resources Management.

References

Castro, M. and K. Erzini. 1987. Comparison of two length-frequency based packages for estimating growth and mortality parameters using simulated samples with varying recruitment patterns. U.S. Fish. Bull. 86(4):645-653.

Corpuz, A., J. Saeger and V. Sambilay, Jr. 1985. Population parameters of commercially important fishes in Philippine waters. University of the Philippines, College of Fisheries. Tech. Rep. Dep. Mar. Fish. 6: 1-99.

Gumbel, E.J. 1954. Statistical theory of extreme values and some practical applications, a series of lectures. National Bureau of Standards, Applied Mathematics Series, 33. US Government Printing Office, Washington.

Ingles, J. and D. Pauly. 1984. An atlas of the growth, mortality and recruitment of Philippine fishes. ICLARM Tech. Rep. 13, 127 p.

Kimball, B.F. 1955. Practical applications of the theory of extreme values. J. Amer. Stat. Ass. 50:517-528.

Lai, Han-Lin and V.F. Gallucci. 1987. Effect of variability on estimates of cohort parameters using length-cohort analysis, with a guide to its use and mis-use. Tech. Rep. 1. USAID Stock Assessment Collaborative Research Support Program.

Mood, A.M., F.A. Graybill and D.S. Boes. 1974. Introduction to the theory of statistics, 3rd ed. McGraw-Hill, Inc., New York.

Pauly, D. 1987. A review of the ELEFAN system for analysis of length-frequency data in fish and aquatic invertebrates, p. 7-34. *In* D. Pauly and G.R. Morgan (eds.) Length-based methods in fisheries research. ICLARM Conf. Proc. 13, 468 p.

Ricker, W.E. 1973. Linear regression in fishery research. J. Fish. Res. Board Can. 30: 409-434.

Manuscript received 6 June 1989; accepted 22 February 1990; corrected galley received 24 May 1991.