Nematologica 30 (1984): 187-192, E. J. Brill, Leiden

STATISTICS IN TAXONOMIC DESCRIPTIONS

BY

R. FORTUNER

California Department of Food and Agriculture, Nematology Lab, Room 340, 1220 N Street, Sacramento, California 95814, U.S.A.

A few basic statistical procedures are explained for use in describing new or known nematode species. The mean $\bar{X}$ and the standard deviation s should be calculated from many specimens (up to and above 30) for every measurement in a sample. Computation of the sample mean ($\bar{X}$) and standard deviation (s) permits an estimate of the confidence intervals to be made for the population mean, $\bar{X} + (txs/\sqrt{n})$, and for a given percentage of the population ($\bar{X} \pm 2$ xs). Observation of several populations is advocated to obtain an estimate of mean and range of a measurement in the species.

Keywords: Computation, statistical procedures, descriptions, confidence limits.

Many scientists have expressed concern at the few specimens from which new species are often described: "We must broaden our concept of species by studying more specimens for the variations and the factors that may influence them. ..... The description of a new species based on one or two specimens ..... differing in one single characteristic in one sex is ridiculous" (Chitwood, 1957). "If the species has sharply distinctive characters 5-10 specimens may be sufficient [if not] at least 20 specimens should be examined and measured." (Franklin, 1970). "Unfortunately descriptions of new species, ... based on studies of single or very few specimens are still being published." (Hooper, 1969).

Hooper's observation still holds, and many new species are described from fewer than ten paratypes. In *Helicotylenchus*, among 83 species described between 1959 and 1971, 30 had ten or less paratypes. The number of specimens studied was not indicated for 12 species. For 65 species described since 1972, 32 had ten or less paratypes. The recording of the mean has shown some progress in recent years, but too many species are still published with only the range (extreme values observed in the sample). As for the standard deviation (S.D.), its use is the exception, rather than the rule in new descriptions. For example, in *Helicotylenchus*, during the period 1959-1971, 62 species were described with only the range, and 21 with range and mean. Since 1972, 30 species were described with range only, and 35 with range and mean. Only one species (*H. depressus* Yeates, 1967) was described including the S.D.

Many taxonomists are probably insufficiently familiar with certain simple statistical concepts to use them with confidence. These are to be found in any

textbook (for example Snedecor & Cochran, 1980), but it may be too difficult to extract from these manuals the few methods appropriate for taxonomic use. The present note is intended to offer a simple, practical account of these methods.

DEFINITIONS

A *species* should encompass the characteristics of all the individuals which belong to it. These individuals are grouped into *populations* scattered over the geographical range of the species. The taxonomist considers only one, or in some cases, a few of these populations. As he cannot observe and measure the large number of individuals included in any population, he takes a *sample* which should be representative of the population, if a sampling procedure has been used that takes a truly random sample from the population. The disparity between the few specimens measured from one locality and the innumerable specimens that constitute the species the sample represents explains why statistical procedures should be followed. The methods explained below require the calculation of two quantities: the *mean* and the *standard deviation.* Any textbook on statistics explains the meaning of these parameters and how to calculate them (see paragraphs 3.2 and 3.3 in Snedecor & Cochran, 1980). Most scientific pocket calculators include keys marked "$\bar{X}$" for the mean and "S DEV" or "s" for the standard deviation, and may be used for calculating these values, which should be provided for all measurements in all descriptions and redescriptions of species.

STATISTICAL PROCEDURES

Sample mean and population mean: the confidence interval.

The mean $\bar{X}$, of a measurement as calculated in a sample, gives an estimate of the mean μ in the population: μ is a fixed value for the population, but $\bar{X}$ varies from sample to sample. The sample $\bar{X}$ is the best estimate of the population mean μ, but it is only an estimate. From the standard deviation s and the sample size n, a confidence interval $i = txs/\sqrt{n}$ (for a normal distribution) can be calculated, where n is the number of specimens in the sample, and t is read from a table at the intersection of the line "n-1" degrees of freedom (D.F.) and the column "95% probability level" or "5% 2-tail" (see paragraph 4.10 and Table A4 in Snedecor & Cochran, 1980). The value of $s/\sqrt{n}$ is called the standard error and is a measure of the variability of the population mean.

The interval $+i$ or $-i$ around the mean $\bar{X}$ observed in the sample has 95% probability of including the true value μ of the mean in the population.*)

*) Only in 5% of the cases, will a sample be found whose mean $\bar{X}$ is too far from the mean μ of the population. To reduce this possibility of error, several representative samples from the population can be drawn or measured, or the 99% probability level considered. This makes use of a larger value for t, and will enlarge the confidence interval. The estimate about the value of μ will be less precise, but there will only be 1 chance in a 100 of being wrong.

For example, in a sample of n = 20 specimens of *Helicotylenchus pseudorobustus*, the mean body length was $\bar{X} = 764$ µm, with a S.D.: s = 58. With n-1 = 19 D.F., t at 95% probability level is equal to 2.093. The confidence interval is $i = 2.093 \times 58/\sqrt{20} = 27$ µm.

The interval 764 ± 27 µm, or 737 to 791 µm, has a 95% probability of including the true mean body length of the population.

The coefficient t depends on the D.F. which are equal to the number of specimens measured (n) minus one. If n were infinite, t would be equal to 1.96 (at 95% probability level), but t increases when n (and n-1) diminishes. As a result, the confidence interval becomes broader and less precise.

For example, when the sample mean 764 µm is obtained with only three specimens (t = 4.3), the confidence interval becomes $i = 4.3 \times 58/\sqrt{3} = 144$. The mean in the population is now somewhere in the interval 620-908 µm. The uncertainty about the actual value of the mean is so great as to render meaningless any comparison of body lengths with other species. The larger the number of specimens in a sample, the more precise the estimate of the mean will be. However, a truly random sample of ten is better than a biased sample of one hundred. Bias is often introduced by unadapted extraction methods (it is difficult to extract old mature females, fatter and more sluggish than young ones), or by the selective picking of the largest specimens under the dissecting microscope.

Theoretical range of individual values within the population

The S.D. can give an estimate of the spread of values typical of the population: in any normal distribution with mean μ and standard deviation σ, approximately 95% of the individuals lie in the interval $\mu \pm 2\sigma$ (see paragraph 4.1 in Snedecor & Cochran, 1980). If the sample studied is large enough ($n \geq 30$), the sample mean ($\bar{X}$) and S.D. (s) can be accepted as reasonable estimates of μ and σ. With the population of *H. pseudorobustus* studied above, 95% of the individuals of the population can be estimated to lie within the limits $764 \pm (2 \times 58) = 648$-$880$ µm (the actual sample range was 666-882 µm).

In the rare case when s is given, it is customary to write: $\bar{X} \pm s$. It would be better written "$\bar{X}$, s" which leaves the reader free to estimate the population limits.

The S.D. can also be used to calculate the coefficient of variability $C.V. = (s/\bar{X}) \times 100$ in the sample. The C.V. should be known for discussions of the taxonomic value of various measurements. The actual value of C.V. need not be given in descriptions of new species except in revisions of genera.

The sample range

Until now, it has been assumed that the distribution of the individual values of a measurement in the sample was normal, that is to say with only one peak and the individual values symmetrically distributed around this peak. Tests exist to

test the normality of a distribution, but mostly common sense should be used. For example, if the sample range is approximately the same as the theoretical range in the population ($\bar{X} \pm 2s$), chances are that the distribution is close to normal. If the sample range extends far above or below the population range, the distribution is either skewed to one side (heavy tailed) or outliers (specimens outside an interval $\pm 3s$ across the mean) may be present. The cause of the deviation from normality should be determined if possible: mixture of species, artefacts (flattened specimens artificially increase the body diameter), biological factors (mature females tend to enlarge in some genera), etc. If the distribution is heavy tailed, a transformation of the data (by taking the logarithm of the individual values for example) may restore the normality. Outliers may be discarded for the computation of the mean. They will only appear in the sample range.

When the distribution of a measurement is normal, the sample range is not very informative. A different sample from the same population would most certainly have a different range. The sample range should never be proposed as a substitute for the mean S.D. of the measurement.

Variation between and within samples

Measurements can be affected by food supply (Goodey, 1952; Wu, 1960; Fortuner & Quénéhervé, 1980; etc.), temperature (Evans & Fisher, 1969), and other external, nongenetic factors (Chitwood, 1957; Thorne & Allen, 1959). Nematodes are generally described from a single sample, which represents a single set of external factors. It is impossible to deduce from that single sample the limits of the variation of the species in other habitats. It is the author's responsibility to try to obtain samples from as many and as varied localities as possible. If the author possesses living specimens, he should also cultivate them on different host-plants. The description of a single population should be the exception: species restricted to a particular habitat, species of probable agricultural importance found in one sample intercepted by quarantine stations, etc.

When many samples, representing different populations, are known for a species they can be considered as successive draws taken at random from the ensemble of populations constituting the species. The successive means calculated from the different samples for one character can be averaged to obtain an estimate of the mean value and the standard deviation of the character within the species. (See paragraph 4.4 in Snedecor & Cochran, 1980.) It would be incorrect to pool all the specimens from the different localities in one big sample, because now an evaluation of the variability among locations is sought not the variability within locations.

For example, Fortuner *et al.* (1984) measured twelve samples of *H. pseudorobustus*. The average body length, calculated from the twelve mean body lengths in the samples was $\bar{X} = 715$, $s = 38.5$ μm.

It can now be said that the mean body length in the species *H. pseudorobustus* is about 691-739 μm, and that populations have a 95% probability of having a mean body length of 638-792 μm.

CONCLUSIONS

It is hoped that this article will induce more taxonomists to follow some simple statistical rules: take truly random samples and try to check the normality of the measurements, always measure samples with as many specimens as possible, calculate mean and S.D. for every measurement, and give measurements from as many different populations as possible.

Chitwood (1957) wrote that "size as a species criterion is useless," and some taxonomists distrust the use of measurements for differentiating species. Morphological differences are the best criteria for specific differentiation, and measurements vary, but this variation is contained within certain limits for each species. These limits can be estimated by statistical methods. Knowledge of intraspecific variations should not be restricted to a few specialized studies, but the concern of every taxonomist in descriptions of new species and redescription of known taxa.

ACKNOWLEDGEMENTS

I thank Dr. M. F. Miller of the University of California, Davis, who reviewed the statistical discussion in this article.

RÉSUMÉ

Statistiques dans les descriptions taxonomiques

L'utilisation de quelques simples procédés statistiques pour la description ou la redescription d'espèces de nématodes est expliquée. La moyenne $\bar{X}$ et l'écart-type s devraient être calculés pour chaque mesure à partir d'un nombre suffisant de spécimens (jusqu'à 30 si possible) de l'échantillon. Il est expliqué comment le calcul de la moyenne ($\bar{X}$) et de l'écart-type(s) de l'échantillon permet d'estimer des intervales de confiance pour la moyenne de la population $\bar{X} \pm (txs/\sqrt{n}$, et pour un pourcentage donné de la population ($\bar{X} \pm 2xs$). L'étude de plusieurs populations est conseillée pour l'obtention d'estimations de la moyenne et de l'étendue d'une mesure dans l'espèce entière.

REFERENCES

Chitwood, M. B. (1957). Intraspecific variation in parasitic nematodes. *Systematic Zoology* **6** (1), 19-23.

Evans, A. A. F. & Fisher, J. M. (1970). The effect of environment on nematode morphometrics. Comparison of *Ditylenchus myceliophagus* and *D. destructor*. *Nematologica* **16** (1), 113-122.

Fortuner, R. & Quénéhervé, P. (1980). Morphometrical variability in *Helicotylenchus* Steiner, 1945. 2: Influence of the host on *H. dihystera* (Cobb, 1893) Sher, 1961. *Revue de Nématologie* **3** (2), 291-296.

Fortuner, R., Maggenti, A. R. & Whittaker, L. M. (1984). Morphometrical variability in *Helicotylenchus* Steiner, 1945. 4: Study of field populations of *H. pseudorobustus* and related species. (To be published in *Revue de Nématologie*).

Franklin, M. T. (1970). Morphological variability and the species concept. (Introduction to a discussion.) *Proceedings IX International Nematology Symposium Warsaw, 1967*, 497-503.

GOODEY, J. B. (1952). The influence of the host on the dimensions of the plant parasitic nematode *Ditylenchus destructor*. *Annals of applied Biology* **39** (4), 468-474.

HOOPER, D. J. (1969). Some problems in the systematics of soil nematodes. In: *The Soil Ecosystem*, Systematics Association Publication No. 8, Sheals, J. G. (Ed.), 131-142.

SNEDECOR, G. W. & COCHRAN, W. G. (1980). *Statistical Methods* (7th edition). Iowa St. Univ. Press, Ames, xvi + 507 pp.

THORNE, G. & ALLEN, M. W. (1959). Variation in nematodes. In: *Plant Pathology, Problems and Progress*, 1908-1958. Univ. Wisc. Press, Madison, 412-418.

WU, L. Y. (1960). Comparative study of *Ditylenchus destructor* Thorne, 1945 (Nematoda: Tylenchidae), from potato, bulbous iris, and dahlia, with a discussion of de Man's ratios. *Canadian Journal of Zoology* **38** (6), 1175-1187.

Accepted for publication: December 6, 1983.