Basic statistics applied to aerobiology

Podstawowe metody statystyczne w aerobiologii

CARMEN GALÁN

Department of Plant Biology, University of Córdoba, Campus de Rabanales, Colonia de San José, nş 4, Córdoba, Spain

Abstract

To properly collect, organize, summarize and analyze data an investigator needs an efficient sampling method. Another key requirement is proper presentation of data (numeric or graphic). The paper describes basic statistical methods and their possible use in the field of aerobiology.

Key words: *data analysis, data presentation, aerobiology.*

Streszczenie

Prawidłowe zbieranie, organizowanie, opracowywanie oraz analizowanie danych wymaga przede wszystkim właściwej metody pobierania próbek. Kolejnym ważnym elementem jest właściwe przedstawienie danych (numeryczne bądź graficzne). W pracy opisano podstawowe metody statystyczne, których zastosowanie sprawdza się w dziedzinie aerobiologii.

Słowa kluczowe: *analiza danych, prezentacja danych, aerobiologia.*

(PDiA 2003; XX, 4: 235–238)

Applied statistics covers the use of scientific methods to collect, organize, summarize and analyze data, to draw valid conclusions and to take reasonable decisions based on such analyses.

A key requirement for data analysis is a thorough understanding of the nature of the data concerned. Aerobiology works with data on airborne biological content, and this content varies over space and time. In this respect, various sampling methods are currently in use, depending on the aims of the study to be performed. The general aim is to ascertain the number of particles in a known volume of air. The sampling method should be widely known, and should assume any error inherent in the method. Potential sources of error include the efficiency of the sampling instrument and the method used for sample slide preparation and reading. There may also be human errors, for example during particle identification and computerized data processing (see chapters on *Basic microscopy*; *Calculating the field of view. Scanning the slides*; *Sources of error*). These do not pose an overwhelming problem, given that the whole function of statistics is to address sampling errors.

Another key requirement is that data be properly presented, either numerically or graphically (see chapter on *Data presentation*). This will provide an idea of the behavior of airborne biological particles over time at a given site. Good data processing is also an essential ingredient of analysis. Each sampling site presents different population-related features, depending on climate, topography and other factors. It is essential to determine the start and length of the local pollen season: decisions taken should not be universal in nature, but will vary as a function of the pollen-type under study and the particular features of the sampling site involved. Curves must be smoothed out. One of the most widely-used methods in aerobiology is the running mean, since it is understood that working data are drawn from a temporal series and are not to be viewed in isolation.

In aerobiology, normal data distribution is very much an ideal model. The aerobiological process, after all, starts with pollen emission, which gives to a progressive rise in airborne pollen content involving transport and dispersal; the peak value is subsequently followed by a decrease in airborne particles, involving processes such as sedimentation and impact. However, in most cases, aerobiological data fail to fit a normal distribution pattern, since airborne particle content is constantly, and sometimes unexpectedly, modified by rain, wind, and the effect of local topography. Depending on whether or not the data fit a normal distribution, parametric or non-parametric statistical tests may be used. Parametric methods do not require that that variables be normal, but

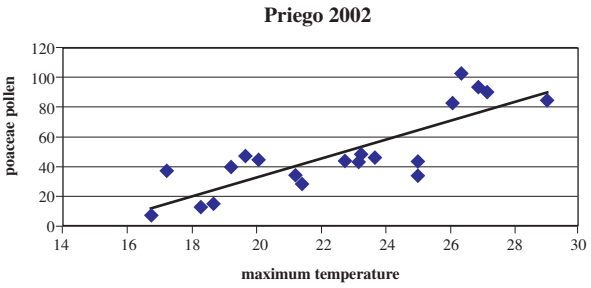

Fig. 1.

Table 1.

	Max T	Min T	Humidity	Rainfall
1995	0.81**	0.59	-0.40	-0.13
1996	0.72**	0.62*	-0.58	-0.34*
1997	-0.14	0.50	-0.80**	-0.42
1998	0.68*	0.78**	-0.60*	-0.59
1999	0.80**	0.75**	-0.37	-0.52
2000	0.45	0.80**	-0.40	-0.26
2001	0.68*	0.70**	-0.55	-0.14
2002	0.57	0.45	-0.78**	-0.22

*Correlation analysis applied to grass pollen grain and some meteorological parameters. * p<0.05; ** p<0.01.*

only that residuals be normal. However, normalization of data has been found to yield better results for the purposes of relating variables to each other, i.e. fewer atypical cases arise. This has prompted the widespread use of various formula-based data transformation methods, including square root transformation, log-normal transformation and the running mean.

Aerobiological hypotheses and limits of significance

We are often forced to take decisions regarding data based on information derived from samples. These decisions are termed statistical decisions. In seeking to reach a decision, it is useful to develop a hypothesis regarding the data involved. Such hypotheses, which may or may not be true, are termed statistical hypotheses.

We often develop a statistical hypothesis with the sole purpose of rejecting it. Thus, for example, if we wish to show that airborne particle concentrations at two different sites differ from each other over time, we formulate the hypothesis that there is *no difference* between them. This hypothesis is referred to as the null hypothesis (H0). Any hypothesis differing from it will be termed an *alternative hypothesis* (Ha). Note that the researcher hypothesis is always the alternative hypothesis. If we reject a hypothesis that should have been accepted, we are said to have made a Type I error; if we accept a hypothesis that should have been rejected, we are said to have made a *Type II error*. Statistical tests are designed to reject the null hypothesis; if this hypothesis is not rejected, it is because we have not found sufficient grounds for rejection. For the same reason, the alternative hypothesis is never rejected, only accepted. In statistical analyses, therefore, we only specify a valid threshold value for rejecting the null hypothesis, i.e. the risk of making a Type I error; this is termed the *significance limit* (α).

In practice, the most widely-used significance limits are 0.05, 0.01 and, though less commonly, 0.001. If we select a significance level of 0.05 (5%) when designing the decision rule, then there are 5 chances in 100 of rejecting the hypothesis when it should have been accepted; in other words, we can be 95% sure of having adopted the correct decision. In the case suggested above, we can state that the difference between airborne particle contents at the two sampling sites was significant to a limit of 95% (*), of 99% (**) or of 99.9% (***).

Correlation between variables

One of the main purposes of scientific research in general is to identify relationships between variables. In aerobiology, it is essential to determine the relationship between airborne particle content and various meteorological parameters, for it air humidity, temperature, rainfall and wind are known to have a decisive effect on airborne particle content. However, the effect of these parameters varies depending on pollen type, sampling site climate, topography and the phenological timing of sampling. Correlation analysis enables us to identify which variables cause the greatest variation in aerobiological data, the strength of the relationship between the variables, and whether that relationship is positive or negative.

When only two variables are involved, we speak of *simple correlation and simple regression*. Supposing X and Y to be the variables in question, a scatter diagram locates the points (X, Y) on a coordinate system formed by an X-axis at right angles to a Y-axis. If the points on the scatter diagram appear to lie in a straight line, the correlation is termed linear. This correlation may be positive or negative, depending on the slope of the line. If all the points appear to lie on a curve, the correlation is termed *non-linear*. By this means, a qualitative assessment may be made of the correlation.

To obtain a quantitative measurement, a *correlation analysis* is required. Depending on whether or not the data fit a normal curve, a parametric test (Pearson) or

a non-parametric test (Spearman) will be used. The degree of association between the variables is expressed as the value of the *correlation coefficient*, which ranges from -1 to +1. The closer this value comes to 1, the closer the relationship between the variables. Plus and minus signs are used to denote positive and negative correlations, respectively.

Once we have identified the meteorological parameters that most affect the production and dispersal of airborne particles, other statistical tests can be used to obtain an equation which will enable us to predict airborne particle concentrations. The elements of the population can be ordered by time, giving a time-related statistic known as a *temporal series*, where „series" implies an ordered succession of values. In seeking an equation enabling prediction, we may opt to use an *analysis of temporal series*, in which time is considered as an independent variable; alternatively, we may use other methods of analysis in which the variable to be predicted is dependent not on time but on some other variable, e.g. meteorological parameters, and time is the framework within which the facts take place (*regression analysis, neuronal networks*).

A temporal series results from four basic components: trend, seasonal variations, cyclical variations and random variations.

Regression analysis

Regression analysis involves a set of techniques, either graphic or analytical, used to determine the relationship between a dependent variable and a number of independent variables.

Regression seeks a line or a mathematical function that will express that relationship, i.e. a line towards which all points on a scatter diagram tend.

There are two types of regression analysis: simple and multiple. In simple regression, the relationship between the two variables is based on a function which may be linear, polynomial, exponential, etc. If all points on a scatter diagram appear to lie on a straight line, this is termed the *regression line*, and is the most common finding in aerobiology. Multiple regression should be based on a linear function, and on a three-dimensional scatter plot forms a plane termed the *regression plane*, which is an extension of the bivariate plot.

Simple linear regression is based on the linear function:

$$Y = a + bX$$

If this function is plotted on a bivariate coordinate system (X and Y axes) it yields a line, where *a* is the Y intercept, or the point at which the regression line crosses the Y axis, i.e. the value of Y when X = 0; *b* is the slope, or angular coefficient; and the tangent of the angle between the regression line and the X-axis indicates whether the relationship is positive or negative.

To obtain an optimum linear model, certain assumptions of regression must be met:

- linearity of the relationship between variables,
- normality and homoscedasticity of errors,
- independence of errors across observations.

For multiple regressions, a fourth assumption must also be met:

- absence of multicolinearity.

In this case, it is also advisable that the number of data for the dependent variable be considerably larger than the number of independent variables.

A regression analysis may yield various models with very similar *determination coefficient* values (R). It is not always easy to determine which of these is the best, i.e. the most practical and reliable. There is no general strategy; rather, selection must be based on *parsimony criteria*:

- the lower the number of independent variables in the model, the more parsimonious it will be;
- since values for meteorological parameters are estimates, forecasting error will have a multiplicative effect;
- a model in which the predictive variables are more readily-measured meteorological parameters, such as temperature, provides greater parsimony;
- variables using cumulative values yield fewer errors, and are therefore recommended.

Stepwise multiple regression analysis is a special application that enables automatic inclusion of variables adding new information, thus improving the determination coefficient. This is a quick, easy-to-use method that obviates the need for prior correlation analysis, since the program itself rejects independent variables not included in the model.

$$Y = a + bX_1 + cX_2 + dX_3 + ... + zX_n$$

Where *a* is the intercept; *b, c*, *d*,... *z*, are coefficients of partial regression; and X_1, X_2, X_3.... X_n are the independent variables.

Neuronal Networks

Neuronal Networks are computational models whose main feature is their ability to learn by example. Thus, when using a neuronal network there is no need to tell it how to obtain output data from given input data; rather, the network is shown examples of the relationship between input and output data, and learns the relationship between the two by

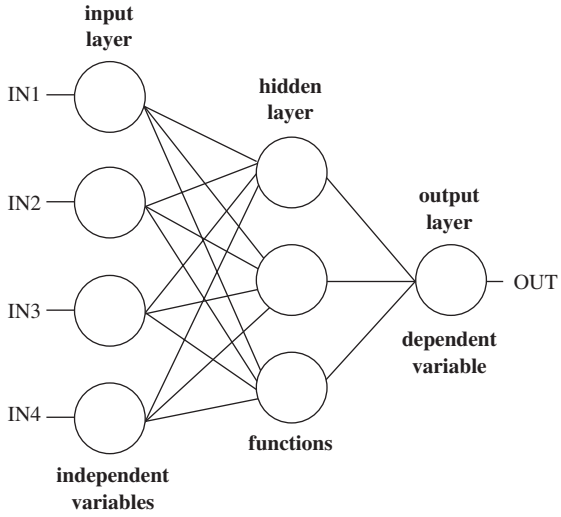

Fig. 2.

means of a learning algorithm. Once the network has been trained to perform the desired function, it can be used, i.e. we can enter input values whose output is unknown, and the neuronal network will calculate the output.

The diagram (fig. 2.) shows a sample neuronal network with 4 input neurons in which input data are entered, 3 hidden neurons which process the information, and 1 output neuron which calculates the output for the given input. All input and output neurons use a transfer function which will differ according to the problem.

There are two types of neuronal networks: feed-forward and recurrent. Feed-forward networks are used when the information to be processed does not involve temporal characteristics such as data order. However, when the problem to be solved involves any sort of time dependence, recurrent neuronal networks are a better option.

The 6th European Course On Basic Aerobiology,
Poznan, Poland
C. Galán