NRES 798 — Statistical Methods for Ecologists
1. Statistics: Introduction and overview

Oscar García

February 13, 2013

Contents

1 Scope

Is has been said that Statistics tends to be used like the drunk uses the lamppost: more for support than for illumination.

The primary function of courses like this is usually to help with the *support* function. Ecologists and other scientists are required to include Statistics in their publications. It tends to follow a ritualized form imposed by reviewers and editors, varying slightly with speciality, and over time with fashion. Statistics is seen as a necessary evil that allows papers to be published and provides some respectability, but little else. Like it or not, this is a reality for professional scientists, and we will try to provide some of the tools of the trade.

Used properly, Statistics can also provide *illumination*, aiding the understanding of situations that involve uncertainty. This may be more for personal consumption than for export. To some extent, we will try to shed some light also, mainly in relation to the whys and limitations of the standard recipes.

In addition, somebody said that the reason to learn economics is to avoid being fooled by economists. Same with statistics, except perhaps that the worst offenders are not usually statisticians

This lecture deals squarely with fundamental concepts. It may seem rather abstract at this stage, but we will keep coming back with specific examples during the course. It is assumed also that you had a first statistics course, from which you can remember some things.

2 Models and Statistics

The human mind is incapable of comprehending the real world in all its complexity. In "reality", assuming that such thing exists, everything is related to everything else. To be able to reason, we use simplified representations, models or theories[1], that include only what are thought to be the most important relationships for the purpose in hand, and ignore the rest. These models may be mental pictures, verbal, mathematical, or of other types. It is crucial to realize that we are always dealing with some kind of model, and not with the "real thing". Failing to make the distinction is the source of most confusion and misunderstandings, in particular when applying Probability and Statistics.

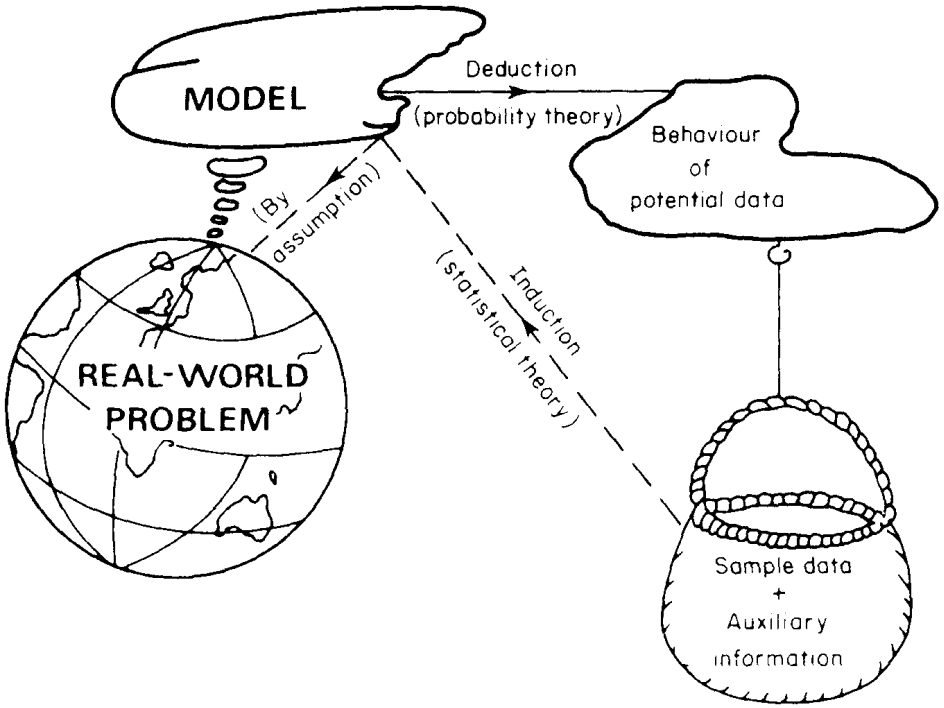
Here we work with mathematical models. Like a mental or verbal model, but expressed in mathematical language. Mathematical notation can be more precise, less ambiguous. More importantly, it allows the recording and re-use of thought processes, in the form of theorems or rules. These can be recycled as building blocks for further deductions, without having to start from scratch every time. The building blocks can be developed by professional mathematicians and used by anybody. Think of musical notation, and how it allowed the storage and re-creation of works written by musicians centuries ago.[2]

Statistics deals with the use of relevant information in situations involving uncertainty. One might distinguish:

1. Descriptive Statistics and Exploratory Data Analysis. Summarizing data, data checking, and searching for patterns.
2. Statistical Inference. Modelling and reaching conclusions. Based on the mathematical Theory of Probability.

[1] Theories are models considered as particularly important, perhaps because of a wider applicability.

[2] For more on models see: `http://www.unbc.ca/assets/nres_graduate_program/nresi_op_06_garcia_2010.pdf`



The left-hand side of the picture[3] depicts the essential aspects of a real-world problem represented in a model. Probability theory tells us about properties of various data that we could obtain. Assuming that the model is "true". Once we obtain specific data, statistical inference can produce statements about certain model properties, and guide decisions and predictions. Important: any conclusions are about the model, not the real world. The application of these results to the original problem requires a leap of faith: there is an implicit assumption that if the model is not too far from the truth, then our conclusions should not be too far from the truth either. Plausible perhaps, but note first that this is rather vague (how do we measure the distances?), and second, that there is no proof of such "continuity" property (a small change in x implies a small change in $f(x)$, remember Calculus?). The justification is that *it usually works*.

[3] Figure 1.2.1 from Barnett, Vic. *Comparative Statistical Inference* — Third Edition, Wiley 1999.

3 Probability

Statistical models differ from other mathematical models in that they include an explicit representation of *uncertainty*, ignorance, or incomplete information. It is assumed that some values are only partially known, and can take values within some set called a *sample space*. The 'random' values may have different 'propensities' of occurring in different parts of the sample space. The propensity for any of the parts is described by a number between 0 and 1, its *probability*. In other words, probability is a function that assigns real numbers between 0 and 1 to subsets of the sample space (named *events*). As is customary in mathematics, there is some ambiguity in the use of language, somewhat confusingly using the same word to refer to a function and to the values that the function takes.

What *is* probability? In the model, it is just a mathematical entity, a function that takes sets into numbers. Probability is a special case of what is known as a *measure*, like the area of parts of the plane that is not restricted to be between 0 and 1. But what does it *really* mean? As always, the connections between the real world and the model are mostly up to the modeller and to the users of the research. There are different views of what randomness might mean or represent. For some, it may be a property of physical objects or processes. In the obligatory example of a coin toss, it would be an intrinsic property of the coin that makes it to fall heads or tails with a certain frequency. This, or something like it, is called the *frequentist* interpretation of probability. For others, or for the same people in other situations, probability may be a purely subjective measure of belief or of our knowledge (or ignorance) about the process, depending on experience and changing over time and from person to person. This is *subjective probability*, of which there is a number of flavours. Or randomness might be intended as a rough representation of the effects of all the interactions that we left out of the model.

In Statistics we almost always deal with sample points (elements of the sample space) that can be represented by one or more numbers, called *random variables*. And probabilities are given through *distribution functions* or through *probability densities* that specify weights for the random variable values. The distribution or density usually depends on one or more unknown *parameters* (in general, a parameter is something that sometimes is treated as a variable, and sometimes as a constant, depending on context).

As a shorthand, it is often convenient to collect individual numbers into a

list or *vector*, and use a single symbol for it (similar to R). For instance, for the random variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, and for the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$. Sometimes, although not always, bold-face or underlining is used to distinguish vectors from simple numbers. A probability density function can then be written as $f_{\boldsymbol{\theta}}(\boldsymbol{x})$, or $f(\boldsymbol{x}; \boldsymbol{\theta})$, or $f(\boldsymbol{x}, \boldsymbol{\theta})$. A Bayesian, as discussed below, might write it as $f(\boldsymbol{x}|\boldsymbol{\theta})$. A vector of random variables can also be called a (vector) random variable, and a vector of parameters called simply a parameter or parameter vector.

A common convention in statistics, that we will not observe right now, is to use capital letters for random variables, seen as functions of sample points, and the corresponding lower-case letter for the function values.

4 Approaches to statistical inference

As shown in the previous figure, a sample of observations is somehow obtained from the population. These data is some list (or table) of numbers that we call $\boldsymbol{x}$. Assuming that the sampling is done "properly", and that the model is "true", $\boldsymbol{x}$ is some (maybe complicated) function of the model random variables, and is therefore also a random variable. Probability theory can produce the density of $\boldsymbol{x}$, which depends also on the unknown parameters:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) \; . \tag{1}$$

This is usually called the statistical model.

We are interested in saying something about $\boldsymbol{\theta}$. For instance, we may want to chose a good estimate $\hat{\boldsymbol{\theta}}$ for use in applications. More specifically, a function of the data, $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$. We may use the same symbol $\hat{\boldsymbol{\theta}}$ for the function, an *estimator*, and its value, an *estimate*.

What is a good estimator? The estimator is a function of the data, which is a random variable, and therefore it is also a random variable. Remember that we are talking about the model, not about the real thing. We would like $\hat{\boldsymbol{\theta}}$ to be close to the true $\boldsymbol{\theta}$, most of the time. We need to be more clear on a couple of things for this to make sense.

First, what do we mean by *close*, and how do we value closeness? In a practical situation the consequences of the error $\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}$ may be different if this is positive or negative, and they may also differ depending on the size of $\boldsymbol{\theta}$. *Decision Theory* represents the consequences by a *loss function*

$L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$, which is 0 if the error is 0, and increases in some way as the values become more different. Now, this is specific to a particular problem and to a particular decision-maker. Science is supposed to provide results for general use, so it is not clear what loss function one should choose. So, somewhat arbitrarily, statistical inferences are usually based on loss functions that are mathematically convenient, and hopefully not too unreasonable. Typical implied loss functions are proportional to $(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^2$, or to $|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|$, or are assumed to be constant for any error different from 0. In any case, we want an estimator that minimizes the (weighted) average or expected loss

$$E[L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})] = \sum_{\boldsymbol{x}} L(\hat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta}) f(\boldsymbol{x}, \boldsymbol{\theta}) \tag{2}$$

(substitute an integral for the sum if $\boldsymbol{x}$ is not discrete).

Second, in general the expected loss (2) depends on the true value $\boldsymbol{\theta}$, which is unknown. We would like it to be small for typical values of $\boldsymbol{\theta}$, averaging according to how plausible different $\boldsymbol{\theta}$ might be. This is easy (in principle) with a subjective view of probability, in the *Bayesian* approach. We have then a *prior* probability density $p(\boldsymbol{\theta})$ that expresses our guess or feeling about the most plausible values, based on prior knowledge, experience, or whatever. We "simply" choose the estimator that minimizes the average expected loss, weighted by the prior (a double expectation).

Or, to put it in a slightly different way, for the Bayesian $\boldsymbol{\theta}$ is a random variable, and therefore the statistical model is a conditional distribution $f(\boldsymbol{x}|\boldsymbol{\theta})$. Through Bayes theorem, this and the prior determine the *posterior* probability density

$$f(\boldsymbol{x}|\boldsymbol{\theta}) \ , \quad p(\boldsymbol{\theta}) \quad \rightarrow \quad g(\boldsymbol{\theta}|\boldsymbol{x}) \ ,$$

describing the probability of $\boldsymbol{\theta}$ given the observations $\boldsymbol{x}$ (hence the name *Bayesian*). Then the mean, median, or mode of the posterior is taken as $\hat{\boldsymbol{\theta}}$. It is found that the mean, median and mode minimize the expected quadratic, absolute, and constant losses, respectively. If the posterior happens to be symmetric, then the three values coincide, and we do not need to agonize over which one to use.

This does not work with the frequentist view of probability, used in what is called Classical Statistics. For the Bayesian, the "true" height of tree number 42 has a prior subjective probability. For the frequentist, that height is a fixed unknown number, it does not make sense to assign a probability to it. Therefore, the decision theorist can define an optimal estimator, and the

Bayesian a "sort of" optimal or approximation to it. But classical statistics has to conform itself with estimators and other statistical procedures that, at best, are good according to certain more or less arbitrary criteria such as consistency, unbiasedness, efficiency, invariance, etc. Or are best within some restricted class, such as estimators that are linear functions of the data. Equation (1), seen now as a function of $\boldsymbol{\theta}$ for the given data $\boldsymbol{x}$, is known as the *likelihood function*, and plays a fundamental role in most statistical procedures.

We may summarize like this:

Approach	Aim	Probability	Information
Decision Theory	Decision-making	Subjective	Data, prior, losses
Bayesian	Inference	Subjective	Data, prior
Classical	Inference	Frequentist	Data

All this is hugely controversial. Bayesian ideas were fringe stuff for many years, but recently have become fashionable and more respectable, vociferously promoted by new converts. Classical Statistics is still the bread and butter for most journal publications, and we will focus mostly on that, but there is also material about Bayesian methods throughout the textbook. The Bayesian approach is generally more intuitive and closer to how most people think; classical arguments tend to be more convoluted. On the other hand, whose prior should be used? It is often said that science is not supposed to make decisions, that its role is to provide "evidence". It should be "objective", so that personal priors or loss functions have no place in it. On the other hand, the choice of model is subjective...

Sometimes Bayesian methods use *uninformative* (pspeudo-)priors, like $p(\boldsymbol{\theta}) =$ constant (not really a density). Then Bayesian and classical results may coincide, and there is no conflict.

A pragmatic view could be that in a scientific study classical statistics provides an incomplete analysis, summarizing the evidence into estimates and standard errors (or other similar forms). It is then up to the reader to informally combine this with her priors and loss functions in order to reach conclusions or make decisions.

There are other topics/approaches, of which here we only mention a couple: Non-parametric Statistics uses models that are not characterized by parameters. It avoids some assumptions, at the cost of weaker inferences. Survey

Sampling (designed-based), and parts of Experimental Design, use a completely different inferential logic, based on repeated occurrences. And there is Descriptive Statistics, and exploratory data analysis (EDA), that will be seen mainly in the labs.

5 Classical statistical inference

Statistics, in its current form, is relatively recent. Amazingly, most of the fundamentals were developed almost single-handedly by Sis Ronald A. Fisher roughly between 1920 and 1935. Above is a photo of Sir Ronald at a seminar organized by the US Forest Service in 1936. Many of the big names in Forest Mensuration are there, foresters were among the first in adopting the new ideas; it has been all down-hill ever since `:-)`.

Classical statistics[4] focuses largely on two main topics: parameter estima-

[4] The Swedish statistician H. Crámer, in his famous 1949 book, refers to Bayesian methods as the "classical approach". The new statistics was developed in response to the

tion, and tests of hypotheses. It works more or less as in this example:

Data: 10 measurements of soil pH. Model: the measurements have a normal distribution with mean μ and standard deviation σ (parameters). Obtain estimators (functions of the measurements) for μ and σ. Use these values for inferences (statements, conclusions) about the parameters (estimation, hypothesis tests). These are based on the probability distribution of the estimators (aka *sampling distribution*).

More generally, as already anticipated above: Data $\boldsymbol{x} = (x_1, \ldots, x_n)$. Parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. Use a statistic[5] or estimator $g(\boldsymbol{x})$ to say something about $\boldsymbol{\theta}$ (estimation: value of $\boldsymbol{\theta}$; hypothesis testing: is a certain value plausible?). Based on the sampling distribution of $g(\boldsymbol{x})$.

In *point estimation* the idea is to later use the estimate as if it were the real value. Some measure of precision based on the sampling distribution may also be given. There is also *interval estimation*, where one estimates bounds for the parameter, in the form of a *confidence region* that contains the parameter with a given probability. Important: it is the region that is random, the parameter is unknown but fixed! Estimators may fulfil criteria such as consistency, unbiasedness, efficiency, etc., and are found by methods such as least-squares, maximum-likelihood, etc.

Hypothesis tests follow the Neyman-Person recipe: Set up a working hypothesis (*null hypothesis*) H_0, alternative to the hypothesis H in which we are interested. Choose a statistic, and calculate its probability if H_0 were true. If the probability for the observed value is less than some magic number such as 0.05 (*significance level*), then "reject" H_0; it is unlikely that the data suggesting H would be obtained by pure chance. Otherwise, "accept" H_0, the test is inconclusive ("more research is needed"). Although the accept/reject language may suggest decision-making, it is not supposed to.

The theory assumes that the model and the hypotheses are formulated without looking at the data, they are supposed to be independent. Remember the "Scientific Method"? Hypothesis $\rightarrow$ testing $\rightarrow$ modified hypothesis $\rightarrow$ testing $\rightarrow \ldots$. Clearly, the assumption of independence fails after the first iteration, unless one uses a completely new data set each time. Same when

perceived deficiencies of that approach in dealing with scientific problems. It is ironic that now Bayesian methods are often presented as a response to the perceived deficiencies of classical statistics.

[5] A *statistic* is any function of the data that does not depend on the parameters. An estimator is a statistic intended to substitute for a parameter.

a model is formulated after an exploratory data analysis. Usually this is cheerfully ignored.

6 Comments and conclusions

So, what's the difference? What most confuses people may be that classical statistical inference does not deal with a specific data set, it deals with the (random) properties of a statistical procedure. Once the data is obtained and the calculations are done, "the dice are cast": the interval either contains the parameter or it does not, the estimate has a certain unknown but fixed error, etc. Descriptive statistics and data analysis deal with a specific data set. Bayesian inference describes degrees of belief *after* the data is observed. The prior belief is updated according to the observations, the results are valid for people with the same prior. The updating often uses the same machinery from classical Statistics, and if the prior is "vague" the numerical results are often the same; the interpretation is different.

Always remember the immortal words of G.E.P. Box: *All models are wrong, but some are useful.* Modelling belongs to the Art part of Science, same as judging results in terms of practical consequences. In-between we can be "objective".

Some healthy disrespect for Statistics might not be a bad thing. But it must be based on knowledge, and do not over-do it! Statistics can be a very useful aid to common sense, although not a substitute. Misuse is common, a little knowledge can be dangerous. It is your professional responsibility to keep your eyes open to this.