Snapshots of modern mathematics
from Oberwolfach

№ 6/2014

Statistics and dynamical phenomena

Howell Tong

A friend of mine, an expert in statistical genomics, told me the following story: At a dinner party, an attractive lady asked him, “What do you do for a living?” He replied, “I model.” As my friend is a handsome man, the lady did not question his statement and continued, “What do you model?” “Genes.” She then looked at him up and down and said, “Mh, you must be very much in demand.” “Yes, very much so, especially after I helped discover a new culprit gene for a common childhood disease.” The lady looked puzzled.

In this snapshot, I will give you an insight into Statistics, the field that fascinated my friend (and myself) so much. I will concentrate on phenomena that change over time, in other words, dynamical events.

1 Chaos and autoregressive models

Consider the following *doubling map*, where n denotes an integer: We choose $y_0 > 0$, and define a series whose n^{th} element is given by

$$y_n = 2y_{n-1} \mod 1. \tag{1}$$

In words this means that we double the previous y-value, throw away the integer part and keep only the decimal part[1]. The map is also called a saw-tooth map because it is equivalent to the iteration defined by the saw-tooth function

$$f(y) = \begin{cases} 2y, & 0 \leq y < 0.5 \\ 2y - 1, & 0.5 \leq y < 1. \end{cases} \tag{2}$$

Plotting the graph of the function shows that it contains two straight lines with a break (threshold) at 0.5. Such a function is nonlinear, although the single pieces are linear.

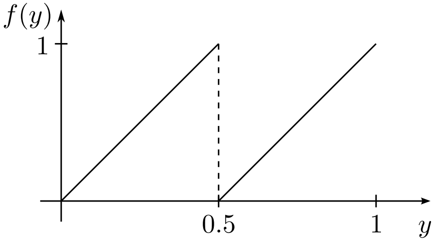

Figure 1: A plot of the saw-tooth function with threshold at 0.5

Now, select any positive number as a starting point and iterate forward to get

$$y_1, y_2 = f(y_1), y_3 = f(y_2), \ldots.$$

We will quickly realize that the output shows no regularity no matter how long we run it, that is, the output looks quite chaotic (hence we need Chaos Theory), almost indistinguishable from a random series. This seemingly puzzling feature of generating randomness from a wholly deterministic mechanism can be understood as soon as we realize that the function f is highly sensitive to initial values: Two initial values differing only in, say, the 8th decimal place will quickly diverge (that is, spread apart) upon repeated application of f. The famous

[1] For example if $y_0 = 0.7$, we get $y_1 = 2y_0 - 1 = 0.4$.

French mathematician Henry Poincaré (1854-1912) listed such sensitivity as one of the sources of randomness. Another curious thing about the doubling (or saw tooth) map is that if we think of n as a parameter for time and run the map "backwards" in time, we will discover that we need to introduce "external randomness" denoted by ε_n. We set

$$X_n = 0.5 \cdot X_{n-1} + \varepsilon_n, \tag{3}$$

where n is an integer, ε_n equals 0 and 0.5 with equal probability, and the connection to the series of the y_n given above is $X_n = y_{-n}$. The ε_n term appears because the inverse of f maps one point to two possible points equally likely. For example we get $f^{-1}(0.5) = 0.25$ or 0.75.
This leads us to yet another curious feature: We started with a *deterministic*[2] nonlinear map (2) and end up with a random (a more fancy word is *stochastic*) linear equation (3).

An equation like (3) defines what is called a *time series model* in Statistics, and it describes the dynamics of a system over discrete time (n). In general, the ε_n can have a more general probability distribution than the uniform distribution from our example, such as the Gaussian distribution (after the famous German mathematical genius Carl Friedrich Gauß (1977-1855)) over the set of real numbers. An equation of the form

$$X_n = a_1 \cdot X_{n-1} + \ldots + a_p \cdot X_{n-p} + \varepsilon_n \tag{4}$$

is called a *(linear) autoregressive model* (AR model for short), which was invented by the British statistician Udny Yule (1871-1951) in 1927 when he studied the annual number of sunspots. Here, the coefficients a_j are the defining parameters to be estimated from observations.

The AR model finds appliciations in a broad variety of fields both in its original form and its multidimensional or piecewise linear generalizations. We will illustrate this by considering some real-life exmaples.

2 Some real examples

- **US hog data**
  Professors George Box and George Tiao from the USA analyzed the data seen in Figure 2 by using a 5-dimesional AR model. They used the variables H_p, H_s, R_p, R_s, W for the hog price, hog supply, corn price, corn supply and

[2] *Deterministic* means that each time we plug in the same value into our function, it gives us the same result. The opposite of deterministic is *stochastic*, such as seen in equation (3): The input 0.5 gives 0.25 (for $\varepsilon = 0$) or 0.75 (for $\varepsilon = 0.5$) with equal probability.

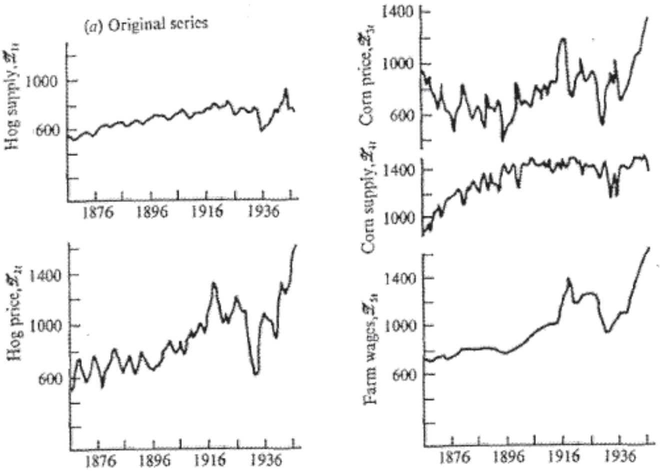

Figure 2: Hog data from the US (Box and Tiao, 1977)

the farmer's wage respectively. They concluded that

$$\frac{H_p H_s}{(R_p R_s)^{0.75} W^{0.50}}$$

is approximately independently distributed about a fixed mean. In other words, using real-life observations, they have justified that

$$\frac{\text{profit of the farmer}}{\text{expense of the farmer}}$$

follows a stable economic law! It is a remarkable historical fact that such a simple empirical relationship had never been discovered for the above classic data set, until the two statisticians invented a new time series method based on another statistical technique called canonical correlations in 1977.

- **Economics and Finance**

  The piecewise linear deterministic model (2) has its stochastic equivalent by adding ε_n. The result is commonly called the *threshold AR* (or *TAR*) model first introduced by the author in 1978. In an even more general form, it looks like

$$X_n = \begin{cases} a_0 + a_1 X_{n-1} + \ldots + a_p X_{n-p} + \varepsilon_n, & X_{n-d} < r \\ b_0 + b_1 X_{n-1} + \ldots + b_p X_{n-q} + c\varepsilon_n, & X_{n-d} \geq r. \end{cases} \tag{5}$$

The variable r is called the threshold parameter and has to be estimated from the observed data, along with the a_j, b_j, p, q, d, c, and the variance of ε_n.

The model has made an enormous impact in economics and finance.
In 2011, the US econometrician Professor Bruce Hansen published a comprehensive review of 75 papers related to TAR models. A typical example for the application of the TAR model is the modelling of aggregate output as measured by GNP growth rates. It has for instance been shown that the GNP of the United States is subject to floor and ceiling effects. The use of the TAR model to study the relationship between long and short interest rates has helped to reveal the strong asymmetric response of interest rate changes to the spread between the long and short rates.

- **Plagues in Kazakhstan**
Early this century, the statistician Kung-Sik Chan and his then doctoral student, Noelle Samia, worked with a team of biologists/epidemiologists led by Professor Nils Chr. Stenseth, formerly President of the Norwegian Academy of Science and Letters, on an extensive scale study of plague epidemics in Central Asia. Concerning the bacterium Yersinia pestis that causes bubonic plague, they concluded in their report in the Proceedings of the National Academy of Sciences (USA) in 2006 that

> *"Y. pestis prevalence in gerbils increases with warmer springs and wetter summers... Climatic conditions favouring plague apparently existed in Central Asia at the onset of the Black Death as well as when the most recent plague pandemic arose in the same region, and they are expected to continue or become more favorable as a result of climate change."*

This conclusion is based on a modified TAR model which the participating statisticians developed and fitted from observed data. In the model to be given below, $N_{t,\ell}$ is the number of great gerbils examined at time t in "large square"[3] ℓ. The number of great gerbils testing positive under a bacteriological test does not follow a Gaussian distribution. Instead, it is more likely to follow a binomial distribution and hence we modify the TAR model slightly while retaining the piecewise linear structure.
We distiguish between data collected in the spring and in the fall and mark their parameters with ^{s} and ^{f} respectively. The binomial distribution model has parameters $(N_{t,\ell}, P_{t,\ell})$, where if t is a spring, the true prevalence rate is $P_{t,\ell} = 0$ when the lag-d^s occupancy (i.e. how many burrows are occupied by the great gerbils d^s units of time ago, namely $X_{t-d^s,\ell}$), is below the spring

[3] The large squares are the result of dividing Kazakhstan into non-overlapping $40 \times 40\,\mathrm{km}^2$ squares.

threshold r^s_ℓ. Otherwise, it follows a logistic regression model shown below. A similar specification applies to the fall data. We should not worry about too many details as long as we get the message that the dynamics underlying bacterium Yersinia pestis operates in 4 regimes depending on the season (spring or fall) and the lag-d^s occupancy rate (above or below the seasonal threshold).

$$P_{t,\ell} = \begin{cases} \begin{cases} 0, & \text{if } X_{t-d^s,\ell} < r^s_\ell \text{ and } t \text{ is a spring} \\ \text{logit}^{-1}\{(\beta^s_0 + b^s_{0,\ell}) + (\beta^s_1 + b^s_{1,\ell})T_{sp,t} + b^s_{2,\ell}R_{sp,t} + \epsilon_{t,\ell}\}, & \\ & \text{if } X_{t-d^s,\ell} \geq r^s_\ell \text{ and } t \text{ is a spring;} \end{cases} \\ \begin{cases} 0, & \text{if } X_{t-d^f,\ell} < r^f_\ell \text{ and } t \text{ is a fall} \\ \text{logit}^{-1}\{(\beta^f_0 + b^f_{0,\ell}) + \beta^f_1 R_{su,t} + \beta^f_2 X_{t-1/2,\ell} + \epsilon_{t,\ell}\}, & \\ & \text{if } X_{t-d^f,\ell} \geq r^f_\ell \text{ and } t \text{ is a fall;} \end{cases} \end{cases}$$

Here, the superscript f signifies fall, X denotes the great gerbil occupancy, $T_{sp,t}$ is the spring temperature, $R_{sp,t}$ is the log spring rainfall, and $R_{su,t}$ is the log summer rainfall. The βs are tuning parameters to be estimated from the observed data. The logistic function $\text{logit}^{-1}(x) = \frac{1}{1+\exp(-x)}$ is just a mathematical trick to transform the binomial model setting into a TAR model format.

3 Looking to the future

The present information age poses many exciting challenges to Statistics. Data collection is so fast and plentiful that suddenly we find ourselves flooded with data.

Let us start with a simple illustration: We saw the example of the hog data, taken at one particular site. Suppose we have similar data at say 100 sites. Does a similar empirical law hold for all of them? Will there be differences? Similar panel time series can and do occur in many situations, e.g. a panel of stock price time series across different stock markets, a panel of death rates due to a particular infectious disease at different locations in the world, and so on. Many interesting questions then need to be answered, e.g. are the different stock markets equally volatile or do they cluster in some way? Is the infectious disease spreading?

It is clear that new statistical methodologies will be necessary in order to cope with the new challenges. This is where fresh thinking is needed. To do so, we may have to re-examine existing statistical methodologies and even our philosophy. Statistics as a scientific discipline has a rather recent history, not much longer than 100 years. Over these 100 years, the discipline has been dominated by a concept called likelihood. It is based on the assumption that the real underlying model is known and the only thing unknown are its tuning parameters. The observed data then enables us to estimate them. Two schools of thought, the frequentist school and the Bayesian school, have been the pillars of Statistics. Despite their sometimes heated and colourful polemics, they share the common ground of likelihood.

But what if the true model does not exist? What if we know that the model given to us by our scientist friend is wrong but it is the best available? I think these are legitimate questions that we need to address. Although some opening shots have been fired by people like Professor Laurie Davies in Germany, Professor Yingcun Xia in Singapore, the author and others, the main contributions will have to come from younger brains like yours!

Image credits

Fig. 2: U.S. hog data, original series. Cited from [1]

References

[1] G. E. P. Box and G. C. Tiao, *A canonical analysis of multiple time series*, Biometrika **64** (1977), no. 2, 355–365, http://biomet.oxfordjournals.org/content/64/2/355.abstract.

[2] K.S. Chan and H. Tong, *Chaos: A statistical perspective*, Springer Series in Statistics, Springer, 2001, ISBN 9780387952802, http://books.google.de/books?id=DurlS0WgVZEC.

[3] P.L. Davies, *Approximating data*, Journal of the Korean Statistical Society **37** (2008), no. 3, 191 – 211, ISSN 1226-3192, http://www.sciencedirect.com/science/article/pii/S1226319208000380.

[4] B. Hansen, *Threshold autoregression in economics*, Statistics and Its Interface **4** (2011), 123 – 128.

[5] N.C. Stenseth, N.I. Samia, H. Viljugrein, K.L. Kausrud, M. Begon, S. Davis, H. Leirs, V.M. Dubyanskiy, J. Esper, V.S. Ageyev, N.L. Klassovskiy, S.B. Pole, and K.S. Chan, *Plague dynamics are driven by climate variation.*, Proceedings of the National Academy of Sciences of the United States of America **103** (2006-08-29 00:00:00.0), no. 5, 13110–5, ISSN 0027-8424.

[6] H. Tong, *On a threshold model*, Pattern recognition and signal processing (C.H.Chen, ed.), NATO ASI Series E: Applied Sc., no. 29, Springer, 1978, pp. 577 – 586.

[7] Yingcun Xia and Howell Tong, *Feature matching in time series modeling*, Statistical Science **26** (2011), no. 1, 21–46, http://dx.doi.org/10.1214/10-STS345.

Howell Tong *is professor emeritus for Statistics at the London School of Economics.*

Mathematical subjects
Probability Theory and Statistics

Connections to other fields
Physics

DOI
10.14760/SNAP-2014-006-EN

Snapshots of modern mathematics from Oberwolfach are written by participants in the scientific program of the Mathematisches Forschungsinstitut Oberwolfach (MFO). The snapshot project is designed to promote the understanding and appreciation of modern mathematics and mathematical research in the general public worldwide. It is part of the mathematics communication project "Oberwolfach meets IMAGINARY" funded by the Klaus Tschira Foundation and the Oberwolfach Foundation. All snapshots can be found on www.imaginary.org and on www.mfo.de/snapshots.

Mathematisches Forschungsinstitut
Oberwolfach gGmbH
Schwarzwaldstr. 9 – 11
77709 Oberwolfach
Germany

Mathematisches Forschungsinstitut Oberwolfach
Member of the Leibniz Association
Klaus Tschira Stiftung gemeinnützige GmbH
KTS
oberwolfach FOUNDATION
IMAGINARY open mathematics