What are today's $\bar{x}$ and s? or Happiness is a new data type

By David R. Brillinger

The sample mean $\bar{x}$ and standard error s have long been computed for data sets of numbers as convenient summary values providing measures of location and scale. With the use of $\bar{x}$ to share restaurant bills equally, it must be the most often employed statistical tool - by far. In contrast s isn't so popular, but it does have a clear presence in the $(\bar{x}, s)$ button on, even elementary, handheld calculators. With today's many stimulating data sets it seems pertinent to wonder just what are the contemporary $\bar{x}$ and s.

Many of these data sets correspond to random or stochastic processes. Of such processes Jerzy Neyman, this Department's founder, once wrote, [5]:

"Currently in the period of dynamic indeterminism in science, there is hardly a serious piece of research which, if treated realistically, does not involve operations on stochastic processes. The time has arrived for the theory of stochastic processes to become an item of usual equipment of every applied statistician."

Having reminded the reader of Professor Neyman's remark, it can be argued that we are now in the era of random process data analysis.

The first random process, as opposed to random variable, studied in any detail is surely the time series. For example, following a study of newspapers and magazines, Tufte, [7], remarked:

"The time series plot is the most frequently used form of graphic design."

So there is a type of random process data that abounds in the world at large. Neuroscientists and exploration seismologists have long had a time series analog of $\bar{x}$. It is what neuroscientists call the average evoked response. As the name suggests a stimulus is applied a number of times and then the responses at lag u time units, after applying the stimulus, are averaged. The average evoked response is the result, $\bar{x}(u)$, graphed as a function of u. A special computer, the ARC (average response computer), was even developed for doing this.

Time series remains an area of active research, but attention has turned to other random processes. Just what is a random process? It is simply a family of random variables or chance quantities. However today's usage often seems to have in mind that the index labeling the family has something extra, like an order or a toplogy or an algebraic structure. Often the index refers to time or space or both.

Under stimulus from physics, engineering, medicine and other fields, basic data elements of concern now include graphs, trees, tesselations, shapes, regions, cellular structures, and other mathematical objects. Dependence is basic and the dimensions are often much increased. Parameters too have taken on more complex forms, eg. curves or measures instead of simply numbers or vectors. One amusing aspect is that early results in estimation and approximation, that some put down at the time as excessive abstraction, are now basic to practice. (Professor LeCam wins!)

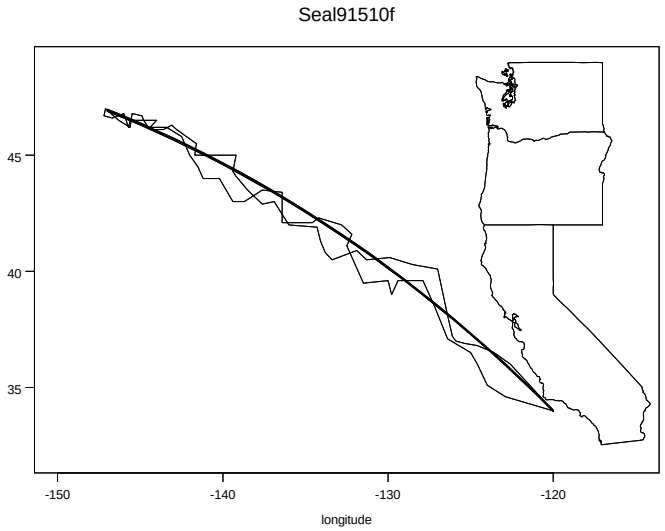
Today's move to the empirical analysis of random processes is made possible in part through the development of useful probability models for unusual mathematical structures and the availability of new devices and ideas for the storage and manipulation of random process data sets. Below an example will be presented of studying a random curve on the surface of a sphere.

Problems continue to be the classical ones of: summarization, model assessment, display, including explanatories, efficiency, robust variants, sanctioning and missing values. New ones arising include: How to get experimental output of special form into a computer for analysis? How to display and make available unusual data types? In connection with this last there are now appearing videos posted on World Wide Web sites, see for example the animation of United States weekly AIDS mortality rate by county, [1] and the applet allowing the study of chip failures as a function of location on wafers in integrated circuit manufacturing, [4]. In the search for insight and unusual occurrences (eg. outliers), descriptive statistics have returned as a force. Tools employed include: laptops, stochastic models, differential equations and complex algorithms. Commonly used data forms are music CDs, images and videos. Special computer data storage methods have been developed for sounds and images in particular.

How does one get involved with all this? A simple way is to take some object and make it random, or make it dynamic or make its values more “abstract”. Journals to look at include: *IEEE Trans. Bio-Medical Engineering, Annals of Applied Probability, SIAM News, JASA Applications Section, IEEE Signal Processing Magazine, Symposium on the Interface of Computer Science and Statistics.*

Because of the difficulty of the problems being studied and the massive amount of subject matter, there is a basic need for collaboration with scientists from other fields. Next follows a brief description of some personal work with a marine biologist (and lawyer!), Brent Stewart, of Hubbs Sea World Research Institute, San Diego.

Twice a year elephant seals set off from along the California coast on lengthy migrations northwest into the Pacific. Sensors are attached and data for journeys become available on the animals' return. The figure below shows one migration taking 75 days. The points plotted are estimated midday positions of the animal and these are the basic data for an analysis. As the figure shows the migrations can cover thousands of kilometers. A surprise is that this animal seemed to have in mind both an inbound and an outbound destination. An initial step of analytically describing such migratory trajectories might help to develop some understanding of this circumstance. One possibility is that the seal is approximately following a great circle route, the shortest route between two points on the surface of a sphere. A great circle path has been included on the figure as a reference. The navigational mechanisms such an animal might employ are as yet unknown, but a great circle route would imply that the animals are able to assess their position relative to some astronomical or global magnetic background and constantly make course corrections. As the elephant seals dive and forage continuously while migrating, there is a need for course corrections.



How might such data be summarized? What might $\bar{x}$ and *s* be here? One approach is to build a stochastic model and thereby be led to pertinent statistics. Historically, paths of particles have been described by differential equations, and indeed the old name for a realization of a time series is a trajectory. In the present case, due to foraging, currents and other unmeasured local effects on the animal it seems pertinent to include random elements in such equations. The stochastic differential equations for a particle wandering randomly on the surface of a sphere were set down by Perrin, [6]. In the case of an animal migrating, a drift term is introduced and Perrin's equations become

$$d\theta_t = (\frac{\sigma^2}{2\tan\theta_t} - \delta)dt + \sigma dU_t$$

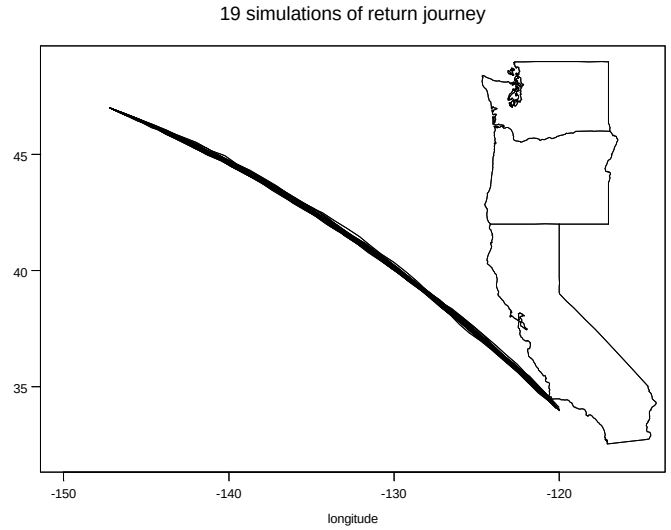
$$d\phi_t = \frac{\sigma}{\sin\theta_t} dV_t$$

where θ and ϕ are latitude and longitude in a coordinate system with the destination the North Pole. For the seals the initial destination may be some point at sea while the return destination is the starting point on the California coast. Here δ is the velocity towards the destination, σ is meant to represent foraging and other variability around a direct heading and *U*, *V* are independent random walks. For example the expected time to get from a distance *x* from the destination to a distance *d* is now a nontractable double integral, not the simple $(x-d)/\delta$ of the noise-free case.

Other aspects of the problem to take account of include: the positions are available only at discrete time points, there are some missing values and there is measurement error. With some further assumptions a likelihood function can be set down and the parameters estimated. The analysis turns up amongst other things that the error of location appears to be more substantial than the foraging variability, σ. This will become apparent in a comparison of the figure above with the figure to be presented below.

So these days there are a variety of novel data types and one can wonder what are today's $\bar{x}$ and *s*? In various circumstances they could be a parameter estimate and a corresponding uncertainty. A candidate for $\bar{x}$ is Cleveland's, [?], loess() function, which has now reached the status of a clickable button in Splus 4.0 . Loess was used to estimate

smooth paths for the seal journeys. As an analog of *s* I propose the figure below. It displays the results of 19 simulations of the process, given by the stochastic differential equations above, employing estimates of δ,σ. This figure gives an indication of the variability of the basic process. The measurement error estimate has not been included in the simulations. It plays another role.



Random process techniques were basic in approaching this problem. A number of my colleagues work on such processes. We can wonder what natural discoveries Aldous's clumps, Evans' local fields, Perez's tree indices and Pitman's windings are leading to. There will surely be some.

[1] AIDS Data Animation Project. http://infoserver.ciesin.org/datasets/cdc-nci/aids.html
[2] Brillinger, D. R. and Stewart, B. S. (1998). Canadian J. Statistics. (Available at http://www.stat.berkeley.edu/~brill).
[3] Cleveland, W. S. (1979). J. Amer. Statist. Assoc. 74, 829-836.
[4] Hansen, M. (1997). http://cm.bell-labs.com/cm/ms/who/cocteau/java/waferApplet/waferApplet.html
[5] Neyman, J. (1960). J. Amer. Statist. Assoc. 55, 625-639.
[6] Perrin, M.F. (1928). Mouvement brownien de rotation. Ann. l'École Norm. Sup. 45, 1-51.
[7] Tufte, E. R. (1983). The Visual Display of Quantitative Information. Graphics Press, Conn.

Postscript. Here's a project for the probabilist readers: develop some properties of a tied down random walk on the sphere, both with and without drift.