Numeracy for Everyone

2. Dice of Life

Anil P Gore and S A Paranjpe

A P Gore is currently Head, Department of Statistics, University of Pune. He initiated teaching and research in ecological statistics at Pune. He has been active for many years in promoting interaction between statisticians and biologists.

S A Paranjpe is Reader in the Department of Statistics, University of Pune. She has long experience of applying statistics in the fields of nutrition, forestry, ecology, agriculture and health.

Lottery tickets are one of the hottest items on sale in India. Common people love to buy dreams of becoming millionaires overnight for a handful of rupees. All living beings play lotteries with nature. At stake are their own lives or their offsprings' lives.

Gambling begins right at conception. In sexually reproducing species that are diploid, an offspring gets one chromosome out of every pair of homologous chromosomes from a parent. Which one? That is a lottery.

In human beings mothers have a pair of X chromosomes of which one is received by a son. Haemophilia is a disease in which blood outflow from a wound does not stop on its own. If a mother is a carrier of haemophilia i.e. has a haemophilia causing gene on one of the two X chromosomes, what is the chance that her son will get it and become haemophilic? 50%. Genetic disorder of colour blindness is a similar situation.

Once born, an organism faces an uncertain environment. Where will it find food? Will it be sufficient? Should a bird in the hand always be preferred to two in the bush? Should an animal forage together with other conspecifics or should it be a loner? Life is an unending string of decisions. Some prove to be right and others wrong. Together they decide the 'fitness' of the organism. This is the evolutionary paradigm.

Now let us try to understand this business of uncertainty and chance. What is chance? Well, the chance of an event (say success in a venture) is a number between zero and unity. If the chance (or probability) of something is zero it means that event is practically impossible. If the chance is one, it means that the event is certain. What is a chance of 50%? It means that in a long series of

repetitions of an experiment, success is expected 50% of the times.

On the face of it, the phrase 'laws of chance' seems like a contradiction. Chance by definition is something that defies all rules. So how can there be laws of chance? The doubt and concern are legitimate but misplaced. Whether a particular conception will lead to a male or a female baby is indeed very difficult to guess. We presume to guess only the proportion of male births in a large set. When we discuss chance, we are not talking about a particular case but about a population. So the statement that the chance of a male birth is 50% means that in a large number of records the proportion of males should be close to 1/2.

Given this background what is the chance of getting two sons in a row? The laws of chance say that the answer is multiplicative i.e. $1/2 \times 1/2 = 1/4$. There are 4 possible outcomes : son-son, son-daughter, daughter-son, daughter-daughter (all equally likely), and we are excluding the last three. It is because of this multiplicative rule that on the race course, the correct guess of outcomes of several races becomes very difficult and hence it is entitled to the Jackpot. If on the other hand you bet that a horse will be at least in the second place, you are right if he is first or second. Here the chances get added, and your bet usually fetches lower rewards. If you toss a six faced die, the possible scores are 1, 2, 3, 4, 5, 6. The chance of any particular score is 1/6. But the chance of getting an even number (i.e. 2 or 4 or 6) is 3/6. It is the sum of the chances of getting 2, 4 and 6. This is the addition law.

For insurance companies, understanding of chance is a matter of survival. Consider a motorcycle costing Rs. 10,000/-. It is to be insured against theft or loss. Suppose the chance of that is 1 in 1000. That means, out of 1000 vehicles insured, one will be lost (on an average) and the insurance company will have to pay Rs. 10,000/- to the insurer. So to recover this, the company will have to charge each of the 1000 insurers a premium of at least Rs. 10/-. Anything above Rs. 10 will go to cover expenses of the company and profit. If the calculation of the chance of theft goes wrong and many vehicles are stolen, the company will have to

Part 1. Why Quantification?, *Resonance*, Vol.4, No.5, 19–30, 1999.

compensate all and its profits will plummet.

A foraging bird must constantly consider trade-offs between risks and gains. The saying 'no risk no gain' applies even for a merrily chirping bird. If the bird devotes itself single mindedly to eating seeds on the ground, it runs a high risk of being captured by a predator. So it must scan the surroundings and fly away if any danger becomes apparent. If the bird wants complete safety it must remain vigilant at every moment. But then it will die of hunger. So do all birds take a course on statistics and calculation of probabilities? Of course not. Natural selection eliminates birds which make bad choices and we are presumably left with those which make correct choices. So quantitative research in evolutionary ecology involves probability calculations to predict which type of bird should thrive in which type of environment.

We generally agree that the chance of the next human birth to be a male is 1/2. Consider families with 8 children. What is the chance that all are males? Using the multiplication law we get the answer $1/2 \times 1/2 \times \ldots$ 8 times i.e. $(1/2)^8 = 1/256$. The chance of all 8 females is the same. For other compositions the values are:

Males	8	7	6	5	4	3	2	1	0
Chance	$\frac{1}{256}$	$\frac{8}{256}$	$\frac{28}{256}$	$\frac{56}{256}$	$\frac{70}{256}$	$\frac{56}{256}$	$\frac{28}{256}$	$\frac{8}{256}$	$\frac{1}{256}$

This is an example of the so called binomial distribution. Data on 53,680 families in England collected around 1920 had the proportions [1] given below.

Are these observed proportions close to what the chance calculations show? (This can be checked using a Chi-square test which is briefly discussed later). The answer is no. When a model does not

Observed proportions

Males	8	7	6	5	4	3	2	1	0
Chance	$\frac{342}{53{,}680}$	$\frac{2{,}092}{53{,}680}$	$\frac{6{,}678}{53{,}680}$	$\frac{11{,}929}{53{,}680}$	$\frac{14{,}959}{53{,}680}$	$\frac{10{,}649}{53{,}680}$	$\frac{5{,}331}{53{,}680}$	$\frac{1{,}485}{53{,}680}$	$\frac{215}{53{,}680}$

match with data, we discard or modify the model. The data remains supreme. In this case the modification possible is that the chance of a male birth is not the same for all couples. Some couples are male prone and some are female prone. For exploring reasons of this kind one would have to go into details of human reproductive physiology.

Here is another example, this time from plant biology. In the species *Lablab niger* each flower contains up to 5 ovules. Suppose the chance that an ovule gets fertilised and becomes a seed is 0.7. Then out of 5 ovules how many seeds should we expect to get? How many seeds would there be in a pod of *Lablab niger*? Actually, there are no certainties. All ovules could fail (no. of seeds = 0) or all could succeed (no. of seeds = 5). A proper use of addition and multiplication laws mentioned above gives the chances of various numbers of seeds as given below.

No. of seeds	1	2	3	4	5
Chance	.0261	.1275	.3058	.3652	.1754

The case of zero seeds is not given because it is not observable. This is an example of what is called a (zero truncated) binomial distribution [2]. Two scientists from Bangalore, R Umaashankar and K N Ganeshaiah checked pods and found the following proportions:

No. of seeds	1	2	3	4	5
Proportion	1/69	4/69	22/69	41/69	1/69

In this case the observed proportions are quite close to the theoretical ones, after taking into account the fact that not all flowers have exactly five ovules.

In genetics, a distinction is made between qualitative and quantitative traits. Eye and hair colours are qualitative traits while human height, milk-yield in cows, per hectare yield in case of rice or wheat are quantitative traits. It is generally believed that quantitative traits are polygenic. A large number of genes, each with a small effect, together determine the phenotype or the actual value. In such cases, the values in a population follow a bell shaped curve

or a normal (Gaussian) distribution. In this distribution, the average and near average values are very common while very large or very small values are rare.

Is it Just by Chance ?

In this section we discuss a rather subtle part of quantitative thinking. It is statistical inference. It becomes important because in ecology we often come across statements such as 'species A grows faster than species B' or 'in raptors females tend to be heavier than males'. How does one verify such claims? The obvious answer is 'by direct measurement'. Take a few cases of each type and check. This works in some cases. Suppose someone wanted to verify that humans are taller than bonnet monkeys. It turns out that every adult human being is taller than any adult bonnet monkey. So a direct measurement will also confirm this. But a statement 'men are taller than women' will fail such a test because some men are shorter than some women. A histogram of heights of men and women shows some overlap.

An alternative statement that is more acceptable is 'Average height of men is greater than average height of women'. How is it verified statistically? We measure the heights of a few men and of a few women, then compare the averages. Suppose the average for men is 162.5 cm and for women it is 162.1 cm. Is this enough evidence to claim that men are taller? Some may say that this difference can arise just by chance. If the difference was very large we would have believed the claim. But how large is 'very large'? Mathematical statisticians have laid down certain rules to decide this. Without explaining the technicalities we will illustrate the approach with an example.

Our interest is to check whether each of the ears in man is equally prone to develop hearing deficiencies (perhaps due to noise pollution). We check a series of case papers in an ear specialist's office. The first relevant case has a problem for the right ear. We say this could be just by chance. Our response is the same when the second and third cases are for the right ear. At the fourth case we get restless. At what point should we declare that the situation is

unexpected? One convention is that a chance of 1 in 100 is small enough to say so. Let us calculate the chances of getting many cases of right ear problems in succession. We use the multiplication law and find:

Event	Chance
First case is of right ear	1/2
First two cases are of right ear	$1/2 \times 1/2 = 1/4$
First three cases are of right ear	$1/4 \times 1/2 = 1/8$
First four cases are of right ear	$1/8 \times 1/2 = 1/16$
First five cases are of right ear	$1/16 \times 1/2 = 1/32$
First six cases are of right ear	$1/32 \times 1/2 = 1/64$
First seven cases are of right ear	$1/64 \times 1/2 = 1/128$

So by the convention, if we get seven cases in a row of right ear problems, we can discard the idea that both ears are equally prone and conclude that for some reason the right ear seems more prone to developing hearing deficiency. The take home message is that it takes a rather substantial tilting of evidence towards one side before it is treated as 'statistically significant' and an old viewpoint is changed.

Where There is Smoke There is Fire

This old Sanskrit saying is a fine description of how we think. 'We' includes humans, birds and bees. A male moth will follow the smell of a sex pheromone of its species to a distance of hundreds of meters, hoping to meet a sexually receptive female at the end of the search. Ivan Pavlov, the Russian scientist who studied animal behaviour, gave his dog some food after ringing a bell. Later, the dog would salivate at the ringing of the bell, expecting to get food right away. Deer in jungle prick their ears on hearing a monkey alarm call because the call may indicate the presence of a dangerous predator around. Something very akin to this has to be done in nature study also. As trees grow old their girth increases. We would like to guess the age of a tree by measuring its girth. That sounds interesting, but why bother? Here is a reason. If we measure girths of all trees of a species in a forest area and guess their ages, we get the age composition of the forest. Generally because of

Suggested Reading for the Series

[1] W B Fairley and F Moste-ller, *Statistics and Public Policy*, Addison-Wesley, London, 1977.

[2] M O Finkelstein and B Levin, *Statistics for Lawyers*, Springer, NY, 1990.

[3] R Hooke, *How To Tell The Liars From The Statisti-cians*, Marcel Dekker, NY, 1983.

[4] A J Jaffe and H F Spirer, *Misused Statistics: Straight Talk For Twisted Numbers*, Marcel Dekker, NY, 1987.

[5] J M Tanur, F Mosteller, WH Kruskal, R F Link, R S Pieters and G R Rising (Eds.), *Statistics, A Guide To The Unknown,* Holden-Day, San Francisco, 1972.

[6] H Zeisel, *Say it with Figures*, Harper, NY, 1957.

[7] H Zeisel and D Kaye, *Prove it with Figures*, Springer-Verlag, NY, 1997.

progressive mortality we expect to see fewer older trees and many younger ones. Suppose we find that five year old trees are very few. That means five years ago conditions were adverse for growth of new seedlings. This may be because all seeds were collected and removed or there was a severe drought or a big forest fire etc. If we find that proportions of young trees are consistently low, it is a cause for concern. There is inadequate regeneration. Is that species declining?

So it is useful to guess the age of a tree from its girth. Such a technique is called regression analysis. First we have to know somehow or the other, by direct record keeping if necessary, the ages of a few trees and their girths. A scatter plot has to be prepared. Finally a line or a smooth curve has to be drawn through the data points. Now if any new girth value is given, the graph can tell us the corresponding age estimate. This is only an estimate. For example some individual trees may get sick and remain thin even at an old age. But still such estimates are better than mere guesses.

If this technique is applied to tree species with very long life spans, then major events over hundreds of years can be guessed. Subash Chandran studied forests of Western Ghats in Uttara Kannada district of Karnataka and calculated proportions of evergreen and deciduous trees in each age class (i.e. girth class) in an area. Whenever the proportion of deciduous trees was high, he argued, forest disturbance and tree felling must have been high. He was able to relate this to various historical events such as timber extraction by the British. In this sense, the history of a forest is recorded in the forest itself.

We can think of many pairs of variables similar to girth and age in terms of relationships. Annual rainfall and standing plant biomass, number of very old dead trees per unit area and number of owls or hornbills or such tree cavity dwelling birds, extent of open water (as opposed to water with lots of vegetation) in a lake and proportion of duck species that love open water, level of pollution in a lake and number of fish species present, moisture levels in a forest and density of tree frogs etc. are examples where a relation-

ship will be revealed if data points are plotted on a graph.

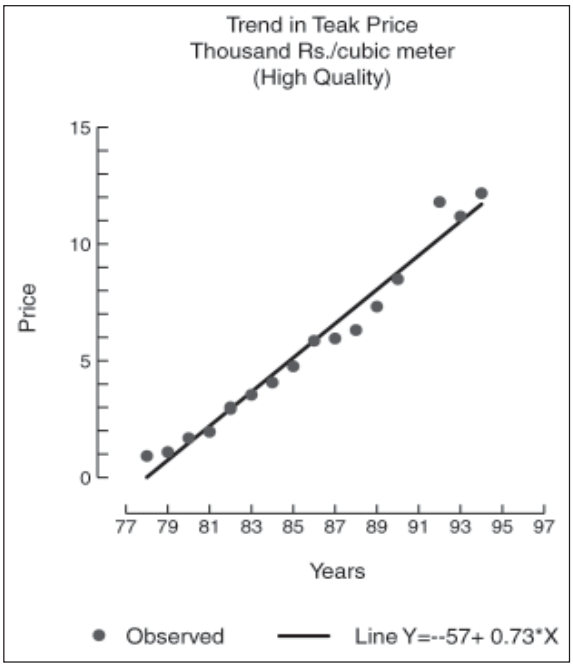
One thing to remember about using regression lines/curves is the risk of extrapolation. There is always a temptation to extend the curve beyond the limits of observed data. That can give misleading results. Consider yield of wheat in one acre for a given dose of nitrogen fertiliser. As the dose level increases, so does the yield. So you plot points, draw a line passing through the points (as many as possible) and extend it freely to the right. It will mean that you can get as much wheat as you want from one hectare just by putting in adequate urea. So the whole country can be fed from one hectare of land which is obviously wrong. This has happened because the line was extended to a region where it was invalid. Increase in wheat yield occurs only at moderate doses. Beyond a point, not only does the yield not rise, it may actually come down.

Sometimes extrapolation becomes necessary inspite of the difficulties. It takes several decades for teak plantations to grow before they can be harvested. What would be the price of teak wood in the market then? *Figure* 1 shows the rising trend in these prices in recent years. One guess of future prices can be obtained by fitting a regression line to the data and extending it as shown in the figure. Separate regression lines will have to be fitted for each quality.

Associations

The techniques of correlation and regression described above are useful to study relationships between characteristics like height, weight, age, temperature, rainfall etc. These are called continuous variables.

But how does one study the relation between two attributes? By 'attribute' we mean a trait that is not quantified. Here are some examples of such traits: Group foraging or individual foraging; chase and hunt like a leopard or wait and catch like a spider; colour of plumage – white like snowy owl or brown like spotted owlet. Let us briefly discuss the so called contingency tables used to study pairs of attributes.



Veena and Lokesha (1993) studied joint occurrence of drongos, common myna and jungle myna. They observed 177 feeding flocks of mynas and recorded the presence or absence of drongos.

Clearly for each type of flock, drongos are found to be around sometimes but not always. If drongos were always present (or absent) the matter would have been simple. Now the question is 'which kind of flock is more likely to have a drongo around?' In flocks of common mynas, drongos are present in (8/54=) 15% cases while in jungle myna flocks, the percentage is (21/34 =) 62 %. In mixed flocks the value is still higher (70/89 =) 79 %. The statistical question is whether these differences are just due to chance or whether there are significant differences between flock types. This is answered using the so called Chi-square test. If there are no differences, then in each case the expected percentage is (99/177 = 56%). The chi-square test checks whether the discrepancies between this expected percentage and the observed values are too large to be ascribed to chance. If the answer is yes, then the next question would be, why do drongos go with mixed flocks more often? The answer seems to be that mixed flocks are larger and hence cause greater disturbance on the ground which causes more insects to scurry around giving greater foraging opportunities to a drongo. In that case a comparison of sizes of flocks in which drongos are

Type of Flock	Drongo		Total
	Present	Absent	
Pure Flocks of Common Myna	8	46	54
Pure Flocks of Jungle Myna	21	13	34
Mixed Flocks	70	19	89
Total	99	78	177

Table 1. Co-occurrence of drongos and mynas.

Box 1. Exercises

1. Simulating family size under the rule 'stop reproduction as soon as a son is born': Almost every couple desires to have children. Most couples in India long for a son. Let us assume that the couple decides to stop having children once a son is born. We simulate a birth by tossing a coin. If it falls head, count it as birth of a son. If it is a tail, treat it as birth of a daughter. So stop tossing the coin as soon as a head is observed. Record the number of times you had to toss a coin (i.e. number of children). Repeat this experiment 50 times. Draw a histogram of number of children (the values will be 1,2,3...). Write a program to do this simulation and simulate family size 1000 times and prepare a histogram. How do the two histograms compare? As the number of simulations increases, the histogram smoothens out.

2. Simulating the dynamics of a farmer's cattle holding: Suppose a farmer has 2 bullocks and 2 cows. The bullocks will work well for the next five years. Will his cows generate their replacement? Many steps (uncertain) are involved. In any year a cow must conceive (probability=.9), must carry the foetus to full term (prob.=.7), must deliver a calf and survive (prob.=.8). The calf must be a male (prob.=.5). The male calf must survive to two years (prob.=.7). Then it can be a new working bull. All probabilities given here are notional. Assume gestation period of 1 year, for simplicity. Simulation has to be separate for each cow. Check how probable it is that the farmer will get two young bullocks to replace his older animals. Can the farmer manage with only one cow? The simulation should throw light on the question of whether we have too many cows.

3. ***Examining writing styles:*** A person is supposed to have a literary style which is reflected in different pieces of writing by that person. One can ask whether the sentence length distribution remains the same for an author. Examine this issue with reference to a) various articles in this series, b) various editorials of *Resonance*, c) two of your favourite writers.

present with those in which they are absent will reveal the same fact. But that will need data on flock sizes.

Suggested Reading

[1] S Kunte and Jeffreys, Lindley Paradox and a Related Problem. In *Bayesian Analysis in Statistics and Econometrics*, Ed. P K Goel and N S Iyengar, Lecture Notes in Statistics, 75, Springer Verlag, pp.249–255, 1992.

[2] V R Prayag, S A Paranjpe and A P Gore, Mixture Models for Dis-tribution of Number of Seeds per Pod in some Multiovulated Plants, in *Recent Advances in Agricultural Statistics Research*, Ed. Prem Narain, V K Sharma, O P Kathuria, Prajneshu, Wiley Eastern, pp.462–466, 1991.

[3] T Veena and R Lokesha, Association of Drongos with Myna Flocks: Are Drongos Benefitted? *J. Biosciences*, Vol.18, No.1, pp.111–119, 1993.

Address for Correspondence
A P Gore & S A Paranjpe
Department of Statistics
University of Pune
Pune 411 007, India.