The IIA-Penn State Astrostatistics School
2-7 July, 2007
Vainu Bappu Observatory, Kavalur

Laws of Probability, Bayes' theorem, and the Central Limit Theorem

Rajeeva L Karandikar
Cranes Software International Limited

Do Random phenomena exist in Nature?

Which way a coin tossed in air will fall is after all completely determined by laws of physics. The only problem in figuring out the trajectory and hence the face of the coin when it is on ground is that we will need to measure too many parameters!

Which way an electron will spin is also not known and is modeled as Random. But we cannot exclude the possibility that sometime in future, someone will come up with a theory that will explain the spin.

If Mathematics and Probability theory were as well understood several centuries ago as they are today but the planetary motion was not understood, perhaps people would have modeled the occurrence of a Solar eclipse as a random event and could have assigned a probability based on empirical occurrence. Subsequently, someone would have revised the model- observing that solar eclipse occurs only on a new moon day. Of course, after more time, the phenomenon would be completely understood and the model changed from a stochastic or random model to deterministic.

Thus we often come across events whose outcome is uncertain. The uncertainty could be:

- Because of inability to observe accurately all the inputs required to compute the outcome. It may be too expensive or even counterproductive to observe all the inputs.
- Due to the current level of understanding of the phenomenon.
- On account of the outcome depending on choices made by a group of people at a future time - such as outcome of an election yet to be held.

So if we need to model the outcome mathematically, one way is to use Probability theory and model it as Random.

A natural question: If the outcome is uncertain, why do we think that it is possible to model it mathematically?

It has been observed that events that are uncertain at micro level appear to be deterministic at macro level. For example,

- While the sex of a child about to be conceived is uncertain, the proportion of boys / girls born in a city / village over a period of time - say an year is stable and this has been observed over centuries across several countries where birth records have been kept.
- The same phenomenon- uncertain at micro level but almost deterministic at macro level has been observed when it comes to weather data: rainfall, high tide levels, maximum and minimum temperatures.
- Yet another example where uncertainty at micro level leads to deterministic behavior at macro level is radioactivity. While which (if any) molecule of a radioactive substance will disintegrate in the next millisecond is uncertain as far as current understanding of physics is concerned, it has been observed that in a fixed time interval -half-life of the substance - the substance reduces to exactly half its initial weight.

Probability theory attempts to capture this phenomenon - of micro level uncertainty giving rise to deterministic behavior at macro level. The macro level observations are a guide to construction of the model for micro level.

Thus we can view Randomness as a model for uncertainty.

Randomness is in the eye of the observer

Suppose an experiment $\mathcal{E}$ results in one of the outcomes $\{e_1, e_2, \ldots e_m\}$. We need to model probability of each of these outcomes- *i.e.* assign a probability p_i to the outcome e_i for every i in such a way that the probabilities add up to 1.

The set $\Omega = \{e_1, e_2, \ldots e_m\}$ is called sample space and each outcome e_i is called an elementary event. A subset $A \subseteq \Omega$ is called an event. Any real valued function from Ω is called a random variable.

Recall: $P(e_i) = p_i$. For an event A, we define

$$P(A) = \sum_{i\,:\,e_i \in A} p_i.$$

Easy to check that if A, B are mutually disjoint, *i.e.* $A \cap B = \phi$ then

$$P(A \cup B) = P(A) + P(B)$$

More generally, we can check that for any two events A, B

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

And for three events A, B, C

$$\begin{aligned} P(A \cup B \cup C) \;=\; & P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ & + P(A \cap B \cap C) \end{aligned}$$

This has generalization to n events.

Question: How do we assign the probabilities p_i to the elementary outcomes?

The simplest case is when due to inherent symmetries present, we can model all the elementary events (*i.e.*outcomes) as being equally likely.

Example 1: Toss of a coin. $\Omega = \{H, T\}$ with $P(H) = .5$ and $P(T) = .5$ This is our model for the probabilities if we believe that there is symmetry.

Example 2: Toss of a dice. $\Omega = \{1, 2, 3, 4, 5, 6\}$ with $P(i) = \frac{1}{6}$ for $i = 1, 2, \ldots 6$.

Example 3: Toss of 5 Rs coin, 2 Rs coin and 1Rs coin together. We list outcomes as

$$\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\},$$

where first letter is outcome of 5 Rs. coin, second of 2 Rs. coin and third of 1 Rs. coin. Here it is reasonable to model all the outcomes as equally likely. Once we agree on this, the model then becomes

$$P(e_i) = \frac{1}{8} \text{ for all } i$$

Example 4: Toss of 5 Rs coin, 2 Rs coin and 1Rs coin together and we note the number of heads. The list of outcomes now is $\{0, 1, 2, 3\}$. Do we have reason to model these as equally likely?

A little thought would convince us otherwise. An experiment would also reveal the same: if we toss the three coins 100 times we would observe that $1, 2$ occur far more than $0, 3$.

So the events are not equally likley.

Question: How do we assign the probabilities?

We have seen that in Example 3, all the 8 outcomes were equally likely. Let us write Z for the outcome of example 3 and Y as the outcome of example 4. Then:

$$Y = 0 \text{ is same as } Z = \text{TTT}$$
$$Y = 1 \text{ is same as } Z = \text{HTT, THT or TTH}$$
$$Y = 2 \text{ is same as } Z = \text{HHT, HTH or THH}$$
$$Y = 3 \text{ is same as } Z = \text{HHH.}$$

Hence it is reasonable to model the probabilities as
$P(Y = 0) = P(Y = 3) = 0.125$, $P(Y = 1) = P(Y = 2) = 0.375$.

This example illustrates that even when the outcomes are not equally likely, we may be able to identify outcomes as combinations of outcomes of another experiment whose outcomes are equally likely and thus obtain a model for the probabilities.

When an experiment $\mathcal{E}$ results in m equaly likely outcomes $\{e_1, e_2, \ldots e_m\}$, probability of any event A is simply

$$P(A) = \frac{\#A}{m}$$

which is often read as ratio of number of favorable outcomes and the total number of outcomes.

Example 5: Consider toss of a usual dice and let X be the number of dots on the side turns up. There are six possible outcomes, reasonable to model them as *equally likely*. Suppose the dice has been tossed but we are unable to observe the outcome. But we are told that X is even. What is the probability that the outcome belongs to $\{1, 2, 3\}$?

Now the six outcomes are no longer equally likely. We know that the outcome is one out of $\{2, 4, 6\}$. So the new model is that these three outcomes are equally likely, since we know nothing more than the fact that the true outcome is one of the three. So the revised probability that the outcome belongs to $\{1, 2, 3\}$ equals $\frac{1}{3}$.

Thus let us consider an experiment with m equally likely outcomes and let A and B be events. If we are given the information that B has happened, what is the probability that A has happened in the changed circumstances? This probability is called conditional probability of A given B, written as $P(A \mid B)$.

Let $\#A = k$, $\#B = n$, $\#(A \cap B) = i$. Then as noted above, given that B has happened, the new probability allocation assigns probability $\frac{1}{n}$ to all the outcomes in B.

Out of these n, $\#(A \cap B) = i$ outcomes belong to A. Hence

$$P(A \mid B) = \frac{i}{n}.$$

Noting that $P(A \cap B) = \frac{i}{m}$ and $P(B) = \frac{n}{m}$, it follows that

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

In general when A, B are events such that $P(B) > 0$, the conditional probability of A given that B has occurred $P(A \mid B)$ is defined by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

This leads to the Multiplicative law of probability:

$$P(A \cap B) = P(A \mid B)P(B)$$

This has a generalization to n events:

$$\begin{aligned} P(A_1 \cap A_2 \quad \cap \quad & \ldots A_n) \\ = \quad & P(A_n \mid A_1, \ldots A_{n-1}) \times \\ & P(A_{n-1} \mid A_1, \ldots A_{n-2}) \times \\ & \ldots P(A_2 \mid A_1)P(A_1) \end{aligned}$$

The Law of Total Probability:

Let A be any event and $B_1, \ldots, B_k$ be a partition of the sample space Ω. Then

$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_k)P(B_k)$$

This follows from the observation

$$P(A) = P(A \cap B_1) + \cdots + P(A \cap B_k)$$

and the Multiplicative law of probability,

$$P(A \cap B_i) = P(A|B_i)P(B_i)$$

Example 6. Suppose an urn has 7 red balls and 7 green balls. A dice is tossed and if the outcome is i $(1 \leq i \leq 6)$, i red balls and $7 - i$ green balls are added to the urn. Now the balls in the urn are mixed and one ball is drawn. Let A be the event that the ball so drawn is red.

To find $P(A)$, take B_i to be the event that i red balls and $7 - i$ green balls are added to the urn. Since given B_i, we know the number of red and green balls in the urn, we can calculate $P(A \mid B_i)$

Suppose now that we have observed that A has happened. Thinking of B_i as possible causes that may have led to the event A, we would like to obtain

$$P(B_i \mid A).$$

Bayes' formula or Bayes' theorem gives the answer (in terms of $P(A \mid B_i)$ and $P(B_i)$). The result is very easy to prove and is the basis of "Baysian Inference".

Bayes' Theorem:

If $B_1, B_2, \ldots B_n$ is a partition of the sample space, then

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)}$$

There are many controversies and apparent paradoxes associated with Conditional probabilities. The root cause generally is incomplete specification of the conditions in the word problems. See the articles

`Three_cards_problem`
`Monty_Hall_problem`
`Bertrand's_box_paradox`

on the site

`http://en.wikipedia.org/wiki/`

Suppose that A, B are events such that

$$P(A \mid B) = P(A)P(B).$$

Now we can verify that

$$P(A \mid B) = P(A).$$

The knowledge that B has occurred has not altered the probability of A. In this case, A and B are defined to be independent.

Let X, Y, Z be random variables each taking finitely many values. Then X, Y, Z are said to be independent if

$$P(X = i, Y = j, Z = k) = P(X = i)P(Y = j)P(Z = k)$$

for all possible values i, j, k of X, Y, Z respectively. This can be generalized to finitely many random variables.

Expectation of a random variable:

Let X be a random variable taking values $x_1, x_2 \ldots x_n$. The expected value of X denoted by $E(X)$ is defined by

$$E(X) = \sum_{i=1}^{n} x_i P(X = x_i).$$

Variance of a random variable is defined by

$$Var(X) = E\{(X - \mu)^2\}$$

where $\mu = E(X)$.

All the statements made above continue to be valid if the sample space contains countably many elementary events and the random variables may take countably many values. One has to be careful when defining $E(X)$.

If X takes non-negative values, $E(X)$ can be defined as above: if X takes values $x_1, x_2, \ldots$ with $x_i \geq 0 \forall i$,

$$E(X) = \sum_{i=1}^{\infty} x_i P(X = x_i).$$

For a general random variable, $E(X)$ is defined (by the same formula) if $E(\mid X \mid) < \infty$.

Examples of random variables:

- **The Binomial distribution:** Consider n independent trials where probability of success in each trial is p and X denotes the total number of successes, then

$$P(X = k) = {}^{n}C_{k}\, p^{k}(1-p)^{n-k}$$

for $k = 0, 1, \ldots n$, $0 < p < 1$. This is known as Binomial distribution, written as $X \sim B(n, p)$. $E(X) = np$ and $Var(X) = np(1-p)$.

Example: (See *A. Mészáros, "On the role of Bernoulli distribution in cosmology," Astron. Astrophys., 328, 1-4 (1997).)*

n uniformly distributed points in a region of volume $V = 1$ unit

X: No. of points in a fixed region of volume p

X has a binomial distribution, $X \sim B(n, p)$

- **The Poisson distribution** Consider a random variable X such that

$$P(X = k) = \frac{\lambda^{k}}{k!} e^{-\lambda}$$

for $k = 0, 1, \ldots$.

This is known as Poisson distribution. Here $E(X) = \lambda$ and $Var(X) = \lambda$.

If X has Binomial distribution $B(n, p)$ with large n and small p, then distribution of X can be approximated by Poisson distribution with parameter $\lambda = np$:

$$P(X \leq a) \text{ approximately equals } P(Y \leq a)$$

where Y has Poisson distribution with parameter $\lambda = np$.

(See: *M. L. Fudge, T. D. Maclay, "Poisson validity for orbital debris ..." Proc. SPIE, 3116 (1997) 202-209, Small Spacecraft, Space Environments, and Instrumentation Technologies*)

Question: International Space Station is at risk from orbital debris and micrometeorite impact. How to assess risk of a micrometeorite impact?

The fundamental assumption underlying risk modeling in this case is that the orbital collision problem can be modeled using a Poisson distribution This assumption found to be appropriate based upon the Poisson as an approximation for the binomial distribution and that is it proper to physically model exposure to the orbital debris flux environment using the binomial distribution.

- **The Geometric distribution** Consider n independent trials where probability of success in each trial is p and X denotes the number of failures observed before the first success, then

$$P(X = k) = (1 - p)^k p$$

for $k = 0, 1, \ldots$. This is known as geometric distribution.

Now, $E(X) = \frac{q}{p}$ and $Var(X) = \frac{q}{p^2}$ where $q = 1 - p$.

- **The negative binomial distribution:** Consider n independent trials where probability of success in each trial is p and X denotes the number of failures before observing r successes, then the possible values of X are $0, 1, 2, \ldots$ with

$$P(X = k) = {}^{r+k-1}C_k \, (1 - p)^k \, p^r$$

X is said to have negative Binomial distributon.

(See *Neyman, Scott, and Shane (1953, ApJ): Counts of galaxies in clusters*).

ν: The number of galaxies in a randomly chosen cluster

Basic assumption: ν follows a negative binomial distribution

In order to consider random variables that may take any real number or any number in an interval as its value, we need to extend our notions of sample space and events. One difficulty is that we can no longer define probabilites for all subsets of the sample space. We will only note here that the class of events - namely the sets for which the probabilities are defined is large enough.

We also need to add an axiom called the Countably additivity axiom:

If $A_1, A_2, \ldots, A_k, \ldots$ are pairwise mutually exclusive events then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

A real valued function X from the sample space Ω is said to be a random variable if for all real numbers a, the set $\{\omega : X(\omega) \leq a\}$ is an event.

We have not said much about the class of events for which probabilites are defined. We only remark that the class is large enough so that if $Y, Z, X_1, X_2, \ldots$ are random variables then U, V, W, defined below are also random variables:

$$U = Y + Z \ , \ V = max(Y, Z)$$

$$W = g(X_1, X_2, \ldots, X_m)$$

where g is a continuous funtion on R^m. Further, if $\lim X_n$ exists and equals X then X is also a random variable.

For a random variable X, the function F defined by

$$F(x) = P(X \leq x)$$

is called its distribution function. If there exists a function f such that

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

then f is called the density of X.

Examples of densities:

- **Exponential density:**

$$f(x) = \lambda \exp(-\lambda x), \quad x \geq 0$$

  and zero otherwise.

- **Normal density:**

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{-\frac{(x-\mu)^2}{2\sigma^2}\Big\}$$

For a random variable $\boldsymbol{X}$ with density $\boldsymbol{f}$, the expected value of $\boldsymbol{g(X)}$, where $\boldsymbol{g}$ is a function is defined by

$$\boldsymbol{E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx.}$$

As in discrete case, $\boldsymbol{E[g(X)]}$ is defined as above if

$$\boldsymbol{g(x) \geq 0 \; \forall x}$$

or if

$$\boldsymbol{E[\,|g(X)|\,] < \infty.}$$

For a random variable $\boldsymbol{X}$ with Normal density

$$\boldsymbol{f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{-\frac{(x-\mu)^2}{2\sigma^2}\Big\}}$$

(written as $\boldsymbol{X \sim N(\mu, \sigma^2)}$),

$$\boldsymbol{E[X] = \mu}$$

$$\boldsymbol{Var[X] = E[(X-\mu)^2] = \sigma^2.}$$

When an experiment is repeated several times, the succesive observations constitute independent random variables with common distribution. Such a sequence is called a sequece of *i.i.d.* (independent identically distributed) random variables. These sequences exhibit the phonomenon refered to at the begining: uncertainity at micro level leading to deterministic behaviour at macro level. This is the consequnce of Law of large numbers:

Law of Large Numbers

Suppose $X_1, X_2, \ldots$ is a sequence of *i.i.d.* random variables with

$$E(\mid X_1 \mid) < \infty.$$

Then

$$\bar{X}_n = \sum_{i=1}^{n} \frac{X_i}{n}$$

converges to $\mu = E(X_1)$: *i.e.* for all $\epsilon > 0$

$$P(\mid \bar{X}_n - \mu \mid > \epsilon) \longrightarrow 0.$$

For a sequence of *i.i.d.* random variables $X_1, X_2, \ldots$, if $E(\mid X_1 \mid^2) < \infty$, then the distribution of $\bar{X}_n$ can be approximated by normal distribution. This was first proven for the case when each X_i takes only two values way back in 1773 by De Moivre.

Central Limit Theorem.

Suppose $X_1, X_2, \ldots$ is a sequence of *i.i.d.* random variables with $E(|\ X_1\ |^2) < \infty$. Let $\mu = E(X_1)$ and $\sigma^2 = E[(X_1 - \mu)^2]$. Let

$$\bar{X}_n = \sum_{i=1}^{n} \frac{X_i}{n}.$$

Then

$$P\Big\{\sqrt{n}\Big(\frac{\bar{X}_n - \mu}{\sigma}\Big) \leq x\Big\} \longrightarrow \Phi(x)$$

where

$$\Phi(x) = \int_{-\infty}^{x} f(t)dt,$$

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\Big\{-\tfrac{x^2}{2}\Big\}.$$

Special case: $X \sim Binomial(n, p)$, n large. Then

$$P(X \leq a)$$

can be approximated by

$$\Phi\Big(\sqrt{n}\big(\tfrac{a-np}{\sqrt{p(1-p)}}\big)\Big)$$