Lecture 11:

Probabilistic & Stochastic Modeling (Part I)

©

The purpose of this and the next lecture is to introduce probabilistic reasoning and stochastic modeling. Historically, dynamical systems, at the end of the 19th century led to a deterministic view of the physical world. The first half of the twentieth century witnessed the overthrow of determinism by the essentially probabilistic quantum theory. Of relevance to laboratory work is that variability in biological preparations and instrument noise, common features of any experiment introduce a probabilistic element. To model experiment, and the data that arises from experiment, it is desirable to be able to characterize both *trend* and fluctuations in the presence of noise. As required this should be accomplished dynamically, i.e., as phenomena evolves in time.

Probability

Classical probability is based on frequency of occurrence. For example if in tossing a coin N times, N_h heads occur then the probability p of tossing a head is

$$p = \lim_{N \uparrow \infty} \frac{N_h}{N} : \tag{11.1}$$

For generality we need to consider *tossing* an unfair coin. A useful substitute model is an urn filled with a large number of identical marbles except for color, which say are black or white. One can then blindly choose a marble, with replacement, as an equivalent to coin tossing. The *odds* only depend on the composition of the urn. Thus we can speak of p_b & p_w as the probabilities of black & white draws and

$$p_b + p_w = 1 \tag{11.2}$$

The urn model can immediately be generalized to any number of different colored marbles therefore many possible outcomes with probabilities $p_1, p_2, ...p_k$ and such that

$$p_1 + p_2 + ... + p_k = 1. \tag{11.3}$$

Modern probability deliberately avoids the above constructive approach in the hope of achieving clarity. It starts with a sample space Ω composed of events E, and a probability function defined on the events so that:

(i) For $E \varepsilon \Omega$, $P(E) \leq 1$

(ii) $P(\Omega) = 1$

(iii) For mutually exclusive events $E_1, E_2, E_3, ...$

$$P(E_1 + E_2 + E_3 + ...) = P(E_1) + P(E_2) + P(E_3) + ... \tag{11.4}$$

Ω is purely a set of events. For example for the toss of 10 equal coins or 10 tosses of one coin there are 2^{10} events in Ω. If we are interested in say the probability of 5 heads in the 10 tosses knowing all the individual cases is not handy. To deal with this and other issues a function from Ω to the real numbers, known as a random variable is introduced. Thus in this example if the random variable X is the number of heads we write

$$p_k = P(X = k) \tag{11.5}$$

as the probability of k heads in the 10 tosses. Therefore if

$$p_j = P(X = j);\ j = 0, 1, ...10 \tag{11.6}$$

denotes the eleven possible outcomes then

$$\sum_j p_j = 1. \tag{11.7}$$

Another example is throwing a pair of dice in which case the interest is not in the 36 possible throws of Ω but rather the sum of the pair, and therefore the random variable X in

$$p_n = P(X = n); n = 2, 3, ...12. \tag{11.8}$$

Enumeration of the outcomes in the game of dice yields the following Table.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Table 11.1: Possible tosses of each die are shown in the outer column and upper row. The inner square matrix indicates the sums.

Coin Tossing

Each outcome of a coin flip is entirely independent of the prior toss, i.e., memory does not enter. (After a run of only heads, to say a tail is due is naive.) The particular run of k tails and $n-k$ heads is illustrated in Table 11.2.

1	2			k	·	·	·	·	n
t	t	·	·	t	h	h	·	·	h

Table 11.2: *A run of k tails, followed by a run of $n-k$ heads.*

has probability

$$q^k p^{n-k}. \tag{11.9}$$

To determine the probability of exactly k tails in n tosses, $B_n(k)$, observe that in principal there are $n!$ ways of rearranging all the n cells of Table 11.2, and that there are $k!$ ways of permuting the first k-cells, and $(n-k)!$ the last $(n-k)$-cells both of which leave the outcome unchanged thus the answer to the question "What is the probability of exact k tails in n tosses?" is

$$B_n(k) = \frac{n!}{(n-k)!k!} p^{n-k} q^k, \tag{11.10}$$

which is known as the Bernoulli probability distribution. Solitary thought is usually needed to absorb this argument. An enumerated concrete case such as $n = 3$ below also can be of help

$P_3(0)$	t t t
$P_3(1)$	h t t t h t t t h
$P_3(2)$	t h h h h t h t h
$P_3(3)$	h h h

Table 11.3: Enumeration of all outcomes of three tosses.

For $B_n(k)$ to be a probability we require

$$\sum_{k=0}^{n} B_n(k) = 1, \tag{11.11}$$

and to see that this is true observe that the binomial theorem, Lecture 1, states

$$(p+q)^n = \sum_{k=0}^{n} \frac{n!}{(n-k)!k!} p^{n-k} q^k \tag{11.12}$$

but $p+q=1$ and therefore (11.11) holds.

A symmetrical way to write (11.12) is

$$(p_1+p_2)^n = \sum_{\substack{n_1+n_2=n \\ n_1,n_2\geq 0}} \frac{n!}{n_1!n_2!} p_1^{n_1} p_2^{n_2}, \tag{11.13}$$

where the summation is for n_1 & n_2 such that $n_1+n_2=n$. We can generalize this to an urn with l different marbles having probabilities $p_1+p_2\cdots+p_l=1$. The counterpart to (11.13) then is

$$1 = (p_1+p_2+\cdots p_l)^N = \sum_{\substack{n_1+n_2\cdots+n_l=N \\ n_k\geq 0}} \frac{N!}{n_1!\cdots n_l!} p_1^{n_1} p_2^{n_2} \cdots p_l^{n_l}. \tag{11.14}$$

The Bernoulli probability (11.10) is binomial and its multinomial generalization from (11.14) is given

$$P(X_1=n_1, X_2=n_2, ...X_l=n_l) = \frac{N!}{n_1!...n_l!} p_1^{n_1} p_2^{n_2} ... p_l^{n_l},\ n_1+...+n_l=N. \tag{11.15}$$

Both are two examples of common probability distributions. Both are discrete examples but the extension to continuous distributions is straightforward.

In all cases of a probability distribution (pdf) $P(x)$ must respect the two formal conditions (i) & (ii) which in present notation are:

$$P(x) \geq 0 \tag{11.16}$$

and

$$\sum_x P(x) = 1 \tag{11.17}$$

where the last is deliberately ambiguous so as to cover both discrete and continuous (integration) cases.

An entire class of continuous pdfs are obtained free of effort from the gamma function (1.57). Clearly the definition of $s!$ the function

$$G_s(x) = \frac{x^s e^{-x}}{s!} \tag{11.18}$$

satisfies (11.16) & (11.17) for any s, and for obvious reasons is known as a gamma distribution. The range of applications of (11.18) can be extended by setting $x = \lambda t$, but to avoid a common error we first observe

$$1 = \int_0^\infty G_s(x)dx \xrightarrow[x=\lambda t]{} \int_0^\infty \frac{(\lambda t)^s e^{-\lambda t}}{s!} \lambda dt \tag{11.19}$$

Therefore

$$G_s(t;\lambda) = \frac{\lambda(\lambda t)^s e^{-\lambda t}}{s!} \tag{11.20}$$

is the appropriate new temporal pdf, and which is said to be of rate λ.

In the general case if $P(x)$ is a pdf, which might be considered as a pdf in the variable y where $x(y)$ is monotonic then denoting the range of integration by $\{\}$

$$1 = \int_{\{x\}} dx P(x) = \int_{\{y\}} \frac{dx}{dy} P(x(y))dy \tag{11.21}$$

and the pdf in y is

$$\mathcal{P}(y) = \frac{dx(y)}{dy} P(x(y)). \tag{11.22}$$

Failure to consider the jacobian dx/dy is a frequent source of error.

Miscellaneous

A handy calculation is

$$\mathcal{E}(x) = \lim_{n\uparrow\infty} \left(1 - \frac{x}{n}\right)^n \tag{11.23}$$

Observe that

$$\ln\left(1-\frac{x}{n}\right)^n = n\ln\left(1-\frac{x}{n}\right) = -n\left(\frac{x}{n}+\frac{x^2}{n^2}+...\right) \tag{11.24}$$

and therefore

$$\mathcal{E}(x) = e^{-x}(1+O(\frac{1}{n})). \tag{11.25}$$

Another useful exercise is the determination of $n!$ for large n

$$n! = \int_0^\infty e^{-t+n\ln t}dt = \int_0^\infty e^{f(t,n)}dt. \tag{11.26}$$

Clearly the integrand has a maximum in the interval of integration. To locate this observe that $\frac{df}{dt}=0$ implies that the extremal is at

$$t = n \tag{11.27}$$

and if we expand the exponent in the neighborhood of (11.27) we obtain

$$f = -n + n\ln n - \frac{1}{2n}(t-n)^2 + O((t-n)^3). \tag{11.28}$$

Next under the variable change $t = sn$ we obtain

$$\sim e^{-n}n^{n+1}\int_0^\infty e^{-\frac{n}{2}(s-1)^2}ds \tag{11.29}$$

Since n is large the integrand of (11.29) has a sharp peak, of width $O(n^{-1/2})$, and the lower limit can be reasonably extended to $-\infty$. With this and the further variable change

$$(s-1) = \frac{x}{\sqrt{n}} \tag{11.30}$$

we can write

$$n! \sim e^{-n}n^{n+1/2}\sqrt{2\pi}\int_{-\infty}^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}}dx. \tag{11.31}$$

Two consequences of interest follow from the fact that

$$\int_{-\infty}^\infty e^{-x^2/2}\frac{dx}{\sqrt{2\pi}} = 1. \tag{11.32}$$

First is Stirling's formula

$$n! \sim e^{-n}n^{n+1/2}\sqrt{2\pi}, \tag{11.33}$$

and second is the fact that

$$\omega = \frac{e^{-x^2/2}}{\sqrt{2\pi}} \tag{11.34}$$

is a pdf on $-\infty < x < \infty$, and is referred to as the normal distribution, or the Gaussian[1].

Exercise 11.1 (a) Go through the steps leading to (11.33) in more careful terms.

(b) Note that no where was it assumed that n is an integer. Compare Stirling's form for a continuous range of $n \geq 1$, use Matlab's **gamma** for this.

Another result of interest is contained in the next exercise.

Exercise 11.2 (a) Consider the Bernoulli distribution in the limit n large, p small such that $np \to \lambda t$ is to show

$$\lim_{n\uparrow\infty} B_n = \frac{(\lambda t)^n e^{-\lambda t}}{n!} = P_n(t) \tag{11.35}$$

This is called the Poisson distribution. Based on our derivation of $P_n(t)$ is the probability of n "events" in a time t defined by a rate λ. Note the contrast with the gamma distribution (11.20) which is pdf in t.

(b) (11.35) is a pdf in n and therefore we should have

$$\sum_{n=o}^{\infty} P_n(t) = 1 \tag{11.36}$$

for any t. Prove this.

Expectation

In general the expected value of any function of x say $f(x)$ is defined as

$$E(f) = \langle f \rangle = \overline{f} = \sum_x f(x)P(x) \tag{11.37}$$

and when money is involved it gives the expected gain. The mean is

$$E(x) = \sum_x xP(x) = \langle x \rangle = \overline{x} = \mu. \tag{11.38}$$

Variance

Once a mean has been determined for $P(x)$ then the variance is defined by

$$\sigma^2 = \left\langle (x-\mu)^2 \right\rangle = \langle x^2 \rangle - \mu^2. \tag{11.39}$$

For non-zero mean the coefficient of variation

$$c_v = \frac{\sigma}{\mu}, \tag{11.40}$$

[1] An historical error since de Moivre a century before Gauss fully demonstrated its central role of (11.34) in science.

is a convenient dimensionless ratio of data.

Characteristic Function

The expectation of e^{ixt} for a pdf $P(x)$ is called the characteristic function

$$\phi(t) = \langle e^{-ixt} \rangle = \sum_x e^{-ixt} P(x) \tag{11.41}$$

which can be very handy in calculations. Clearly

$$\begin{aligned} \phi(0) &= 1 \\ \phi'\big|_{t=o} &= -i\langle x \rangle \\ \phi''\big|_{t=o} &= -\langle x^2 \rangle \end{aligned} \tag{11.42}$$

which is enough to calculate mean and variance, and if we continue high moments of interest can be evaluated. For example if $P = B_n(k)$ then since $e^{ikt} = (e^{it})^k$

$$\phi(t) = \sum_{k=o}^{n} \frac{n!}{(n-k)!k!} p^{n-k} q^k e^{-ikt} = (p + qe^{-it})^n \tag{11.43}$$

and $P = \omega(x)$, (11.34), then

$$\phi(t) = \int_{-\infty}^{\infty} e^{-ixt} \omega(x) dx. \tag{11.44}$$

Exercise 11.3. (a) Use (11.43) to obtain the mean and variance of B_n.

(b) Show that (11.44), which is the fourier transform of the gaussian, is

$$\phi(t) = e^{-t^2/2}. \tag{11.45}$$

Exercise 11.4 (Special).

Suppose we are tossing a fair silver dollar for *real*, so that each time we get a head we win, and each time we toss a tail we lose, a dollar, so that the gain, G_n, in dollars after n tosses is

$$G_n = (n-k) - k = n - 2k = 2(\frac{n}{2} - k) \tag{11.46}$$

where k is the number of tails thrown, so that G_n can be negative.

(a) Show that the expected gain after N tosses is zero.

(b) Explore the drift in G_N by considering

$$\sqrt{\langle G_N^2 \rangle} = 2\sigma = \sqrt{N} \tag{11.47}$$

(c) Set up a Matlab program to perform 100 trials of 100 tosses for a fair coin and calculate $B_n(k)$ and calculate by averaging over the 100 trials. Is

$$\frac{\sigma}{N} = \frac{1}{2\sqrt{N}} \tag{11.48}$$

a reasoned measure of the fairness of a coin?

(d) Find mean and variance for any coin, (11.10)

Generalities

In a general setting we are confronted with the finite sampling of some totally unknown probability function $\mathcal{P}(x)$ (say through experiment). It is useful to pretend that we know $\mu = \langle x \rangle$ and $\sigma^2 = \langle (x-\mu)^2 \rangle$, based on $\mathcal{P}(x)$. If $\{x_j\}$, $j = 1, ..., N$ represents N samples chosen from $\mathcal{P}(x)$, then

$$\overline{x} = \frac{1}{N} \sum_{j=1}^{N} x_j \tag{11.49}$$

might be regarded as an approximation to μ. In fact on this basis the expectation of the righthand side of (11.49) shows that

$$\langle \overline{x} \rangle = \mu. \tag{11.50}$$

Formally the expectation of (11.49) requires $\mathcal{P}(x_1, x_2, ...x_N)$, but in the present circumstance this is clearly given by $\mathcal{P}(x_1)\mathcal{P}(x_2)...\mathcal{P}(x_N)$. (11.49) is said to be an unbiased estimator of μ, i.e., the expectation of the quantity is the quantity being estimated. We pause to mention that at this point we have passed the blurry boundary between Probability and Statistics[2].

A likely candidate for estimating σ^2 is

$$\widetilde{s}^2 = \frac{1}{N} \sum_{j=1}^{N} (x_j - \overline{x})^2, \tag{11.51}$$

but as the following exercise demonstrates this is not true.

Exercise 11.5. Show that

[2] "Statistics is a funny subject. The first time you go through it, you don't understand it at all. The second time you go through it, you think you understand it, except for one or two small points. The third time you go through it you know you don't understand it, but by that time you are so used to it, it doesn't bother you anymore." With a tip of the hat to Arnold Sommerfeld who first said this for Thermodynamics.

$$\langle \widetilde{s}^2 \rangle = \frac{N-1}{N}\sigma^2. \tag{11.52}$$

It therefore follows that

$$s^2 = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \overline{x})^2 \tag{11.53}$$

is an unbiased estimator of σ^2.

On the basis of (11.47) we can say

$$\frac{X_1 + X_2 + ... + X_N}{N} - \mu = O(\frac{1}{\sqrt{N}}) \tag{11.54}$$

This embodies what is:

The Law of Large Numbers

This states that one can always choose N large enough so that

$$Pr\left(\frac{X_1 + ... + X_N}{N} - \mu\right) \tag{11.55}$$

is as small as we please. This is one of the two cornerstones of probability. The other is the remarkable:

Central Limit Theorem:

Suppose $P(x)$ is an arbitrary pdf with mean and variance μ and σ^2 and for which nothing more is known except that higher moments exist. Further imagine $X_1, X_2, ..., X_N$ are random variables taken from $P(x)$. Then

$$Pr\left(\sum_{j=1}^{N} X_j = N\mu\right) = \frac{e^{-(\sum_{j=1}^{N} x_j - N\mu)^2/2N\sigma^2}}{\sqrt{2\pi N\sigma^2}} \tag{11.56}$$

Alternately in terms of $\overline{x}$, (11.49)

$$P(\overline{x}) = \frac{e^{-(\overline{x}-\mu)^2/2(\sigma/\sqrt{N})^2}}{\sqrt{2\pi(\sigma/\sqrt{N})^2}}. \tag{11.57}$$

Proof: Consider the characteristic function of $P(x_1, ..., x_N) = P(x_1)...P(x_N)$

$$\Phi(t) = \int e^{-it(\sum_{j=1}^{n} x_j)} f(x_1)...f(x_N)dx_1...dx_N = (\phi(t))^N \tag{11.58}$$

Therefore

$$P(x_1, x_2, ..., x_N) = \int e^{it(\sum_{j=1}^{N} x_j)}(\phi(t))Ndt \tag{11.59}$$

But from (11.42) we know

$$\phi = 1 - i\mu t - \langle x^2 \rangle t^2/2 + ... \tag{11.60}$$

Therefore

$$\ln \phi \sim -i\mu t - \langle x^2 \rangle \frac{t^2}{2} + \frac{1}{2}\mu^2 t^2 + ... \tag{11.61}$$

and from (11.39)

$$\langle x^2 \rangle = \sigma^2 + \mu^2 \tag{11.62}$$

and therefore

$$\ln \phi \sim -i\mu t - \sigma^2 t^2/2. \tag{11.63}$$

From this it follows that

$$N \ln \phi \sim -iN\mu t - N\sigma t^2/2 + ... \tag{11.64}$$

For N large this shows that the area under the integral is almost entirely in the neighborhood of the origin and

$$\begin{aligned} P(x_1, x_2, ..., x_N) &= \int_{-\infty}^{\infty} e^{it(\sum_{j=1}^{N} x_j - N\mu) - N\sigma^2 t^2/2} dt \\ &= \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{(\sum_{j=1}^{N} x_j - N\mu)^2}{2N\sigma^2}}}{\sqrt{N\sigma^2}} \end{aligned} \tag{11.65}$$

which demonstrates (11.56).

The Central Limit Theorem states that under mild conditions, the mean of N samples taken from any pdf is distributed as a Gaussian, as the number is increased.

Figure 11.1 below shows the result of random selection from the uniform distribution on $(0, 1)$.

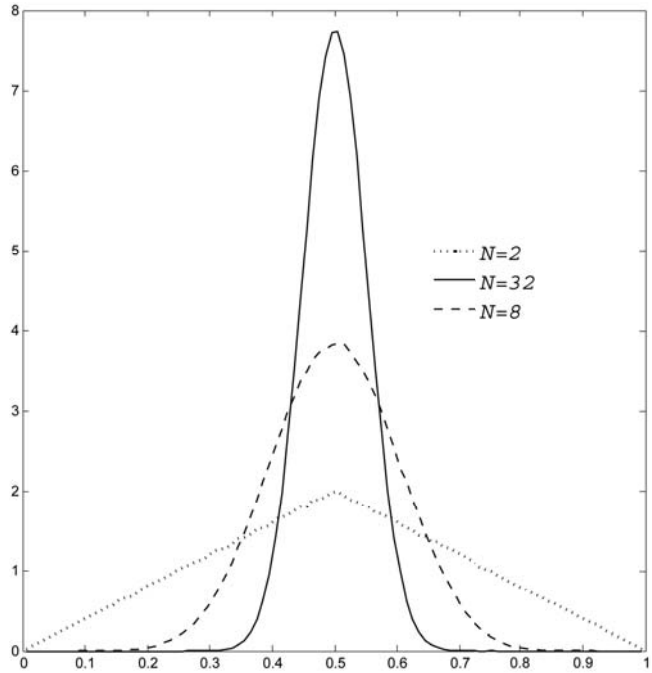

Fig. 11.1: *Direct evaluation of (11.49) for $N = 2, 8, 32$, if the underlying pdf is the uniform pdf.*

Only when $N = 32$ does the actual distribution of $\frac{1}{N}\sum_{j=1}^{N} x_j$ become Gaussian

Exercise 11.6 (a) Reproduce Figure 11.1

(b) Consider the probability $P(x) = 12\big(x - \frac{1}{2}\big)^2$ $0 \le x \le 1$ and verify the approach to the Gaussian for $\frac{1}{N}\sum_{j=1}^{N} x_j$. Hint. Consider the cumulative pdf

$$C = \int_0^x P(x)dx = 4\left(x - \frac{1}{2}\right)^3 + \frac{1}{2} \tag{11.66}$$

which since $0 \le C \le 1$ satisfies the uniform distribution and therefore consider **rand**(C) and solve for x.

The Central Limit Theorem is due to de Moivre. It has applications well beyond the statement that the arithmetic mean is distributed normally. If you expand your view any attribute due to many (summed) random variables is a candidate for a Gaussian description. A more concrete example appears if we return to the Rogue's Gallery Problem of Lecture 5 There we might assume that the gray levels of any pixel is a random variable. Then since the coefficients in the expansion of faces in terms of eigenfaces are each integrals (sums) over pixels there is an expectation of the normality of their description.