Application Example 7

DISTRIBUTION MIXTURES

One frequently deals with random variables the distribution of which depends on various factors. One example is the distribution of atmospheric parameters such as wind speed, barometric pressure, temperature, precipitation intensity etc. The distribution of these quantities varies with weather conditions and possibly time of the year. For example, wind speed has a very different distribution during inter-storm periods, in extratropical storms, and in tornadoes.

Another example is the weight of vehicles on a highway. The distribution of the weight of a generic vehicle depends on the type of vehicle (say car, truck, or bike), as well as the day (weekend versus weekday) and hour of the day. The last two factors affect the mixture of vehicle types in a traffic stream. For example, in an urban area, the fraction of trucks is higher during the night than during the day.

A third example is the strength of concrete, the distribution of which depends on the amount of water used in the mixture, the amount and type of aggregates and the curing conditions, among other factors.

Let X be the variable of interest and denote by Θ a factor that influences the distribution of X. For simplicity, suppose that Θ is discrete, with possible values θ_i, i = 1, ..., r, and denote by A_i the event $\Theta = \theta_i$. An effective technique to obtain the distribution of X is through conditioning, i.e. one first determines the conditional distribution of X under each A_i and assesses the probabilities $P[A_i]$ and then one finds the cumulative distribution of X by using the Total Probability Theorem as

$$F_X(x) = P[X \le x]$$

$$= \sum_{i=1}^{r} P[A_i] P\left[X \le x \middle| A_i\right] \qquad (1)$$

$$= \sum_{i=1}^{r} P[A_i]\, F_{X|A_i}(x)$$

This method of conditioning can be extended to the case when the distribution of X depends on several controlling factors. What one needs in this multi-factor case is the conditional distribution of X under each combination of the controlling factors and the probabilities of such combinations.

By differentiating both sides of Eq. 1 with respect to x, one obtains the corresponding expression for the probability density function, which is

$$f_X(x) = \sum_{i=1}^{r} P[A_i] f_{X|A_i}(x) \qquad (2)$$

A distribution obtained as the weighted sum of other distributions, as in Eqs. 1 and 2, is called a distribution mixture.

As an example, consider the time D it takes to commute in the morning from a suburb to downtown Boston. For given traffic and weather conditions, respectively T and W, the distribution of D might be lognormal, with mean value $m_D(T,W)$ and coefficient of variation $V_D(T,W)$. The lognormal probability density function is usually written in terms of two other parameters, $m_{\ln D}$ and $\sigma^2_{\ln D}$, which are the mean value and variance of ln(D). This density has the form

$$f_D(d) = \frac{1}{d} \frac{1}{\sqrt{2\pi}\ \sigma_{\ln D}} e^{-(\ln d - m_{\ln D})^2 / 2\sigma^2_{\ln D}} \qquad (3)$$

In turn, $m_{\ln D}$ and $\sigma^2_{\ln D}$ can be calculated from m_D and $\sigma^2_D = m^2_D V^2_D$ as

$$m_{\ln D} = \ln(m^2_D) - \frac{1}{2}\ln\left(\sigma^2_D + m^2_D\right)$$
$$\sigma^2_{\ln D} = -\ln(m^2_D) + \ln\left(\sigma^2_D + m^2_D\right) \qquad (4)$$

The reason why the distribution in Eq. 3 is called lognormal is that ln(D) has normal distribution, with mean value and variance given by Eq. 4.

Problem 7.1

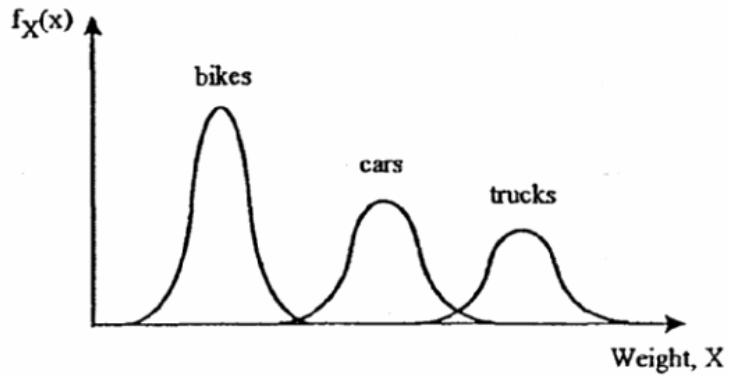
For a given suburb of Boston and a certain route, values of the distribution parameters $m_D(T,W)$ and $V_D(T,W)$ and probabilities of different (T,W) combinations might be as follows (m_D in minutes):

W	T	P[T∩W]	m_D (minutes)	V_D
good	light	0.24	25	0.05
good	normal	0.40	30	0.05
good	heavy	0.16	40	0.10
bad	light	0.02	30	0.07
bad	normal	0.08	40	0.10
bad	heavy	0.10	50	0.15
		∑ = 1.00		

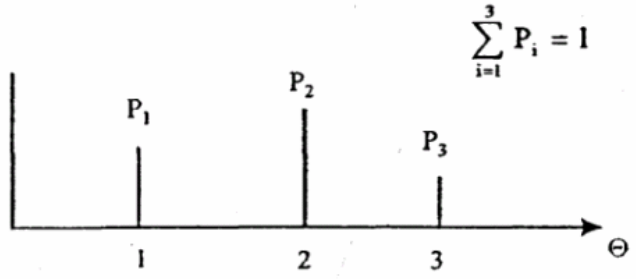
Use Eq. 2 to find the marginal probability density function of D. Plot this density function. Is it a lognormal density? Using the fact that, for any given W and T, ln(D) has normal distribution with parameters in Eq. 4 and using tables of the normal distribution, find the unconditional probability that D > 60 minutes.

An important feature of the conditioning approach is that, while the probability distribution of the mixing variable Θ may depend on the application, the conditional distributions of $(X|\Theta)$ usually do not and can be obtained from large data sets.

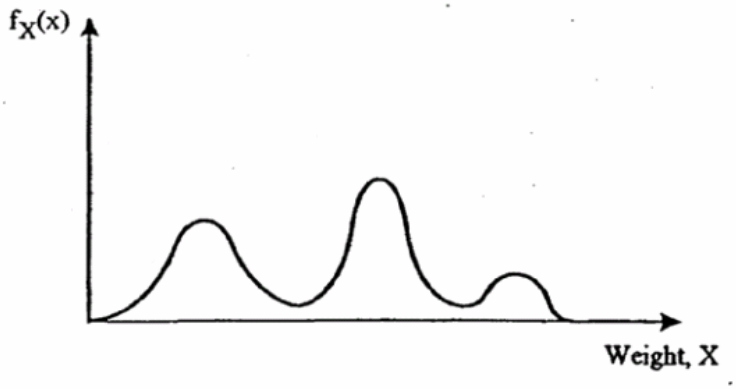
For example, consider the case when X = vehicle weight and let Θ be a discrete variable with values 1, 2, and 3 for "truck", "car", and "bike", respectively. Then $(X|\Theta)$ is the weight of a randomly chosen truck, car or bike, depending on Θ. The distribution of these conditional variables is rather constant nationwide and may be accurately estimated using extensive surveys from various parts of the country. These conditional distributions are schematically illustrated in Figure 1a. On the other hand, the traffic mixture (the distribution of the discrete variable Θ) varies from location to location and must be determined locally (Figure 1b). Doing so is a much easier task than having to determine the distribution of X exclusively from local data. Obtaining the distribution of X from Eq. 1 makes best use of all available data, at both the local and national levels.



(a) Conditional probability densities of vehicle weight.



(b) Probability that a generic vehicle is a bike, a car, or a truck.



(c) Unconditional probability density of vehicle weight obtained as a distribution mixture.

Figure 1: Illustration of distribution mixture for vehicle weight