L E C T U R E 14

§57. Poisson-Dirichlet Distribution

It is often necessary to consider random vectors

$$p = (p_1, p_2, ..., p_n) \tag{1}$$

that form a discrete probability distribution, i.e. satisfy the following conditions:

$$p_j \geq 0, \quad j = 1, 2, ..., n, \quad \sum_{j=1}^{n} p_j = 1. \tag{2}$$

For example, p_j may specify the jth of n possible species in the biological population. Random probabilistic vectors of this type also often arise in the Bayesian approach to statistics.

The simplest non-trivial example of a probability distribution on a simplex Δ_n of vectors satisfying conditions (2) is the Dirichlet distribution $D(\alpha_1, \alpha_2, ..., \alpha_n)$, the density of which (with respect to $(n-1)$-dimensional Lebesgue measure on Δ_n) is given by the formula

$$f(p_1, p_2, ..., p_n) = \frac{\Gamma(\alpha_1 + \alpha_2 + ... + \alpha_n)}{\Gamma(\alpha_1)\,\Gamma(\alpha_2)\,...\,\Gamma(\alpha_n)}\; p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} ... p_n^{\alpha_n - 1}. \tag{3}$$

Parameters α_j can take any strictly positive values, and the nature of the distribution significantly depends on these parameters. If $\alpha_j = 1$ for all j, then we have a uniform distribution on the simplex Δ_n. For large values of α_j, distribution (3) is concentrated far from the borders of the simplex, which corresponds to more or less uniform discrete distributions of vector p. On the other hand, for small values α_j, the Dirichlet distribution is concentrated near the borders of the simplex, which corresponds to extremely non-uniform distributions of p, which are some large p_j and the rest are small.

In particular, if all α_j equal to some small value α, then, from symmetry, all p_j will have the same mathematical expectation $1/n$, but probability that at least one of p_j will be much greater than the average; for which j or which value p_j will be great - only a matter of chance.

It is difficult to use directly formula (3), because the linear dependence of the components p_j. It has long been known that it is much more convenient to describe the Dirichlet distribution in terms of independent gamma quantities.

Let $Y_1, Y_2, ..., Y_n$ be independent, positive random variables, and Y_j have the following density function:

$$g_\alpha(y) = \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)}, \quad y > 0, \tag{4}$$

where $\alpha = \alpha_j$. Suppose that $Y = Y_1 + Y_2 + ... + Y_n$. It is not difficult to verify that a vector p with components

$$p_j = \frac{Y_j}{Y} \tag{5}$$

has distribution $D(\alpha_1, \alpha_2, ..., \alpha_n)$ and does not depend on Y. The proof consists in a direct calculation using the change of variables carried out using the function acting from R^n to R^n and given by the formula

$$(Y_1, Y_2, ..., Y_n) \mapsto (Y, p_1, p_2, ..., p_{n-1}).$$

As a consequence of this calculation, we obtain that the random variable Y also has a distribution (4) with the parameter

$$\alpha = \alpha_1 + \alpha_2 + ... + \alpha_n.$$

Another way to prove this is to use the Laplace transform:

$$\int_0^\infty g_\alpha(y)e^{-\theta y}\,dy = \frac{1}{(1+\theta)^\alpha}, \quad \theta > -1, \tag{6}$$

which also shows that Gamma distribution $G(\alpha)$ defined by formula (4) is infinitely divisible and has a Levi-Khinchin representation

$$\frac{1}{(1+\theta)^\alpha} = \exp\{-\alpha \int_0^\infty (1 - e^{-\theta z})z^{-1}e^{-z}\,dz\}. \tag{7}$$

Representation (7) corresponds to a subordinator, known as the Moran gamma process. This process is given by the parameters

$$\beta = 0, \quad \gamma(dz) = z^{-1}\,e^{-z}\,dz. \tag{8}$$

In this case, the increment $\varphi(t) - \varphi(s)$ has a distribution $G(t-s)$. Note, that

$$\gamma(0, \infty) = \int_0^\infty z^{-1}e^{-z}\,dz = \infty,$$

and therefore the jumps of the process φ are everywhere dense.

For $\alpha_1, \alpha_2, ..., \alpha_n > 0$ we set

$$t_0 = 0, \quad t_j = \alpha_1 + \alpha_2 + ... + \alpha_j, \quad 1 \leq j \leq n. \tag{9}$$

Then random variable $Y_j = \varphi(t_j) - \varphi(t_{j-1})$ has a distribution $G(\alpha_j)$, and all random variables Y_j are independent. Since

$$Y = Y_1 + Y_2 + ... + Y_n = \varphi(t_n),$$

we see that formula (1) in which

$$p_j = \frac{\varphi(t_j) - \varphi(t_{j-1})}{\varphi(t_n)} \tag{10}$$

sets a random vector from Δ_n with distribution $D(\alpha_1, \alpha_2, ..., \alpha_n)$.

§56. CURVILINEAR REGRESSION

So far we have studied only the case where the regression curve of Y on x is a straight line; that is, where for any given x, the mean of the distribution of Y is given by $\alpha+\beta x$. In this section we first investigate cases where the regression curve is nonlinear but where the methods of can nevertheless be applied; then we take up the problem of polynomial regression; that is, problems where for any given x the mean of the distribution of Y is given by

$$\beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p$$

Polynomial curve fitting is also used to obtain approximations when the exact functional form of the regression curve is unknown. It is common practice for engineers to plot paired data on various transformed scales such as square root or logarithm scale, in order to determine whether the transformed points will fall close to a straight line. If there is the transformation that suggests a straight-line regression equation, the necessary constants (parameters) can be estimated by applying the method of section 11.1 to the transformed data. For instance, if a set of paired data consisting of n points (x_i, y_i) "straightens out" when $\log y_i$ is plotted versus x_i, this indicates that the regression curve of Y on x is **exponential**, namely, that for any given x, the mean of the distribution of

values of Y is given by $\alpha \cdot \beta^x$. If we take logarithms to the base 10 (or any convenient base), the predicting equation $y = \alpha \cdot \beta^x$ becomes

$$\log y = \log \alpha + x \cdot \log \beta$$

and we can now get estimates of $\log \alpha$ and $\log \beta$, and hence of α and β, by applying the method of Section 11.1 to the n pairs of values $(x_i, \log y_i)$.

Before we extend the methods of preceding sections to problems involving more than one independent variable, let us point out that the curves obtained(and the surfaces we will obtain) are not used only to make predictions. They are often used also for purposes of optimization - namely, to determine for what values of the independent variable (or variables) the dependent variable is a maximum or minimum. Statistical methods of prediction and optimization are often referred to under the general heading of response surface analysis. Within the scope of this text, we will be able to introduce two further methods of response surface analysis: Multiple regression and related problems of factorial experimentation. In multiple regression, we deal with data consisting of n $(r+1)$-tuples $(x_{i1}, x_{i2}, ..., x_{ir}, y_i)$, where the x's are again assumed to be known without error while the y's are values of random variables. Data of this kind arise, for example, in studies designed to determine the effect of various climatic conditions.

As in the case of one independent variable, we first treat the problem where the regression equation is linear, namely, where for any given set of values x_1, x_2,..., and x_r, for the r independent variables, the mean of the distribution of Y is given by

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_r x_r.$$

Example 92. Observations $y_1, y_2, ..., y_n$ are assumed to come from a model with

$$E(Y_i|x_i) = \theta + 2 \ln x_i,$$

where θ is a unknown parameter and $x_1, x_2, ..., x_n$ are given constants. What is the least square estimate of the parameter θ?

SOLUTION: The sum of the squares of errors is

$$\varepsilon(\theta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \theta - 2\ln x_i)^2.$$

Differentiating $\varepsilon(\theta)$ with respect to θ, we get

$$\frac{d}{d\theta}\varepsilon(\theta) = 2\sum_{i=1}^{n} (y_i - \theta - 2\ln x_i)(-1).$$

Setting $\frac{d}{d\theta}\varepsilon(\theta) = 0$, we get

$$\sum_{i=1}^{n} (y_i - \theta - 2\ln x_i) = 0$$

which is

$$\theta = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - 2\sum_{i=1}^{n} lnx_i\right).$$

Hence the least square estimate of θ is

$$\hat{\theta} = \overline{y} - \frac{2}{n}\sum_{i=1}^{n} lnx_i.$$