The Identification of Distribution Form

Summary

A notion of distribution form with an intuitive interpretation has had a long presence in statistical writings: "the variable has a normal distribution", "the lifetime is Weibull". This paper examines this notion of distribution form and presents a formalization of it. For purposes of statistical modelling (Fraser, 1979) the objectivity of the distribution form is of central concern. Three criteria are examined which relate to this objectivity. Each is shown to require the same restriction on the determination of distribution form, namely, that the class of possible response presentations should have closure properties under composition, that is, be group like. This paper examines the foundational support for the use of the transformation model investigated in Brenner and Fraser (1979).

1. Introduction

We examine the ubiquitous notion of distribution form as encountered throughout the statistical literature. As instances consider the following. A response variable is normally distributed with unknown location $\theta$ and known scaling $\sigma_0$. The distribution is approximately Student (6) with location $\mu$ and scaling $\sigma$. The response vector $y$ has location $X\beta$ relative to the design matrix $X$ with error that is approximately normal with mean 0 and scaling $\sigma$.

Are the very clear references to the normal, and Student (6) in these examples just a convenience? Or do we have an element of substance, with clear aspects of objectivity for an investigator? We examine the notion of an identifiable distribution form and investigate three criteria that relate to its basic objectivity. Roughly expressed these criteria are observability, samplability, and compatibility. Each leads to certain explicit closure properties among the presentation (or location) functions of the response variable, essentially that these form a group.

2. The Definition of Distribution Form

Consider a few simple examples. A response variable is normally distributed with unknown location $\theta$ and known scaling $\sigma_0$. For the system under investigation there is in fact just one distribution and we can refer to points on the range of this
distribution: for example, the upper 10 % point, the median, the lower 10 % point. Correspondingly, to the degree that the model is valid, we can then refer to the same points in terms of "the normal distribution" or more specifically in terms of each of the possible normal distributions for the response.

Consider a distribution that is approximately Student (6) with location μ and scaling σ. For the system being investigated there is just one distribution and we can refer to points on its range as in the preceding example the upper 10 % point, the median, the lower 10 % point. And again, to the degree that the model is valid, we can describe these same points in terms of "the Student (6) distribution" or in terms of each of the possible Student (6) response presentations.

As a third example consider a simple factor-analysis situation involving a bivariate response. We suppose that the background information supports a single linear factor with say a symmetric triangular distribution convolved with a rotationally symmetric Student (6) error distribution. For the response presentation we suppose that this can appear as any positive linear deformation. There is just one distribution for the application and to the degree of validity of the model we can refer to points on its range directly in terms of the same points for "the standard case" or in terms of the same points for each of the possible response possibilities.

We now formalize the notion of distribution form. Let Y be the sample space for the response variable and let τ in T be a parameter indexing the various possible response distributions. Let $P = \{p\}$ be a space of labels for referring to the points of the distribution, and correspondingly, for each $\tau \in T$, let $p_\tau:Y \rightarrow P$ be a one-one onto function that presents the label $p = p_\tau(y)$ for each point $y$ in the sample space of the $\tau$-distribution.

The preceding presents identifying labels for points on each of the possible distributions. It does not, however, tell us how the labels were engendered or derived from the physical ingredients of the problem.

Pursuing this, we now consider points in the response space and a means for "scanning" the various possible distributions. Let $Y_e$ be an examination space for this purpose and let $T = \{t\}$ now denote a class of one-one transformations $t:Y_e \rightarrow Y$; we scan possibilities on the response space $Y$ by examining them on $Y_e$. Appropriately the elements of $T$ will be referred to as scans.

Note that we are using the same symbol $T$ for both the parameter space and the class of scans where a scan $t$ examines a distribution for which $\tau = t$. Of course the transformations $T$ must necessarily embody the same continuity and differentiability characteristics and properties that are either implicit or explicit in the modelling of the particular physical system.