Remarks on how to describe a (parametric) statistical model

R. Labouriau

2020

I am aware that describing even a simple statistical model precisely is not an easy task, specially for a public of non-mathematicians. Anyway, I do believe that it is important to learn to describe the statistical models one uses, because in doing so we are sure that we understand the mathematical objects we are using and we are in a position to critically judge them and their applicability. When describing a model, the first step is to establish the basic notation. In the (simple) contexts we study in this course, we usually start by stating the notation for the random variables that represent the observed responses (which are viewed as realisation of those random variables). To be more concrete, let us consider the situation where we are modelling differences between two treatments, say A and B, and that the design of the study we are considering is such that we have, say 3 observations for each treatment, all in all 6 observations. In this case, we might setup the basin notation by defining 6 random variables that will represent the 6 results. We will label these six random variables by appending to each of them two sub-indices, say t and r, representing the treatment and the repetition number respectively. In this way, our notation will allow us to uniquely identify each of the six observations and, at the same time, inform us which treatment was allocated to each of the six observations. To set up the notation we may write: *"For $t = A, B$ and $n = 1, 2, 3$, Y_{tr} is the random variable that represents the n^{th} repetition of the t^{th} treatment."*. Once done that me have established the very

basic notation. Next, we should state that the observations are independent. To do that we may write:
"We assume that $Y_{A1}, Y_{A2}, Y_{A3}, Y_{B1}, Y_{B2}, Y_{B3}$ *are independent"*
or in short,
"We assume that $Y_{A1}, \ldots, Y_{B3}$ *are independent".*
Note that this statement refers to all the observations (or more precisely, the 6 random variables representing the 6 observations).

Suppose now that we want to state that the 6 random variables in play are normally distributed, have the same variance and an expectation depending on the treatment (being equal to all the observations that received the same treatment). In this case, it suffices to write:
"For $t = A, B$ *and* $r = 1, 2, 3$ *,* $Y_{tr} \sim N(\mu_t, \sigma^2)$*"* .
Note that according to what is written, for each possible value of t (say A or B) and each possible value of r (say $1, 2$ or 3) we have that Y_{tr} is normally distributed; moreover, the variance each Y_{tr} is σ^2 (here σ^2 is an unknown quantity) and the expectation of Y_{tr} (denoted $E(Y_{tr})$) is equal to μ_t , which is a quantity depending on the treatment applied to the corresponding observation as indicated by the subindex. Since the statement "$Y_{tr} \sim N(\mu_t, \sigma^2)$" is valid for each possible value of t (A or B) and each possible value of r ($1, 2$ or 3), then in fact me have specified that all the random variables representing the observations are normally distributed (so we expressed the here the assumption of normality). In the same way, we expressed that all the random variables in play have the same variance (so we implicitly stated the variance homogeneity assumption) and that the random variables representing the observations that received the same treatment are all equal (homogeneity of the means)
Summing up, one way to describe the model in play precisely is to write: *"For* $t = A, B$ *representing the treatments and* $r = 1, 2, 3$ *labelling the repetitions ,* Y_{tr} *represents the value of the* r^{th} *repetition of the experimental units that revived the* t^{th} *treatment. It is assumed that* $Y_{A1}, \ldots, Y_{B3}$ *are independent. Moreover,*

$$Y_{tr} \sim N(\mu_t, \sigma^2), \text{ for } t = A, B \text{ and } r = 1, 2, 3."$$