Note on statistic and random variable

stma

March 15, 2024

We realized that in a mathematical point of view, there seems to be some ambiguity with the rigorous treatment of concepts in statistics. While we are standing at where rigorousness does not always help, we put a faith on clarification being favourable to avoid a violated use of terminology.

Here we referred the definitions of *statistic* and *random variable* to two books for comparison. Both of texts are classical and well-reputed, and we respect for their own manners.

Definition. Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the random variable or random vector $Y = T(X_1, \ldots, X_n)$ is called *statistic*. [1]

Definition. A *statistic* is a random variable or random vector that is measurable function T from $(\mathfrak{X}, \mathscr{A})$ into some $(\mathfrak{Y}, \mathscr{B})$. [2]

Definition. A *random variable* is a function from a sample space S into the real number. [1]

Definition. A *random variable* is a measurable real-valued function $X : (\mathfrak{X}, \mathscr{A}) \to (\mathbb{R}, \mathfrak{B})$ from a measurable space into the real number. [2]

Restricted to statistical application, we could find no reasonable explanation of taking non-real numbers for value space of *random variable*, and the measurability is almost always assumed to avoid unnecessary complexity, though they may be implicit.

For the purpose of defining *statistic* distinct from *random variable*, we assume that a finite series of random variables are provided at the first place. Furthermore, despite a statistic is not restricted to a real (vector) valued function, an ordinary example shows that a collection of arithmetic operations are to be equipped in the parameter spaces, as in $s(\mathbf{X}) = \langle \frac{1}{n}(X_1 + \ldots + X_n), X_1 \cdot X_n \rangle$, for example.

Hence we may be at a good starting point to define a statistic by (the image of) a function:

$$s : [\mathfrak{X}, \mathbb{R}]^I \to [\mathfrak{X}, \mathfrak{Y}];\ \langle X_i \rangle_{i \in I} \mapsto \langle s_j(X_1(-), \ldots, X_n(-)) \rangle_{j \in J},$$

where I and J are finite sets with $|I| = n$, $|J| = m$.

Provided that a series of random variables are given, the essential data to determine a statistic is how the real numbers are transformed and arranged, denoted by s_j, which can be seen explicitly in the following.

$$\begin{tikzcd}
\mathfrak{X} \arrow[d, dotted, "X^I"'] \arrow[dr, "X_i"] & \\
\mathbb{R}^n \arrow[r, "\mathrm{pr}_i"] & \mathbb{R}
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathfrak{X} \arrow[r, dotted, "s"] \arrow[d, "X^I"] & \mathfrak{Y} \arrow[r, equal] \arrow[d, "\mathrm{pr}_j"] & \prod_{j \in J} \mathfrak{Y}_j \\
\mathbb{R}^n \arrow[r, "s_j"] & \mathfrak{Y}_j .
\end{tikzcd}$$

The induced dotted arrows are universal in the category **Meas** of measurable spaces.

Note that with this definition, it is consistent both for why a statistic is represented as a "function of random variables" in practice and for withstanding the use in two contrasting contexts [1] and [2]; statistic s is a random variable or a random vector when $\mathfrak{Y}_j = \mathbb{R}$ for all $j \in J$, which factors through real transformation (a function between measurable spaces, not necessarily measurable, is called *transformation*).

References

[1] George Casella and Roger Berger. *Statistical Inference*. Duxbury Resource Center, June 2001. ISBN: 0534243126.

[2] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Third. Springer Texts in Statistics. New York: Springer, 2005, pp. xiv+784. ISBN: 0-387-98864-5.