

Identifiability of certain family of distributions based on their first moment and c-statistic

Mohsen Sadatsafavi

2024-01-15

Consider a parametric family of probability distributions with support on $[0,1]$, with the following characteristics:

- the CDF is strictly monotonical;
- the distribution is quantile-identifiable: being fully identifiable by knowing a pair of its quantile values.

We note that common two-parameter distributions for probabilities, such as beta ($\pi \sim \text{Beta}(\alpha, \beta)$), logit normal ($\text{logit}(\pi) \sim \text{Normal}(\mu, \sigma^2)$, where $\text{logit}(\pi) := \log(\pi/(1 - \pi))$) and probit-normal ($\Phi^{-1}(\pi) \sim \text{Normal}(\mu, \sigma^2)$ where $\Phi(x)$ is the standard normal CDF) satisfy the above criteria. All these distributions are smooth. The quantile-identifiability of the beta distribution is proven in Shih et al (doi:10.1080/00949655.2014.914513). For the logit-normal and probit-normal distributions, it is immediately deduced from the monotonical link to the normal distribution and the quantile-identifiability of the latter.

Lemma: For a family of probability distributions with the above characteristics, the combination of expected value and c-statistic uniquely identifies the distribution.

Proof

Let F be the CDF from the family of distributions of interest. Let m be its first moment, and c the c-statistic, defined as $c := P(\pi_2 > \pi_1 | Y_2 > Y_1)$ where $\pi_i \sim F$ and $Y_i \sim \text{Bernoulli}(\pi_i)$ a realization of response value given the probability. We shall prove that F is uniquely identifiable from $\{m, c\}$.

First, we apply Bayes' theorem to the definition of c : $c = P(Y_2 > Y_1 | \pi_2 > \pi_1) \frac{P(\pi_2 > \pi_1)}{P(Y_2 > Y_1)} = P(Y_2 = 1, Y_1 = 0 | \pi_2 > \pi_1) \frac{0.5}{P(Y_2=1, Y_1=0)} = \frac{\mathbf{E}\pi_{max}(1 - \mathbf{E}\pi_{min})}{2m(1-m)} = \frac{\mathbf{E}\pi_{max}(1 - 2m + \mathbf{E}\pi_{max})}{2m(1-m)}$,

where π_{max} and π_{min} are, respectively, the maximum and minimum of a pair of independent RVs from F . $\mathbf{E}\pi_{max}(1 - 2m + \mathbf{E}\pi_{max})$ is monotonical with respect to $\mathbf{E}\pi_{max}$ in its range $([m, 1])$. Further, let $f(x) := dF(x)/dx$ be the PDF of F . We have $\mathbf{E}\pi_{max} = \int_0^1 2xf(x)F(x)dx = xF^2(x)|_0^1 - \int_0^1 F^2(x)dx = 1 - \int_0^1 F^2(x)dx$; i.e., $\mathbf{E}\pi_{max}$ is monotonically related to $\int_0^1 F^2(x)dx$. As such, the goal is achieved by showing that $\{m, \int_0^1 F^2(x)dx\}$ uniquely identifies F .

We show this by proving that two different CDFs F_1 and F_2 with the same m cannot have the same $\int_0^1 F^2(x)dx$.

To proceed, we note that for probability distributions with support on $[0,1]$, the equality of means indicates the equality of the area under CDFs, as (by integration by parts - proof is by Harry Lee) $m = \int_0^1 xf(x)dx = xF(x)|_0^1 - \int_0^1 F(x)dx = 1 - \int_0^1 F(x)dx$.

Given that both CDFs are anchored at $(0,0)$ and $(1,1)$, are strictly monotonical, and have the same area under the CDF but are not equal at all points, they must cross. However, they can only cross once, given the quantile-identifiability requirement (if they cross two or more times, any pairs of quantiles defined by the crossing points would fail to identify them uniquely).

Let z be the unique crossing point of the two CDFs, and let $z^* = F_1(z) = F_2(z)$ be the CDF value at this point. We break $\int_0^1 (F_1^2(x) - F_2^2(x))dx$ into two parts, after removing the only three points $x \in \{0, z, 1\}$ where $F_1(x) - F_2(x) = 0$:

$$\int_0^1 (F_1^2(x) - F_2^2(x))dx = \int_{x \in (0,z)} (F_1^2(x) - F_2^2(x))dx + \int_{x \in (z,1)} (F_1^2(x) - F_2^2(x))dx = \int_{x \in (0,z)} (F_1(x) - F_2(x))(F_1(x) + F_2(x))dx + \int_{x \in (z,1)} (F_1(x) - F_2(x))(F_1(x) + F_2(x))dx.$$

Without loss of generality, assume we label F s such that $F_1(x) > F_2(x)$ when $x \in (0, z)$. In this region, $F_1(x) - F_2(x) > 0$, and (due to F s monotonically increasing) $0 < F_1(x) + F_2(x) < F_1(z) + F_2(z) = 2z^*$. As such, replacing $F_1(x) + F_2(x)$ by the larger positive quantity $2z^*$ will increase this term. As well, in the $x \in (z, 1)$ region, $F_1(x) - F_2(x) < 0$, and $0 < F_1(x) + F_2(x) < F_1(z) + F_2(z) = 2z^*$. As such, replacing $F_1(x) + F_2(x)$ by the smaller positive quantity $2z^*$ will also increase this term. Therefore we have

$$\int_0^1 (F_1^2(x) - F_2^2(x))dx < 2z^* (\int_{x \in (0,z)} (F_1(x) - F_2(x))dx + \int_{x \in (z,1)} (F_1(x) - F_2(x))dx),$$

and the term on the right hand side is zero because of the equality of the area under the CDFs. Therefore, $\int_0^1 (F_1^2(x) - F_2^2(x))dx < 0$.