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

Mohsen Sadatsafavi

2024-01-18

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 ($\pi \sim Beta(\alpha, \beta)$), logit normal ($logit(\pi) \sim Normal(\mu, \sigma^2)$, where $logit(\pi) := log(\pi/(1-\pi))$) and probit-normal ($\Phi^{-1}(\pi) \sim Normal(\mu, \sigma^2)$ where $\Phi(x)$ is the standard normal CDF) satisfy the above criteria. All these distributions have strictly monotonical CDFs. 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 its c-statistic, defined as $c := P(\pi_2 > \pi_1 | Y_2 = 1, Y_1 = 0)$ where $\pi_i \sim F$ and $Y_i \sim Bernoulli(\pi_i)$. c is the probability that a random draw from the distribution of π among 'cases' (those with $Y = 1$) is larger than a random draw from its distribution among 'controls' (those with $Y = 0$). We shall prove that F is uniquely identifiable from $\{m, c\}$.

First, applying the Bayes' rule to the distribution of π among cases ($P(\pi|Y = 1)$) and controls ($P(\pi|Y = 0)$) reveals that the former has a PDF of $xf(x)/m$ and the latter $(1-x)f(x)/(1-m)$, where $f(x) := dF(x)/dx$ is the PDF of F. Thus we have:

$m(1-m)c = \int_0^1 [xf(x)\int_0^x (1-y)f(y)dy]dx = \int_0^1 [xf(x)\int_0^x f(y)dy]dx - \int_0^1 [xf(x)\int_0^x yf(y)dy]dx = \int_0^1 xf(x)F(x)dx - \int_0^1 g(x)G(x)dx$, where $g(x) = xf(x)$ and $G(x) = \int_0^x g(y)dy$. Integrating by parts for both integrals results in

$m(1-m)c = \frac{1}{2}xF^2(x)|_0^1 - \frac{1}{2}\int_0^1 F^2(x)dx - \frac{1}{2}G^2(x)|_0^1 = \frac{1}{2} - \frac{1}{2}\int_0^1 F^2(x)dx - m^2/2$

i.e., c 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 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 Fs 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 Fs 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, $\int_0^1 (F_1^2(x) - F_2^2(x))dx < 0$.

Remarks

This proof can easily be extended to other metrics of central tendency and discrimination. For example, if instead of mean, the median (or any quantile) is available, the proof immediately applies because the quantile becomes the sole crossing point of any two Fs, and again the equivalence of $\int_0^1 F^2(x)dx$ guarantees the equivalence of the two distributions.

Similarly, if one knows the Gini index of a distribution, then due to the unique relationship between the c-statistic and gini[REF], the proof is applicable.

Finding the solutions

We note that for some distributions including beta and probit-normal, knowing m immediately solves for one parameter. As such, numerical optimization reduces to one-dimensional optimization. For beta distribution, the relationship is algebraic. For probit-normal, it requires evaluating the Φ^{-1} which cannot be expressed in closed-form, nonetheless there are numerically tuned algorithms widely implemented in several software

For logit-normal distribution, where there are no