BASIC PROBABILITY: THEORY
Master of Logic, University of Amsterdam, 2016
TEACHERS Christian Schaffner and Philip Schulz TA Bas Cornelissen

Some notes on sufficient statistics

The exercise In exercise 1 of this weeks' board questions, you were given a set $x_1^n = (x_1, \dots, x_n)$ of n i.i.d. observations that were all geometrically distributed. So they are observations of RV's $X_1, \dots X_n$ where

$$P(X_i = x_i \mid \Theta = \theta) = \text{Geom}(x_i \mid \theta) = (1-\theta)^{x_i}\theta.$$

We had to show that $t := T(x_i) = \sum_{i=1}^n x_i$ is a sufficient statistic.

What does that even mean? By the Factorization Theorem, it suffices to find two functions $g(\theta, t)$ and $h(x, t)$ such that

$$P(X_1^n = x_1^n \mid \Theta = \theta) = g(\theta, t) \cdot h(x_1^n, t). \tag{1}$$

So what is our joint distribution? For legibility, we'll drop the random variables and write e.g. $P(x_1^n \mid \theta) := P(X_1^n = x_1^n \mid \Theta = \theta)$. By independence this is:

$$P(x_1^n \mid \theta) = \prod_{i=1}^n (1-\theta)^{x_i}\theta = (1-\theta)^{\sum_{i=1}^n x_i} \cdot \theta^n. \tag{2}$$

The answer Now observe that this is simply $(1-\theta)^t\theta^n$, so when we choose $g(\theta, t) := (1-\theta)^t\theta^n$ and $h(x, t) := 1$ we have found a factorization of the joint. By the Factorization Theorem, t is thus a sufficient statistic.

But why? True as that may be, this feels a bit unsatisfactory. After all, the idea was that given the value of the sufficient statistic, it should be possible to write the PMF without using the parameter. The Factorization Theorem told you *that* this is possible, but it doesn't tell you *how* to do it.

Or does it? In fact, the proof does. We essentially have to expand the conditional distribution of x_1^n given t and θ:

$$P(x_1^n \mid t, \theta) = \frac{p(x_1^n, t \mid \theta)}{p(t \mid \theta)} = \frac{p(x_1^n \mid \theta)}{p(t \mid \theta)} = \frac{p(x_1^n \mid \theta)}{\sum_{z_1^n : t(z_1^n) = t} p(z_1^n, t \mid \theta)}. \tag{3}$$

In the second equality we used the fact that t is a deterministic function of x_1^n so the probability of x_1^n and t is exactly the same as the probability of x_1^n. In the third equality we used a little trick, writing a marginal as a marginalized joint.

Recall that we actually had a factorization of $p(x_1^n \mid \theta)$, which we can now fill in in (3) to get

$$P(x_1^n \mid t, \theta) = \frac{g(\theta, t) \cdot h(x_1^n, t)}{\sum_{z_1^n} g(\theta, t) \cdot h(z_1^n, t)} = \frac{h(x_1^n, t)}{\sum_{z_1^n} h(z_1^n, t)} \tag{4}$$

And since we know $h(x_1^n, t) = 1$, we can actually calculate this as

$$P(x_1^n \mid t, \theta) = \frac{1}{\sum_{z_1^n : T(z_1^n) = t} 1} = \frac{1}{|\{z_1^n : T(z_1^n) = t\}|} \tag{5}$$

— if you manage to count the set in the denominator, that is.

The lesson Taking a step back, consider the conditional probability of x_1^n given t, as expressed in the first equality of (3). That is the thing we want to write without using θ, and we can do so if we somehow manage to cancel out the θ in the numerator against the θ's in the denominator. This is precisely what happened in the last step of (4). Working with actual distributions, this might however be very difficult. Also, finding the actual distribution *without* the θ need not be easy: you have to deal with the sum in (4).

What else should now be clear? For example: if we have data x_1^n and y_1^n with the same sufficient statistic $T(x_1^n) = T(y_1^n) = t$, drawn from two distributions, with parameters θ and θ', then by (4)

$$P(x_1^n \mid t, \theta) = P(y_1^n \mid t, \theta').$$

We can also say something about the original distributions, not conditioned on t. The distributions of x_1^n and y_1^n differ from another only in the normalizing constant. We can make that more explicit as follows:

$$
\begin{aligned}
P(x_1^n \mid \theta) &= P(t \mid \theta) \cdot P(x_1^n \mid t, \theta) && (6)\\
&= P(t \mid \theta) \cdot P(y_1^n \mid t, \theta') && (7)\\
&= \frac{P(t \mid \theta)}{P(t \mid \theta')} \cdot P(t \mid \theta') \cdot P(y_1^n \mid t, \theta') && (8)\\
&= \frac{P(t \mid \theta)}{P(t \mid \theta')} \cdot P(y_1^n \mid \theta') && (9)
\end{aligned}
$$