Sampling a contact relation within the same age interval

Stijn Vansummeren

December 4, 2020

Assume that we are given a set D of n persons, all with the same age. Further assume that we know the contact rate r, that is, we know that on average, every person in D will meet with r other people in D.

Our goal is to use a sampling-based algorithm to derive a *contact relation* over D. Formally a *contact relation over* D is a binary relation $R \subseteq D \times D$ such that:

- if $(a, b) \in R$ then $a \neq b$ (people don't meet themselves); and
- if $(a, b) \in R$ then also $(b, a) \in R$ (having a contact is symmetric).

The following algorithm allows us to derive a (random) contact relation R over D. For now, assume that $p \in [0, 1]$ is given. We will determine its correct value below.

Sampling algorithm.

1. Initialize R to be empty.
2. For each x in D:
   (a) Draw $K_x \sim B(n - 1, p)$
   (b) Let Y_x be a sample of K_x elements from $D \setminus \{x\}$.
   (c) For every $y \in Y_x$, add (x, y) and (y, x) to R

Here, $B(n - 1, p)$ is the Binomial distribution with parameters $n - 1$ and p.

It is straightforward to verify that the resulting relation R will be valid a contact relation.[1]

[1] For the purpose of simulating epidemics we of course don't want to actually build R, but instead immediately process the pairs (x,y) and (y,x) of R as soon as they are generated. By explicitly building R in these notes, we can reason over the correctness of the sampling algorithm.

Let ρ_a denote the expected number of contacts that a person $a \in D$ has in the generated contact relation R. Our goal now is to derive a value for p such that ρ_a equals the contact rate r. Hereto, we reason as follows.

The probability space. We will need to reason over the probability space (Ω, P), where the set Ω of possible outcomes is the set of all possible runs of the algorithm. Concretely, each outcome $\omega \in \Omega$ is hence a tuple of the form

$$\omega = (R^\omega, (K_x^\omega)_{x\in D}, (Y_x^\omega)_{x\in D}),$$

that records the contact relation R^ω constructed by the run ω, as well as the family of natural numbers $(K_x^\omega)_{x\in D}$ chosen during the run and the family of samples $(Y_x^\omega)_{x\in D}$.

In what follows, if $x \in D$ then we write K_x for the discrete random variable that, given outcome $\omega \in \Omega$ returns the actual value K_x^ω chosen in line (2.a). The random variables R and Y_x are defined similarly. Furthermore, if x and b are elements of D then we overload notation and denote by the expression $x \in Y_b$ the event consisting of all outcomes ω for which $x \in Y_b(\omega)$. That is,

$$(x \in Y_b) = \{\omega \in \Omega \mid x \in Y_b(\omega)\}.$$

For a given event $\alpha \subseteq \Omega$, $\mathbf{1}_{\{\alpha\}}$ denotes the indicator random variable, i.e., the binary random variable such that

$$\mathbf{1}_{\{\alpha\}}(\omega) = \begin{cases} 1 & \text{if } \omega \in \alpha \\ 0 & \text{otherwise .} \end{cases}$$

Counting the number of contacts in one particular outcome. Consider an arbitrary outcome $\omega = (R^\omega, (K_x^\omega)_{x\in D}, (Y_x^\omega)_{x\in D})$. What is the number of contacts that person a has in R^ω? Because R^ω is symmetric by definition, the number of contacts of a person a is simply the number of times that a occurs in the first column in R^ω, i.e., it is the number of tuples in $(a, b) \in R^\omega$ with $b \in D \setminus \{a\}$. Note that such a tuple (a, b) may have been added to R^ω because $b \in Y_a^\omega$, or because $a \in Y_b^\omega$, or both. Therefore, the number of times that a occurs in the first column in R^ω equals

$$\begin{aligned}
\sum_{b\in Y_a^\omega} 1 + \sum_{\substack{b\in D\setminus\{a\}\\ a\in Y_b^\omega}} 1 - \sum_{\substack{b\in D\setminus\{a\}\\ a\in Y_b^\omega, b\in Y_a^\omega}} 1 &= |Y_a^\omega| + \sum_{\substack{b\in D\setminus\{a\}\\ a\in Y_b^\omega}} 1 - \sum_{\substack{b\in D\setminus\{a\}\\ a\in Y_b^\omega, b\in Y_a^\omega}} 1 \\
&= K_a(\omega) + \sum_{b\in D\setminus\{a\}} \mathbf{1}_{\{a\in Y_b\}}(\omega) - \sum_{b\in D\setminus\{a\}} \mathbf{1}_{\{a\in Y_b, b\in Y_a\}}(\omega).
\end{aligned}$$

Determining the expected number of occurrences. The expected number of occurrences of a is then

$$\begin{aligned}
\rho_a &= E[K_a + \sum_{b \in D \setminus \{a\}} \mathbf{1}_{\{a \in Y_b\}} - \sum_{b \in D \setminus \{a\}} \mathbf{1}_{\{a \in Y_b, b \in Y_a\}}] && (1)\\
&= E[K_a] + \sum_{b \in D \setminus \{a\}} E[\mathbf{1}_{\{a \in Y_b\}}] - \sum_{b \in D \setminus \{a\}} E[\mathbf{1}_{\{a \in Y_b, b \in Y_a\}}] && (2)\\
&= E[K_a] + \sum_{b \in D \setminus \{a\}} P(a \in Y_b) - \sum_{b \in D \setminus \{a\}} P(a \in Y_b, b \in Y_a) && (3)\\
&= (E[K_a] + \sum_{b \in D \setminus \{a\}} P(a \in Y_b) - \sum_{b \in D \setminus \{a\}} P(a \in Y_b)\, P(b \in Y_a) && (4)\\
&= (n-1)p + \sum_{b \in D \setminus \{a\}} p - \sum_{b \in D \setminus \{a\}} p\, p && (5)\\
&= (n-1)p + (n-1)p - (n-1)p^2 && (6)\\
&= 2(n-1)p - (n-1)p^2. && (7)
\end{aligned}$$

Here, (2) is due to linearity of expectation; (3) is because $E[\mathbf{1}_{\{\alpha\}}] = P(\alpha)$ for every indicator variable $\mathbf{1}_{\{\alpha\}}$ and every event α; (4) is because Y_a and Y_b are independent random variables; and (5) is because:

- $K_a \sim B(n-1, p)$ by definition of the algorithm; therefore K_a follows a Binomial distribution and as such it is known that $E[K_a] = (n-1)p$;
- $P(b \in Y_a) = P(a \in Y_b) = p$, as we will formally demonstrate futher below.

Determining p. We want ρ_a to equal the contact rate r.

$$\begin{aligned}
& \rho_a = r \\
\iff \quad & 2(n-1)p - (n-1)p^2 = r \\
\iff \quad & p^2 - 2p + \frac{r}{n-1} = 0
\end{aligned}$$

As such, we can determine p by finding the roots of a quadratic polynomial. There are at most two such roots, given by

$$p_1 = \frac{2 + \sqrt{4 - 4\frac{r}{n-1}}}{2} \qquad \text{and} \qquad p_2 = \frac{2 - \sqrt{4 - 4\frac{r}{n-1}}}{2}.$$

Note, however, that these roots only exist if the quantity under the square root is positive, i.e., if

$$\begin{aligned} & 4 - 4\frac{r}{n-1} \geq 0 \\ \iff \quad & 1 - \frac{r}{n-1} \geq 0 \\ \iff \quad & 1 \geq \frac{r}{n-1} \\ \iff \quad & n - 1 \geq r \end{aligned}$$

This is expected, since we can never hope to achieve a contact rate of r if D has strictly less than $r+1$ persons. (Every person in D can meet with at most $n-1$ persons.)

Further note that, of the two possible roots p_1 and p_2, it suffices to consider only p_2. Indeed, we are only interested in those values of p that lie within the interval $[0, 1]$. Note that $p_1 \geq 1$, and when $p_1 = 1$ (which is achieved when $r = n-1$) then $p_2 = p_1 = 1$ (i.e., in that case there is only a single root). Therefore, the correct value of p is

$$p = \frac{2 - \sqrt{4 - 4\frac{r}{n-1}}}{2}.$$

Why the probability that $b \in Y_a$ is p. We next prove formally that, for any $a \in D$ and any $b \in D \setminus \{a\}$ the probability that $b \in Y_a$ in step (2.b) of the algorithm, is p.

For the sake of the development, fix $a \in D$ and $b \in D \setminus \{a\}$. Note that, in particular, this hence implies that $n \geq 2$.

We will need the following known equalities.

- If X is a random variable such that $X \sim B(n-1, p)$, then the probability that X equals a specific value $k \in \mathbb{N}$ is as follows.

$$P(X = k) = \begin{cases} \binom{n-1}{k} p^k (1-p)^{n-1-k} & \text{if } 0 \leq k \leq n-1 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

- For any real values $x, y \in \mathbb{R}$ and any natural number $m \in \mathbb{N}$ it holds that

$$(x + y)^m = \sum_{l=0}^{m} \binom{m}{l} x^l y^{m-l} \tag{9}$$

Now note that $P(b \in Y_a, K_a = k)$ is the probability that $b \in Y_a$ and the value of K_a drawn in step (2.a) is k. Equivalently, $P(b \in Y_a, K_a = k)$ is the probability that both (i) $b \in Y_a$ and (ii) $|Y_a| = k$.

Then the quantity that we seek, $P(b \in Y_a)$, is simply $P(b \in Y_a, K_a)$, marginalized over the possible values of K_a:

$$P(b \in Y_a) = \sum_{k \in \mathbb{N}} P(b \in Y_a, K_a = k) \tag{10}$$

$$= \sum_{k \in \mathbb{N}} P(b \in Y_a \mid K_a = k) P(K_a = k) \tag{11}$$

$$= \sum_{k=0}^{n-1} P(b \in Y_a \mid K_a = k) P(K_a = k) \tag{12}$$

The second equality is by definition of conditional probability; the third is because $K_a \sim B(n-1, p)$ and therefore $P(K_a = k) = 0$ for $k \geq n$ by (8).

Let us next calculate the conditional probability $P(b \in Y_a \mid K_a = k)$. This is the probability that $b \in Y_a$ when $K_a = |Y_a| = k$. I.e., it is the probability that b is in a random k-sized subset of $D \setminus \{a\}$. Now note:

- there are $\binom{n-1}{k}$ possible subsets of $D \setminus \{a\}$ of size k;
- when $k = 0$, there is no subset of $D \setminus \{a\}$ that contains b, since the only subset of size 0 is the empty set.
- when $k \geq 1$ there are $\binom{n-2}{k-1}$ possible sets $Y_a \subseteq D \setminus \{a\}$ of size k that contain b.

  Indeed, observe that every set $Y_a \subseteq D \setminus \{a\}$ that contains b is of the form $Y_a = \{b\} \cup Y'_a$ with $Y'_a \subseteq D \setminus \{a, b\}$. If Y_a is of size k, then Y'_a is of size $k - 1$. Therefore, the number of k-sized sets Y_a with $b \in Y_a$ equals the number $k - 1$ sized subsets of $Y'_a \subseteq D \setminus \{a, b\}$, of which there are $\binom{n-2}{k-1}$.

Combining these three points, we conclude that, for every $0 \leq k \leq n - 1$:

$$P(b \in Y_a \mid K_a = k) = \begin{cases} 0 & \text{if } k = 0 \\ \frac{\binom{n-2}{k-1}}{\binom{n-1}{k}} & \text{if } 1 \leq k \leq n-1. \end{cases} \tag{13}$$

We now continue our derivation of $P(b \in Y_a)$, and plug in (8) and (13) in

(12):

$$
\begin{aligned}
P(b \in Y_a) &= \sum_{k=0}^{n-1} P(b \in Y_a \mid K_a = k) P(K_a = k) && (14) \\
&= \sum_{k=1}^{n-1} \frac{\binom{n-2}{k-1}}{\binom{n-1}{k}} \binom{n-1}{k} p^k (1-p)^{n-1-k} && (15) \\
&= \sum_{k=1}^{n-1} \binom{n-2}{k-1} p^k (1-p)^{n-1-k} && (16) \\
&= \sum_{l=0}^{n-2} \binom{n-2}{l} p^{l+1} (1-p)^{n-1-(l+1)} && (17) \\
&= p \sum_{l=0}^{n-2} \binom{n-2}{l} p^l (1-p)^{n-2-l} && (18) \\
&= p(p + (1-p))^{n-2} && (19) \\
&= p1^{n-2} && (20) \\
&= p && (21)
\end{aligned}
$$

Here, (17) is by re-indexing with $l = k - 1$, and (19) is by application of (9) with $x = p$, $y = 1 - p$ and $m = n - 2$.