1. Family sizes

Alex Average started work as a statistician at the ONSO (the Office for National Statistics for Outopia) on 1st January 2020. Alex's first job was to find the distribution of family sizes in Outopia. After discussion, Alex decided to define the family of an individual, x, as meaning the number of people (including x) who are related to x and live in the same dwelling as x.

Alex decided to initially survey a random sample of 20 individuals and ask their family size, recording the answers as $F_1, \ldots, F_{20}$.

The survey results were $6, 1, 6, 6, 5, 3, 4, 3, 2, 4, 5, 6, 6, 5, 6, 5, 5, 6, 3, 6$. This gives a mean family size of 4.35. Alex is very surprised by this, since the last full census of Outopia (in December 2019) found an average family size of 3.75.

Explain what's gone wrong.

Hint Try taking a random sample of size 3 from a population with just two families of sizes 1 and 5 (you can use a single die to do this).

Answer The problem is that Alex's sampling method is *size-biased*; that is the chance of a *family* being sampled is biased by the size of the family. A family of size 6 contributes six people to the population of Outopia, whereas a family of size 2 only contributes two. So Alex's method will over-represent larger families and so will overestimate average family size.

Extensions

(1) Obtain an estimate of Outopia's average family size from Alex's data.

Hint If the *proportion* of families in Outopia of size j is p_j, what is the proportion, s_j, of people that are in a family of size j?

Answer Because of the size-bias, s_j should be proportional to jp_j. So to scale these numbers to add to one we must divide by $\mu = \sum_i ip_i$ – the mean family size.

For each i, we have an estimate $\hat{s}_i$ of $s_i = \frac{ip_i}{\mu}$ from Alex's sample. So, if we take $\sum \frac{\hat{s}_i}{i}$, we get an estimate of $\sum \frac{s_i}{i} = \frac{1}{\mu}$.

Ordering the sample by size we get $\{1, 2, 3, 3, 3, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6\}$ or

Size	Frequency
1	1
2	1
3	3
4	2
5	5
6	8

From this we get $\sum \frac{\hat{s}_i}{i} = \frac{32}{120} = \frac{4}{15}$ so our estimate of μ is $\hat{\mu} = \frac{15}{4} = 3.75$.

(2) What does this tell you about the time you spend waiting for a bus at the bus-stop? Can the average waiting time be very bad?

Hint You might think of seconds between buses (interarrival times) as family size.

Answer Using the hint, we can see that, if we think of our arrival time as a method of sampling seconds, then we are doing a size-biased sample of the interarrival times.

Using the same notation as before, if an interarrival time I has $\mathbb{P}(I = j) = p_j$ then the value of a (size-biased) interarrival time, S has $\mathbb{P}(S = j) = s_j = \frac{jp_j}{\mu}$ and mean $\nu = \sum \frac{j^2 p_j}{\mu}$.

If we assume that our arrival time is distributed uniformly during a (size-biased) interarrival time S, then, conditional on the value of the (size-biased) interarrival time, S, our waiting time W is uniformly chosen from $\{1, \ldots, S\}$. So our waiting time has conditional mean $\frac{S+1}{2}$ and so our mean waiting time is $\frac{\nu+1}{2}$.

This waiting time could have a very bad average.

To take 3 examples:

Distribution	Mean	Size-biased mean	Average waiting time
Binomial(n,p) $p_j = \binom{n}{j} p^j q^{n-j}$	np	$np + q$	$\frac{np+q+1}{2}$
Geometric(p) $p_j = q^{j-1}p$	$\frac{1}{p}$	$\frac{1+q}{p^2}$	$\frac{(1+q)}{2p^2} + \frac{1}{2}$
Modified Zipf(3) $p_j = \frac{4}{j(j+1)(j+2)}$	2	∞	∞