Question 1

David Hrabovszki

2/7/2020

Question 1: Cluster sampling

An opinion poll is assumed to be performed in several locations of Sweden by sending interviewers to this location. But since it is unreasonable from the financial point of view to visit each city, we use random sampling without replacement to select 20 cities, where the the probability of a city getting selected is proportional to its number of inhabitants.

1.

The data containing the cities and populations is found in population.csv, which we import into R.

2.

We create a function `sampler()` that selects 1 city from the whole list using a uniform random number generator. This function essentially performs weighted random sampling on our data, where the weight corresponding to each city is its population.

Our implementation is based on the pseudo-code written by Peter Kelly [1]. The function first generates an integer random number from the uniform distribution between the range [1,Total Population]. Then, it iterates through all the cities and updates the random number by subtracting the population of the current city from it. After that, but in the same iteration step, it compares this value to 0. If it is smaller than or equal to 0, then the function returns the current city and the loop breaks. If it is larger, then it continues into the next iteration step and updates the value with the subtraction of the next city's population from it until it finds a city, where this value is <= 0.

Since the generated random number gets lower and lower with every iteration, it will eventually return a result in every case. More populated cities get selected more often, because the larger the weight, the more likely to be larger than a randomly generated number in the range [1,Total Population], which means that their difference is <= 0.

3.

In this step, we use our `sampler()` function to randomly select 20 cities from the list without replacement. The function ensures that more populated cities get selected with a higher probability.

The method is as follows:

(a) Apply `sampler()` to the list of all cities and select one city

(b) Remove this city from the list

(c) Apply `sampler()` again to the updated list of the cities

(d) Remove this city from the list

(e) … and so on until we get exactly 20 cities.

[1] https://medium.com/@peterkellyonline/weighted-random-selection-3ff222917eb6

4.

Using `sampler()` function and the sampling method described in step 3, we obtain the following list of selected cities:

Table 1: Selected cities

	Municipality	Population
5	Huddinge	95798
8	Nacka	88085
16	Stockholm	829417
19	Tyresö	42602
51	Söderköping	14042
80	Hultsfred	13855
107	Kristianstad	78788
112	Malmö	293909
154	Lidköping	37989
155	Lilla Edet	12773
173	Tjörn	14961
177	Uddevalla	51518
178	Ulricehamn	22753
221	Västerås	135936
230	Mora	20146
247	Härnösand	24675
249	Sollefteå	20442
266	Nordmaling	7205
278	Arvidsjaur	6622
285	Luleå	73950

The average population of the 20 selected cities is 94273, while the average of all cities is 32209. This indicates that our random selection contains more populated cities with a higher probability.

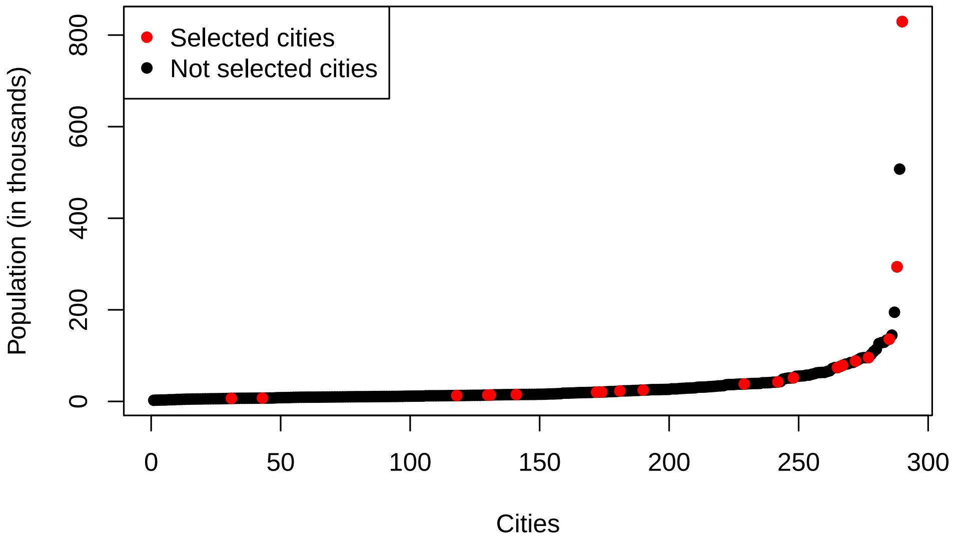

Figure 1: Weighted random selection of cities in Sweden

Figure 1 visualizes the cities as points in ascending order by their populations. The red points represent cities that were selected by our sampling method, the black ones represent the ones that were not. We can observe that more populated cities make up the majority of our selection, as they were more likely to be selected. In fact, out of the 20 most populated cities, we picked 5, while we chose 0 out of the 20 least populated ones.

5.

Figure 2 shows the histogram of the size of all Swedish cities, while figure 3 is the histogram of the size of the 20 selected cities. We can observe that both histograms look similar, this is due to the randomness of the sampling method. We can also see that cities with large population are very rare in the first case, but much more frequent in the second one. This is because our `sample()` function selects cities with large populations with a higher probability. Cities with smaller populations are still in majority, since there were a lot more of them originally, but their proportion is much smaller than before sampling.

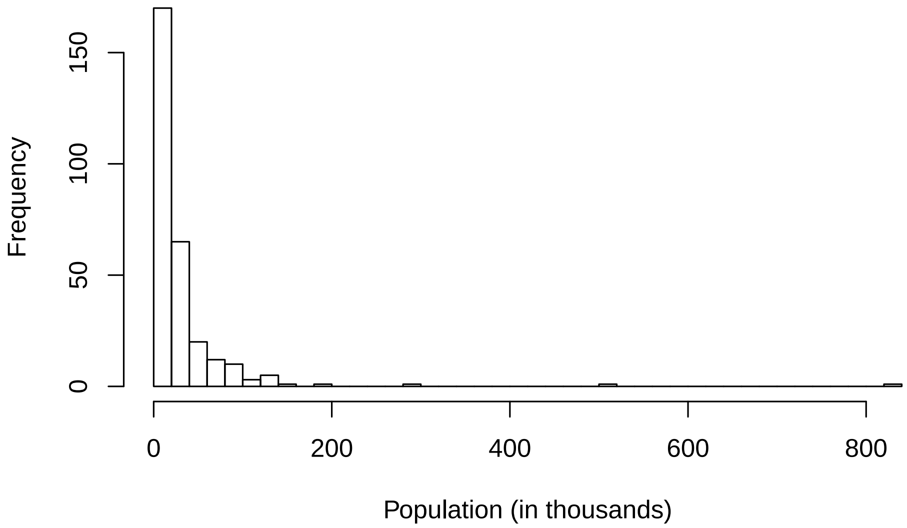

Figure 2: Histogram of the sizes of cities in Sweden

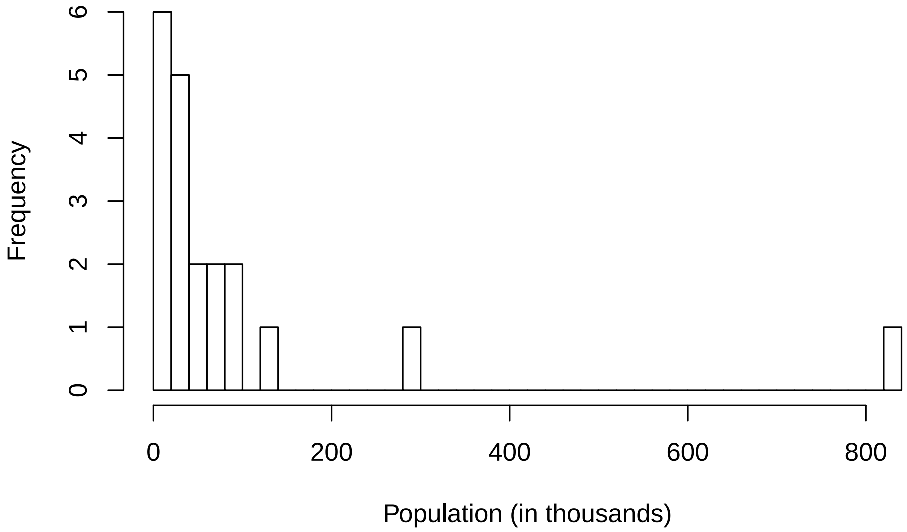

Figure 3: Histogram of the sizes of the selected cities