UCD School of Mathematics and Statistics

Statistical analysis of the formation of the Citizens' Assembly

Alan Benson, Nial Friel

9 November 2016

1 Introduction

The Citizens' Assembly (the Assembly) is a body comprising a government appointed Chairperson and 99 citizens of the Republic of Ireland, randomly selected to be "broadly representative" of the Irish electorate. The Assembly will meet on a regular basis from October 2016 to deliberate various topics, with the first discussion meeting taking place on 25 November 2016.

In this short report, we consider how the Assembly was constructed. RED C Research and Marketing Ltd (RED C) was appointed on 22 August 2016 to provide the representative sample of 99 members for the Citizens' Assembly and to also recruit 99 substitutes. In our analysis, we define "broadly representative" to mean that each person on the electoral register has an equal chance of appearing as one of the 99 members of the Assembly. This differs from the stratified sampling approach of RED C where a form of quota sampling was employed using Census 2011 and Quarterly National Household Survey (QNHS) population estimates.

2 Data

The Irish electorate is split up into 40 parliamentary constituencies, which do not exactly overlap the 26 counties of the Republic of Ireland. For this reason we form 22 areas by combining constituencies into areas to match as close as possible to the 26 counties, this data is shown in Table 1. Some counties are combined however these 4 combinations do not affect the results, as the number of citizens come from at most one county in each of the combinations. There are a total of 16 areas represented and 6 areas not represented on the Assembly. The 6 areas not represented incorporate 10 counties; Carlow, Cavan, Kerry, Kilkenny, Laois, Leitrim, Longford, Offaly, Sligo and Tipperary. The count of electors is taken from electors `http://www.electionsireland.org/` and the Citizens' Assembly counts are taken from `http://www.citizensassembly.ie`.

Area	Electors	Citizens
Carlow-Kilkenny	107,023	0
Cavan-Monaghan*	90,618	5
Clare	83,660	7
Cork	380,499	14
Donegal	117,675	4
Dublin	851,198	26
Galway	172,136	3
Kerry	112,751	0
Kildare	136,771	3
Laois	63,295	0
Limerick	143,201	6

Area	Electors	Citizens
Longford-Westmeath**	89,241	4
Louth	104,696	1
Mayo	92,958	4
Meath	130,188	8
Offaly	65,636	0
Roscommon	64,235	5
Sligo-Leitrim	95,911	0
Tipperary	112,615	0
Waterford	81,819	2
Wexford	109,861	1
Wicklow	97,858	6

Table 1: 22 areas constructed from the 26 counties of the Republic of Ireland (*5 from Monaghan, **4 from Westmeath). The number of registered electors and citizens from each area are shown.

3 Our sampling scheme for recruiting the Assembly participants

To allow each elector to have an equal chance of appearing on the Assembly, we sample without replacement 99 citizens from the 3,303,845 available electors. Then we map each elector to their area and analyse how many of the 22 areas are represented and how many are not represented. This sampling without replacement method prevents any person erroneously appearing on the Assembly more than once. This method is also known as simple random sampling and the outcome from this method is a case of the multivariate hypergeometric distribution. Let $N_1, \dots N_{22}$ be the counts of electors from each area shown in column 2 of Table 1 and let $a_1, \dots, a_{22}$ be the observed counts from each area (column 3 of Table 1). The probability mass function is then

$$P((A_1, A_2, \dots, A_{22}) = (a_1, a_2 \dots, a_{22})) = \frac{\binom{N_1}{a_1}\binom{N_2}{a_2}\dots\binom{N_{22}}{a_{22}}}{\binom{N_1+N_2+\dots N_{22}}{a_1+a_2+\dots a_{22}}} = \frac{\binom{N_1}{a_1}\binom{N_2}{a_2}\dots\binom{N_{22}}{a_{22}}}{\binom{3{,}303{,}845}{99}}. \quad (1)$$

This function is quite easy to understand, $\binom{3{,}303{,}845}{99} \approx 2.58 \times 10^{489}$ are the number of different possible groups of 99 different people we can make from 3,303,845 people. $\binom{N_1}{a_1}$ is the number of ways we can choose a_1 people from the $N_1 = 107{,}023$ electors in the Carlow-Kilkenny area, $\binom{N_2}{a_2}$ is the number of ways we can choose a_2 people from the $N_2 = 90{,}618$ electors in the Cavan-Monaghan area and so on for the 20 other areas. The sum $a_1 + \dots + a_{22}$ is exactly 99.

The above describes the method, however the question we would like answered is not how many citizens fall into each area but rather after the 99 citizens have been chosen at random, how many areas have no citizen on the Assembly by this method? This problem is well known in statistics as the classical occupancy problem and is closely related to another problem called the coupon collector's problem in combinatorial probability. In order to analyse this occupancy problem, where we are interested in the occupancy of each of the 22 areas, we perform a Monte Carlo simulation of 1,000,000 different assemblies, each comprising 99 people. Over the 1,000,000 simulated assemblies we will calculate the probability distribution of the number of areas not represented. From this distribution we will be able to calculate the chances no areas are left out, 1 area is left out, 2 areas are left out, etc. Computing the exact distribution, without simulation, is difficult and complicated further by the fact we are sampling without replacement. However for a large population, in our case 3,303,845, the difference between with replacement (multinomial) and without replacement (multivariate hypergeometric) sampling is negligible.

3.1 Notes on the approach of RED C Research

We stress again that the approach we are using is not the approach of RED C Research, the company appointed to form the Assembly. Their approach does not directly give equal chance of each person of the Irish electorate appearing on the assembly. Their approach was targeted instead to form a representative sample of the Irish electorate using Census 2011 figures and QNHS population estimates. Their methodology is described in a document available at `http://www.citizensassembly.ie`. Many of the people we randomly select may not be available to appear on the Assembly, nor may not be eligible to appear on the Assembly. However the number excluded does not seem be a large proportion of the electorate. RED C initially used a random address based scheme sampled from population estimates and after this confirmed that the people selected were eligible to vote. The random address based scheme was to select District Electoral Divisions (DEDs), pick a first house and then pick every nth house, where n was not specified in the methodology document. It is not clear in which direction (North, South, East, West) the nth house was selected nor what happened if the nth house had previously been sampled. When they arrived at the house they would ask to speak the person who was next to have a birthday and communicated with this person. Although this is a randomisation step to ensure only 1 person was selected per household it does jeopardise larger households. We do not take a household approach.

4 Monte Carlo simulation results for 1,000,000 assemblies

Shown in Figure 1 are the results of simulating 1,000,000 assemblies. The most likely (modal) number of areas not represented is 1. The mean number of areas not represented is 1.1432. The current Assembly omits 6 areas. The probability of observing 6 or more areas not represented based on this simulation is 0.0155% or about 1 in 6500.

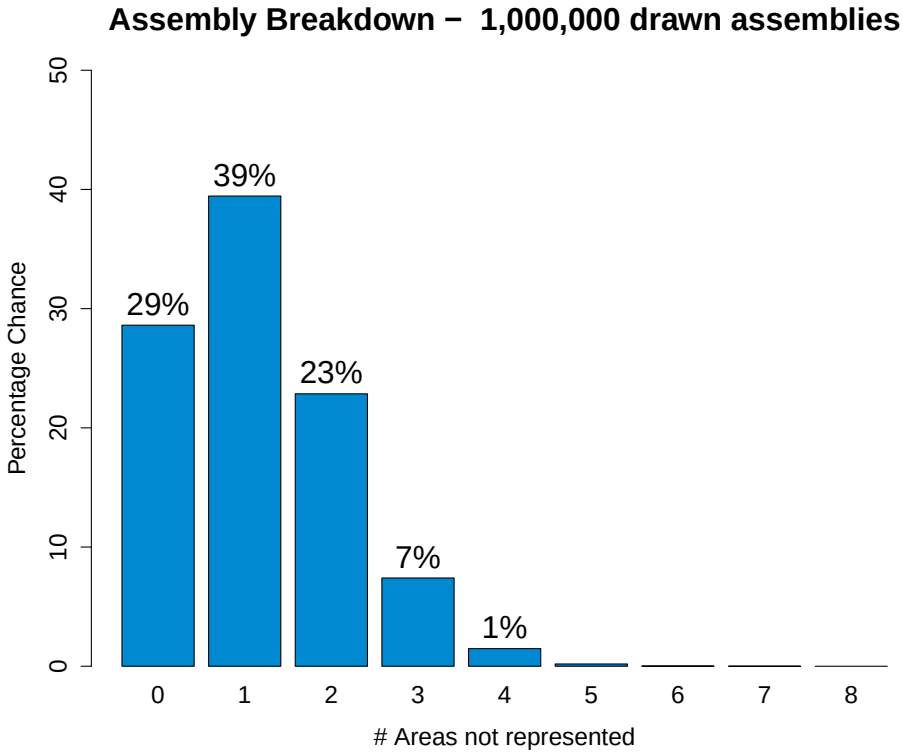

Figure 1: 1,000,000 assemblies drawn by sampling the electoral population (without replacement). This plot shows the distribution of the number of areas not represented. The mode is "1 area not represented". For the 2016 Citizens' Assembly, 6 areas were not represented. The probability of observing greater than or equal to 6 areas not represented, based on this simulation, is 0.0155%. ***Note*** *- labels less than 1% are not shown, but* ***not*** *assumed to be exactly 0%.*

5 How large should the assembly be?

An interesting question to ask is how many people should be on the assembly be in order to ensure it is more likely that every area is represented than any areas not represented. This is the coupon collector's problem which studies how long a person collecting coupons (or football stickers) must wait to collect each coupon from a finite set. The first coupons are easy to collect but after a while duplicates start appearing and the last few coupons take much longer to find! For the Assembly, the first few areas are easy to fill but it will take many more people to ensure we have a least one from each area. Imagine rolling a die with 22 sides and waiting until you have observed every number. How long should you wait to see every number at least once? How long should you wait given the die is biased and lands on each side with different probabilities proportional to the size of the electorate?

5.1 Simulation

We will keep picking people for the assembly and wait until we have at least 1 person from each area then record how large the resulting assembly is. The median assembly size from the simulation was 121. This means if we choose an assembly of size 121 or more we will more than likely have each area represented. If we want higher certainty of having each area represented the table below shows a range of certainty values and required assembly size.

Certainty	Size Required
10%	79
20%	91
30%	101
40%	111
50% (median)	121
60%	133
80%	165
100%	3,240,551

Table 2: Assembly size required to ensure a certain probability of full area representation when hypergeometric sampling is used.

The only way we can be certain (100% probability) that each county is represented would require in the an assembly of size 3,240,551 (which is the total electorate 3,303,845 less the size of the smallest area (Laois) + 1). Other restrictions such as ensuring we have a male and a female from each area, a person from each age group etc. would require much larger assemblies but with available data we could calculate these required sizes accurately.

Appendix

A Data sorted by electorate size

Area	Electors	Citizens	Area	Electors	Citizens
Laois	63,295	0	Carlow-Kilkenny	107,023	0
Roscommon	64,235	5	Wexford	109,861	1
Offaly	65,636	0	Tipperary	112,615	0
Waterford	81,819	2	Kerry	112,751	0
Clare	83,660	7	Donegal	117,675	4
Longford-Westmeath**	89,241	4	Meath	130,188	8
Cavan-Monaghan*	90,618	5	Kildare	136,771	3
Mayo	92,958	4	Limerick	143,201	6
Sligo-Leitrim	95,911	0	Galway	172,136	3
Wicklow	97,858	6	Cork	380,499	14
Louth	104,696	1	Dublin	851,198	26

Table 3: The data in Table 1 sorted by number of electorate size (*5 from Monaghan, **4 from Westmeath).

B How random is our assembly picker?

Across the world many assemblies are created using a method known as sortition. The method uses random numbers from lotteries all over the world so that the data is verifiable. The method is know as Publicly Verifiable Nominations Committee Random Selection or NomCom. More information can be found here `https://tools.ietf.org/html/rfc3797`.

Ideally our method of sampling needs to be completely random. To ensure complete randomness of the assembly picker and to prevent external interference, the source of randomness for picking the members was taken from a statistically verifiable hardware random number generator (TrueRNG v3 - Figure 2). The entropy of this generator was sampled during the running of our algorithms and found to have entropy of 7.999961 bits per byte (8 bits) with a serial correlation coefficient of -0.001072.

Figure 2: Our random number generator. Thanks to Ubld Electronics LLC.