Purposive Sampling in the Analysis of Count Data

Paolo Victor T. Redondo
School of Statistics
University of the Philippines Diliman

Purposive sampling is a non-probability sampling method oftentimes used whenever probability samples are not efficient, costly, or simply not feasible. Many measurements are result of counting specific attributes, hence, Poisson Regression is often used in modeling. The goal of this study is to determine whether location-based purposive sampling can help improve the estimates produced by Poisson Regression and if it can facilitate reduction in the sample size requirement while maintaining optimality of estimates. A simulation study involving several partitions of the predictor values is implemented. In some partitions, the model works better even in small sample size, say 50, while in other partitions, the model works at least as good as their respective simple random sample counterpart.

Keywords: purposive sampling, poisson regression, sample size

1. Introduction

Count responses can be observed in almost every field, whether in medicine (count of non-healthy cells), community-based monitory system indicators (number of malnourished children under age 5 per barangay), telecommunications (number of subscribers per month), and many others. A well-known model used to analyze such data is poisson regression where the assumption of equality of mean and variance (*equidispersion*) must hold true. Also, it is appropriate to model count data using poisson pegression especially if some conditions are met, e.g., logarithm of the mean of the response variable follows a linear function of explanatory variables, independence, (Ang and Barrios, 2015). However, in practice, assumption of equidispersion, not to mention independence, is oftentimes false whenever the variance is either smaller (*underdispersion*) or larger (*overdispersion*) than the mean. These deviations from the assumptions could generate estimates from poisson regression that are still consistent but standard errors are usually unstable and hence, inferences based on standard errors becomes invalid, (Husain and Bagmar, 2015). Furthermore, underdispersion and overdispersion are often assessed using models that incorporates extra parameters and hence, estimation procedures are usually more complicated. On the other hand, the distribution of the count response in the possion regression model around the expected value approaches normality as the conditional mean increases (Williams, 2015).

In the sense of data collection, whenever deemed efficient and possible, probability sampling is highly recommended because randomization reduces biases and allows the generalization of conclusions generated. However, random sampling is not always feasible and at the same time efficient, e.g., highly dispersed population may induce higher cost for researchers (Bernard, 2002). In these scenarios, non-probability sampling is commonly exercised. Purposive sampling, a non-probability sampling method, entails the categorization of subjects in accordance with the specified criteria based on the research problem (Mack et al., 2005). Cresswell and Plano Clark stated that this procedure involves identifying and selecting individuals that are experienced with or are especially knowledgeable about the phenomenon of interest (as cited by Palinkas et al., 2015). Also, purposive sampling enables an initial understanding of that phenomenon (ACAPS, 2011). There are numerous purposive sampling designs, of which includes the selection of extreme values or deviant cases, selection of cases with maximum variation, selection of homogenous cases, etc. (Palinkas et al., 2015). Some studies suggest that this non-probability sampling method can lead to representative samples, i.e., when either sample characteristics or inferences from the obtained sample approximates the population values well (van Hoeven et al., 2015). Hence, another hurdle to be answered is "How much information do we need?"

Sample size determination continues to be one of the most important areas in statistics; too small sample size can lead to under-powered studies while too large sample size wastes precious resources (Sides et al., 2015). In modeling, it is always necessary to have more observations than predictors, at least from a classical point of view (Hyndman and Kostenko, 2007). Oftentimes, relying on the assumption of (approximate) normality, sample size calculation is typically based on analytical formulas, (Baio et al., 2015). Although theories exist as to calculating sample sizes, these does not address the issue of minimum sample size requirements in the sense of the statistical power of the hypothesis test on the estimated coefficients (de Bekker-Grob, 2015).

This study evaluates whether purposive sampling can help improve the parameter estimates of poisson pegression, whether it will lessen the sample size in order to achieve such results, and under which situation can poisson regression still capture (in terms of parameter estimates) the behavior of a data generating process which produces overdispersed count responses.

2. Methodology

Implementing purposive sampling

Suppose the distribution and center of an explanatory variable is known in the literature, can obtaining the sample near or far from the center help improve the estimates of a model? Thus, the concentration of observations on a specific partition determined by relative locations is a possible perspective of purposive sampling hypothesized in this study.

Different partitions are defined where purposive samples are drawn from. Five partitions are created separated by the 2^{nd}, 4^{th}, 6^{th} and 8^{th} decile labelled A, B, C, D and E. For each partitions, different concentration of observations are specified as shown below:

1. Equal Partition – observations of the sample are equally distributed in the five partitions, i.e., 20% belongs in A, another 20% in B and so on up to partition E. (20-20-20-20-20)
2. Light Mid – concentration of the observations are mostly near the center of the distribution. (10-20-40-20-10)
3. Heavy Mid – concentration of the observations are very near the center of the distribution without representative for extreme values. (0-25-50-25-0)
4. Light End – concentration of the observations are slightly far from the center of the distribution having no representative for central data. (15-35-0-35-15)
5. Heavy End – concentration of the observations are very far from the center of the distribution having no representative for central data. (35-15-0-15-35)

Data generating process

Different scenarios with varying data generating process, importance of explanatory variables, from which variable will the proposed purposive sample is obtained and sample sizes are constructed.

Five data generating processes, namely poisson distribution and negative binomial distribution with varying dispersion parameter α (0.5, 1.0, and 5.0) are used to evaluate whether purposive sampling can improve the estimates of parameters in poisson regression whether overdispersion (different magnitude) is present or absent.

Normally distributed independent random variables, uniformly distributed independent variables and explanatory variable following Poisson distribution are considered. For the case of Poisson, two possible means, ($\lambda = 5$ and $\lambda = 20$) are specified to visualize the effect of small and large conditional mean of the count responses. For coding purposes, let the first variable to be normal (X_1), the second to be Poisson (X_2) and the last to be uniform (X_3). Thus, the heterogeneous mean of the count process is given by:

$$\mu_Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad (1)$$

Importance of variables

Based on the magnitude of coefficient in Equation (1), greater importance of variable is associated with a larger value of parameter. When the normally distributed is deemed to be the most important variable, we assign the following values of the parameters in Equation (1) as follows: β_1= -0.75, β_2= 0.25 and β_3= -0.5. When the most important variable is Poisson distributed, the coefficients are: β_1= -0.25, β_2= 0.75 and β_3= -0.5. And if the uniformly distributed variable is the most important: β_1= -0.25, β_2= 0.5 and β_3= -0.75.

Sorting

Since three independent variables are considered, purposive sampling is done in three different ways, from which explanatory variables the relative location-based samples come from. Thus, for each data generating process, depending on the most important variable, purposive samples are obtained based on which variable is sorted, that is, if the data is sorted using the normally distributed explanatory variable, the sample that will be obtained follows the concentration stated (depending on the partition) for the sorting variable.

Sample Size

Six cases of sample size are used; 50, 100, 500, 1000, 5000 and 10000. This is to verify if estimates improve for each proposed purposive sample as the size of the sample increases. All sample sizes are drawn from a simulated population of size 100,000.

Assessment

To assess the outcome of purposive sampling, the estimates are compared to the estimates obtained using a simple random sample with same sample size. The difference between the value of the estimate from the true value of the parameter being estimated are observed.

3. Results and Discussion

Results per sample sizes in each scenario in all partition are omitted to facilitate the summarization, however, the sample size where the estimates no longer significantly change are indicated for comparison.

Table 1. Poisson Data, High Mean, Important Predictor is Normal

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	1000	1.000455795	-0.75003271	0.249968215	-0.499725807
Normal	Heavy End	500	0.999870468	-0.749868261	0.250004344	-0.499752873
Poisson	Heavy End	500	1.000185912	-0.750075799	0.249993396	-0.500235261
Uniform	Heavy End	500	0.999467529	-0.750250666	0.249997125	-0.499067139

Note that using Heavy End partition in any of the variables, based on the deviation from the true value of the parameter, the estimates are at par or better than that of the SRS even for smaller sample size.

Table 2. Poisson Data, High Mean, Important Predictor is Poisson

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	500	0.999998066	-0.249999846	0.750000056	-0.499999838
Normal	Heavy End	50	0.999996432	-0.250000731	0.749999914	-0.499995229
Poisson	Heavy End	50	0.999991557	-0.249999556	0.750000384	-0.500001488
Uniform	Heavy End	100	0.999994247	-0.249999872	0.750000284	-0.500002022

Regardless of the sorting variable, using Heavy End partition, based on the magnitude of bias, the estimates are at par or better than that of the SRS even for smaller sample size.

Table 3. Poisson Data, High Mean, Important Predictor is Uniform

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	100	1.000128703	-0.249986364	0.499997436	-0.75008728
Uniform	Equal	50	1.000103517	-0.250040229	0.499992859	-0.749964186
Uniform	Heavy End	100	0.999912702	-0.249989585	0.500005068	-0.750045824

The importance of regressors being considered, sorting using the most important variable (uniformly distributed), and obtaining the sample under Equal partition improves the estimate for the sorting variable while maintaining the quality of other estimates for a smaller sample size. On the other hand, under Heavy End partition, the estimates are as good as those of their SRS counterpart.

Table 4. Poisson Data, Low Mean, Important Predictor is Normal

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	1000	1.005269386	-0.749819594	0.249545565	-0.506245068
Normal	Heavy End	500	0.99302172	-0.752770571	0.250358267	-0.494687647
Poisson	Heavy End	100	1.003298852	-0.751031747	0.249812392	-0.502609816
Uniform	Heavy End	500	1.006833106	-0.750552156	0.249049894	-0.502718496

It can also be seen from Table 4 that the estimates for this scenario using Heavy End partition is almost at par with the estimates of the SRS with smaller sample size.

Table 5. Poisson Data, Low Mean, Important Predictor is Poisson

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	500	1.000678403	-0.250133818	0.750003294	-0.501528559
Normal	Heavy End	500	0.998436922	-0.249736269	0.750073361	-0.499335199
Poisson	Heavy End	500	0.998025581	-0.24991987	0.750148416	-0.499380032
Uniform	Heavy End	500	0.999730941	-0.24984088	0.749983781	-0.49948238

Similar with the above scenarios, regardless of the sorting variable used, obtaining the sample under the Heavy End partition, estimates are almost at par with that of the SRS with the same sample size.

Table 6. Poisson Data, Low Mean, Important Predictor is Uniform

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	1000	0.99661007	-0.250161598	0.500275868	-0.749626956
Poisson	Heavy End	100	0.999702681	-0.250242865	0.499998832	-0.750977098
Uniform	Equal	100	0.997722766	-0.249768212	0.500433713	-0.751649934

Given that most important explanatory variable is uniformly distributed, drawing the sample under the Heavy End partition after sorting with the Poisson variable gives estimates as good as their SRS counterpart but for smaller sample size. Also, when sorting using the uniform variable, equal partition improves the estimates if important variables are Poisson or uniform but slightly increases the bias for the estimate in the parameter associated normal variable as important predictor.

Table 7. Negative Binomial (0.5) Data, High Mean, Important Predictor is Normal

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	1000	1.00688757	-0.741215511	0.24941286	-0.502924566
Normal	Heavy End	1000	1.046528094	-0.742805647	0.249038256	-0.547703315
Poisson	Heavy End	1000	1.023199946	-0.744098896	0.249879794	-0.558112767

Even for slightly overdispersed data, given that the most important variable is normally distributed, Poisson regression can somehow capture the behavior of the data generating process in terms of parameter estimates. Furthermore, using the Heavy End partition for Poisson and Normal variables works as good as the SRS sample.

Table 8. Negative Binomial (0.5) Data, High Mean, Important Predictor is Poisson

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	10000	11.16412368	-0.104644615	0.33805481	-0.246484516
Poisson	Heavy Mid	50	1.103325902	-0.218625759	0.742085752	-0.465546264

Note that for the SRS, Poisson regression seemed to not capture the behavior of the slightly overdispersed data even for large sample size. However, had the Poisson variable to be the most important, sorting with the said variable and obtaining the sample under the Heavy Mid partition does not only improve the result but enables Poisson regression to give good estimates for the parameters of interest with a much smaller sample size.

Table 9. Negative Binomial (0.5) Data, High Mean, Important Predictor is Uniform

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	10000	1.065059612	-0.276931992	0.496407515	-0.767844747
Poisson	Heavy Mid	50	1.033724755	-0.246717877	0.498193557	-0.751874782

Using Heavy Mid Partition after sorting with the Poisson variable, the estimates generated are better than that of SRS with smaller sample size.

Table 10. Negative Binomial (0.5) Data, Low Mean, Important Predictor is Normal

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	5000	0.998669031	-0.752933172	0.249486492	-0.497227424
Poisson	Heavy Mid	50	0.91963768	-0.748551288	0.259825493	-0.506974751
Uniform	Heavy End	1000	0.984087934	-0.750015376	0.253221096	-0.507634258

In Table 10, using Heavy Mid partition after sorting with Poisson variable and using Heavy End partition after sorting with uniform variable helps improve the estimates given by the Poisson regression, at much smaller sample size, even when the data is slightly overdispersed.

Table 11. Negative Binomial (0.5) Data, Low Mean, Important Predictor is Poisson

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	1000	1.028528315	-0.235874498	0.744728299	-0.523172968
Poisson	Heavy Mid	100	1.005953307	-0.252963387	0.745639277	-0.500040495
Poisson	Heavy End	100	1.01356999	-0.244922579	0.742296664	-0.488662057

Though the parameter estimates are slightly far from the true value of the parameter, the results obtained using the Heavy Mid and Heavy End partitions are comparable to the results given by their respective random sample counterpart but with a much smaller sample size. Note that only sorting with Poisson variable are stated due to the non-improvement set by the other cases. This might be due to the greater importance of the Poisson variable than others.

Table 12. Negative Binomial (0.5) Data, Low Mean, Important Predictor is Uniform

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	1000	0.997115614	-0.250016553	0.499659206	-0.752508412
Normal	Heavy Mid	500	1.028554055	-0.253795351	0.496933255	-0.766948564
Poisson	Heavy Mid	100	0.980582147	-0.253767952	0.503334659	-0.745298417
Uniform	Equal	100	1.011390141	-0.251981854	0.493603863	-0.739317803
Uniform	Heavy End	100	0.940907649	-0.243151946	0.505594751	-0.734083904

Note that for normal and Poisson variables, Heavy Mid partition seemed to works at par or better than the SRS sample but for even smaller sample size. Also, sorting with the uniform variable, Equal partition and Heavy End partition both perform as good as the SRS.

Table 13. Negative Binomial (1) Data, High Mean, Important Predictor is Normal

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	5000	0.990579645	-0.763277178	0.250425637	-0.511010222
Normal	Heavy End	500	0.929485636	-0.765870126	0.251890774	-0.500833213
Poisson	Heavy Mid	1000	0.983655113	-0.753728436	0.251298407	-0.503774497
Uniform	Equal	500	1.135242821	-0.743503527	0.243961548	-0.517207113

Sorting with normal variables then using Heavy End partition, sorting with Poisson variable then using Heavy Mid partition and sorting with uniform variable then using Equal partition work almost at par or even better than the SRS at smaller sample size for this given scenario.

Table 14. Negative Binomial (1) Data, High Mean, Important Predictor is Poisson

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	10000	11.28691401	-0.130919596	0.328481471	-0.158963116
Poisson	Heavy Mid	100	0.944910053	-0.286591544	0.74974257	-0.492417638

Whenever the Poisson regressor is deemed to be the most important variable with its mean being high, estimates form the SRS does not capture the behavior of the data generating process while sorting using the said variable improves the estimates of the Poisson Regression up to the point that it somehow captures the behavior of the overly dispersed data even for a small sample size.

Table 15. Negative Binomial (1) Data, High Mean, Important Predictor is Uniform

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	10000	1.352734829	-0.23187834	0.485930109	-0.630843247
Poisson	Heavy Mid	500	1.169122052	-0.249399079	0.491803197	-0.747601409
Uniform	Equal	100	0.788411195	-0.246859465	0.504761788	-0.759964229

Even though the estimates under Heavy Mid partition after sorting with Poisson variable and Equal partition after sorting with uniform variable are still slightly far from the true value of the estimates, it is comparable to or even better than the results obtained from the SRS but with much smaller sample sizes.

Table 16. Negative Binomial (1) Data, Low Mean, Important Predictor is Normal

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	1000	0.976901436	-0.74795959	0.256322633	-0.520442352
Normal	Heavy End	100	1.006340837	-0.748907412	0.241970813	-0.497628295
Poisson	Heavy End	500	0.980454713	-0.749329828	0.252400082	-0.485323797
Uniform	Equal	500	0.982919779	-0.750956829	0.251573593	-0.499778636

Given the stated scenario, Heavy End partition seems to work better or at par with the SRS when sorting is done with normal and Poisson variables while Equal partition for sorted uniform variable all for smaller sample size.

Table 17. Negative Binomial (1) Data, Low Mean, Important Predictor is Poisson

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	5000	1.099399083	-0.256859792	0.736616048	-0.509301659
Normal	Heavy End	500	1.084596829	-0.252381336	0.732370962	-0.502972244
Poisson	Heavy End	100	1.008325528	-0.236196945	0.735249891	-0.473205012

Note that Heavy End partition improves the estimates given by the Poisson regression when sorting is done with the normal and Poisson variable. Also, it somehow makes the regression model appropriate for the overly dispersed data with sample size smaller than that of the SRS.

Table 18. Negative Binomial (1) Data, Low Mean, Important Predictor is Uniform

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	1000	1.066870723	-0.246592619	0.491257173	-0.766679991
Normal	Heavy End	50	1.003572344	-0.253045281	0.487221946	-0.769492883
Poisson	Heavy End	50	0.964204814	-0.224622915	0.490784296	-0.706955045
Uniform	Heavy End	50	1.025256261	-0.251245479	0.482740603	-0.760913396

For the given scenario, regardless of the sorting variable, Heavy End partition improves the estimates produced while considering a much smaller sample size.

Table 19. Negative Binomial (5) Data, Low Mean, Important Predictor is Normal

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	10000	0.89902744	-0.735496621	0.254021022	-0.503336026
Uniform	Equal	5000	0.892794394	-0.744034701	0.254111966	-0.506902024

Sorting with the uniform variable and obtaining the sample under the Equal partition works better than the SRS counterpart for smaller sample sizes.

Table 20. Negative Binomial (5) Data, Low Mean, Important Predictor is Poisson

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	10000	11.00897317	-0.111814463	0.326178746	-0.20291408
Poisson	Heavy Mid	1000	0.748179513	-0.255264844	0.761172782	-0.490731115

Whenever the Poisson regressor is believed to be most important and its mean is known to be high, sorting using this variable and getting the sample under the Heavy Mid partition enables Poisson Regression to capture the behavior of a highly overdispersed data for relatively smaller sample size than SRS.

Table 21. Negative Binomial (5) Data, Low Mean, Important Predictor is Uniform

Sorting Variable	Partition/ SRS	Sample Size	β_0	β_1	β_2	β_3
	SRS	10000	1.099191157	-0.200843355	0.492124661	-0.632795144
Poisson	Heavy Mid	500	0.828568144	-0.22789842	0.505748128	-0.757413385

Even if uniform is the most important regressor, for as long as the mean of the Poisson variable is high and the data is greatly overdispersed, sorting with the Poisson variable and obtaining the sample under the Heavy Mid partition works best in improving the result of the regression with a relatively small sample size than that of the SRS.

4. Conclusion

Based on the results obtained from the simulation study, estimate for the parameter associated with the sorting variable often improves regardless of the partition used whenever no overdispersion or slight overdispersion (up to α = 1) is present in the data even for small sample sizes. In most cases, Heavy End partition is either at par with SRS or even better for small sample sizes regardless of the importance of variables and sorting variable for as long as only slight overdispersion is evident. Given that the data is greatly overdispersed, sorting with normally distributed and uniformly distributed variables does not contribute to the improvement of estimates, this might be due to the relatively small magnitude of values considered since both of their respective mean and variance are smaller than the Poisson variable. On the other hand, sorting by the Poisson variable under the Heavy Mid Partition does not only improve the estimates but also enables Poisson Regression to capture the behavior of the overly dispersed data for some cases of high overdispersion. For small sample sizes such as 50 and 100, the proposed partitions often improve the estimates of the Poisson regression. Lastly, whenever sorting is reckoned to be useful, based on the obtained results, Heavy End partition works better for normal and Poisson variables for slightly overdispersed data while equal partition is best for uniformly distributed variables especially for highly overdispersed data and small sample size.

ACKNOWLEDGEMENT

I would like to thank Dr. Erniel Barrios for guiding me in the direction of this study and for giving insightful comments about the result of the simulation.

REFERENCES

ACAPS, 2011, Technical Brief: Purposive Sampling and Site Selection in Phase 2. Available at: http://www.acaps.org/img/documents/purposive-sampling-and-site-selection-purposive-sampling-and-site-selection.pdf

ANG, M. and BARRIOS, E., 2015, Cross-country analysis of maternal mortality ratio, Working Paper Series, 2015(03), University of the Philippines Diliman School of Statistics. Available at: http://www.stat.upd.edu.ph/docs/research/working%20papers/2015/working%20paper%202015-03.pdf

BAIO, G., COPAS, A., AMBLER, G., HARGREAVES, J., BEARD, E. and OMAR, R., 2015, Sample size calculation for a stepped wedge trial, Available at: http://trialsjournal.biomedcentral.com/articles/10.1186/s13063-015-0840-9

BERNARD, H. R., 2002, *Research Methods in Anthropology: Qualitative and Quantitative Methods*, 3rd ed., Walnut Creek, California: AltaMira Press.

DE BEKKER-GROB, E., DONKERS, B., JONKER, M. and STOLK, E., 2015, Sample size requirements for discrete-choice experiments in healthcare: A practical guide, *Patient* 8:373-384.

HUSAIN, M. and BAGMAR, S.H., 2015, Modeling under-dispersed count data using generalized poisson regression approach, *Global Journal of Quantitative Science* 2(4): 22-29.

HYNDMAN, R. and KOSTENKO, A., 2007, Minimum sample size requirements for seasonal forecasting models, *Foresight* 6 (Spring):12-15. Available at: http://robjhyndman.com/papers/shortseasonal.pdf

MACK, N., WOODSONG, C., MACQUEEN, K. M., GUEST, G. and NAMEY, E., 2005, *Qualitative Research Methods: A Data Collector's Field Guide*, Family Health International, North Carolina, USA.

PALINKAS, L., HORWITZ, S., GREEN, C., WISDOM, J., DUAN, N. and HOAGWOOD, K., 2015, Purposeful sampling for quantitative data collection and analysis in mixed method implementation research, *Science+Business Media* 42(5): 533-544, New York.

SIDES, R., KAHLE, D. and STAMEY, J., 2015, Bayesian sample size determination in two-sample poisson models, *Biometrics & Biostatistics International Journal* 2(1): 00023.

VAN HOEVEN, L., JANSSEN, M., ROES, K. and KOFFIJBERG, H., 2015, Aiming for a representative sample: simulating random versus purposive strategies for hospital selection, *BMC Medical Research Methodology*. Available at: http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-015-0089-8

WILLIAMS, R., 2015, *Models for Count Outcomes*, University of Notre Dame. Available at: https://www3.nd.edu/~rwilliam/stats3/CountModels.pdf