Estimating the underlying infant mortality rates for small populations: A case study of counties in Estonia.

David A. Swanson

Center for Studies in Demography & Ecology

University of Washington

dswanson@ucr.edu

and

Department of Sociology

University of California Riverside

Riverside, California USA 92521

+951 827 6398

dswanson@ucr.edu

ABSTRACT

Infant mortality is an important population health statistic that is often used to inform health policy decisions. For a small population, an infant mortality rate is subject to high levels of stochastic uncertainty and may not indicate the intrinsic mortality regime affecting the population. This situation leads some agencies to either not report infant mortality for these populations or report infant mortality aggregated over space, time, or both. A method is presented for estimating infant mortality rates that reflect the intrinsic mortality regimes underlying small populations. The method is described, tested for validity, and illustrated in a case study by estimating IMRs for the 15 counties in Estonia. The study suggests that the method can produce reasonable estimates of the "underlying" infant mortality rates for small populations subject to high levels of stochastic variation and for which infant deaths may not even be reported. In this regard, the method described here may assist in the generation of information about the health status of spatially concentrated immigrant and ethnic minority populations in Europe, particularly in terms of infant mortality.

KEYWORDS

Policy Decisions, Monitoring Health Status, Migrants, Small Population, Beta-Binomial Model

INTRODUCTION

The infant mortality rate (IMR) is widely used. It is an indicator not only as a measure of the risk of infant death but as an indicator of the availability and quality of health care services, poverty levels, and socio-economic status differentials (Hummer, 2005; Kitagawa and Hauser, 1973; Link and Phelan, 1995; Stockwell, Goza and Balisteri 2005; Stockwell et al., 1987).[1]

Because statistical data are often used to guide health policy decisions, it is not surprising that the IMR also is used in this regard (Chen, Oster, and Williams, 2016; Kleinman, 1996; Misra et al., 2004; Stockwell et al. 1987). Moreover, as observed by VanEenwyk and Macdonald (2012), questions concerning health outcomes and related health behaviors and environmental factors often are studied within small subgroups of a population, because many activities to improve health affect relatively small populations. Fortunately, the advent of geographic information systems and high volume, fast computer-based information systems often involving the matching of records from different sources means that this type of information is technically feasible. However, the demand for this information along with the technical feasibility of obtaining it is not always compatible with the need of for data confidentiality or the realities of stochastic uncertainty concerning small populations.[2] This means that even when it is possible to provide data for a small population, it is not always the case that they are, a situation not often encountered when dealing with large populations (Office for National Statistics, 2015). The Centers for Disease Control, the unit within the US National Center for Health Statistics that reports vital statistics, for example, does not present or publish death or birth counts of nine or fewer or rates based on counts of nine or fewer (in figures, graphs, maps, table, etc.) at the subnational level (US National Center for Health Statistics, no date).

Data representing small populations are not only subject to limitations posed by confidentiality concerns, they also are subject to higher levels of measurement bias and lower levels of precision than those typically found in larger populations (Reeske and Razum, 2011; Swanson and Tayman, 2012: 216). These issues are typically associated with the stochastic uncertainty that affects small populations. A typical strategy for dealing with the combination of these issues is to aggregate data for small populations and generate what amounts to an arithmetic average from them. Another strategy is to gain permission to access individual level records, match them and then construct statistics (Kinge and Kornstad, 2014). However, unlike the strategy of aggregation, this approach inevitably requires administrative approval and requires both a substantial amount of time and personnel costs implement.

The issues concerning data availability for small populations directly affect immigrant populations because these groups tend to be clustered in the country of destination by ethnicity and country of origin and spatially segregated from the citizens of the country of origin (Andersson, 2013; Blom, 1999; Boal, 1996, Bråmå, 2008; Clark and Ware, 1997; Dunn, 1998, Kalandides and Vaiou, 2012; Kempen and Ozuerkren, 1998; Ljunggren and Andersen, 2015; and Massey, 1985). By virtue of being so clustered and segregated, immigrants represent small populations in the context of the countries in which they are found. Moreover, it is often not easy to obtain health and data on these same populations, especially in Europe. Rechel, Mladovsky and Devillé (2011), for example, find that accurate data on the health of migrants are lacking in many European Union countries.

Unfortunately, data representing small populations are subject to high levels of stochastic uncertainty, which implies that reported IMRs for small populations can vary dramatically over time even though there is no substantive change in their respective intrinsic mortality regimes,

as reflected in their "underlying" infant mortality rates. Awareness of this situation has led to a range of methods used in developing estimates of the underlying IMRs for small populations. One approach is "non-reporting," which is to simply not report IMRs for small populations, as is the case with the Centers for Disease Control (US National Center for Health Statistics, no date). Unfortunately, this approach discards related information (e.g., reported births) that may be of use in estimating IMRs for small populations – a point to which I later return.

Another general approach is to provide an estimate by embedding small population information within a larger context, which takes us back to the "aggregation strategy" discussed earlier. This approach is used by, among other agencies, the US National Center for Health Statistics (2018), for which the "larger context" is defined both in terms of time and space. In terms of time, the NCHS data on infant mortality rates by county are aggregated for the period 2007-2015 and in terms of space, counties with small populations are aggregated. One drawback to both approaches is that neither is specific to the time and county of interest, which implies that in terms of the IMRs constructed from them are, in fact, simple arithmetic averages. Related to this issue is the fact that these averages are biased unless appropriate weights or other procedures are used to reduce bias (Voss et al., 1995), steps that may not be feasible in a given situation. Another "large context" approach, one taken in this paper, which, unlike the "non-reporting" approach, has the potential to provide estimates of the IMRs underlying small populations, while also avoiding the drawbacks found in the aggregated approach. Another benefit of this approach is that it is a statistical estimator and, as such, is not in conflict with confidentiality issues. To this end, a publication by Link and Hahn (1996) was used as a starting point in generating the approach described, tested, and applied here.

Estonia is used as a case study because it is an example of a small population in and of itself in that only 13,197 births were reported by Statistics Estonia (2017) for the country as a whole in 2015.[3] Among its 15 counties (see Exhibit 1), the number of births ranges from a low of 70 births in Hiiu County to a high of 6,864 in Harju County (where the capital city, Tallinn, is located). Making Estonia even more suitable as a case study is the fact that Estonia's National Institute for Health Development reports infant deaths not only for the country as a whole, but by county. For example, it reports 35 infant deaths for the country as a whole in 2015.Among its 15 counties, zero infant deaths were reported in seven of them for 2015, while for those reporting at least one infant death, the number ranged from one in Jõgeva County to 15 in Harju County. Having the ability to compute IMRs for these counties (including those that are zero) allows one to compare IMRs based on reported data to the estimates aimed at portraying the underlying IMRs and their intrinsic mortality regimes.

(Exhibit 1 About Here)

Methods

There are two major components of the method introduced here. The first part of the methods section discusses the fundamental binomial nature of Infant mortality rates in that they are the proportion of births that result in deaths during the first year of life that constitute a beta binomial process. The second part of the methods section looks at the second component by extending the beta binomial process to a set of two estimates constituting samples of the mean and variance of the underlying process and argues that by averaging them one can produce a superior estimate of the mean proportion of births that result in deaths during the first year of life.

Part I: Infant Mortality Rates as a Beta Binomial Process

Infant mortality rates measure the proportion of births that result in deaths during the first year of life. As such, they measure the relationship between events (deaths) and trials (births) with the distribution of infant deaths in a given area i at a given time t is (approximately) binomial, with parameter d, where

$$d_{i,t} = D_{i,t}/B_{i,t} \qquad [1]$$

where

i = area (i = 1 to n)

t = time

D = infant deaths

B = births

and is typically described as a beta binomial random process with a probability mass function defined by two parameters: α and β. The first parameter, α, can be interpreted as the count of the event of interest, which in our case is the number of infant deaths, the number of births in which the infant dies before achieving the first year of life. The second parameter, β, can be interpreted as the count of "non-events," which in our case is the number of children born who survive to reach one year of age. Note that "rate" = α/(α + β), which in our case is equivalent to" infant mortality rate" = infant deaths/(infant deaths + survivors to age 1), which reduces to infant deaths/births. Thus, parameter α is the numerator in the expression defining a rate, and when added together, the parameters α and β represent the denominator. Together, the IMR may be re-

expressed the IMR as the compound distribution of α and β captured in the beta-binomial probability model:

$$IMR = \alpha/(\alpha + \beta) = \text{infant deaths}/(\text{infant deaths} + \text{infant survivors}) \quad [2]$$

Since the IMR may be conceptualized directly using the beta-binomial model, IMRs may be thought of as stochastic processes that occur within each county while also contributing to higher-level meta-populations within which they are nested (Taylor and Karlin, 2001; Graham and Talay, 2013).

Part II: An Indirect Estimator of IMR Using Averaging of Samples from a Beta-Binomial Stochastic Process

A potential number of strategies exist for dealing with small sample size dynamics or confidentiality suppression in making estimates of infant mortality rates. First, one might simply use the national IMR in place of highly-uncertain localized estimates of IMR. This would stabilize estimates for IMR on the local level, but at the expense of potentially masking heterogeneity in IMRs across geographic units. For purposes of capturing spatial patterns in IMR, a main priority in smaller-level analyses, this solution is less acceptable. A second alternative might be to make local adjustments based on judgment. While this may improve estimates overall, especially when judgments are made by applied demographers with significant experience, this approach is subject to the criticism that non-standard methods are applied across different geographies and/or population groupings. With resource allocation decisions often tied to demographic estimates, this solution may not be satisfactory either. An ideal approach would be to utilize a principled method for adjusting local estimates of IMR. Simple model averaging, based on the beta-binomial model represents a viable approach for achieving this goal.

Because it has been established that the IMR constitutes a beta-binomial probability process, think of two estimates of this process as constituting samples of the mean and variance of the underlying process. Therefore, these can be considered as samples obtained from the same underlying mortality process and in averaging them it can be anticipated that a superior estimate of the mean proportion is obtained (Graham and Talay, 2013; Gardiner, 1983; Taylor and Karlin, 2001). As such, the averages of two estimates based on the model may also be averaged as:

$$IMR_{averaged} = (\alpha_1 + \alpha_2)/ ((\alpha_1 + \beta_1) + (\alpha_2 + \beta_2)) \quad [3]$$

where the subscripts (1,2) now represent estimates of death and survivorship counts for two groups. This method can, of course, be extended to k groups as desired. Such model averaging yields an estimate where a larger-scale and representationally-appropriate model IMR is leveraged to make smaller-scale estimates more precise in a manner similar to that observed in the literature on indirect estimation in demography (Brass, 1968; Moultrie et al. 2013, Siegel and Swanson 2004, UN, 1967). Recent attempts to extend indirect estimation based on stochastic process theory have been introduced (Baker et al., 2011) and here this idea is leveraged further in developing indirect estimates of IMR based on model averaging.

Before turning to a discussion of the data, it is appropriate here to discuss in some detail the averaging process just described. Because an IMR is typically expressed per 1,000 births, it can be turned into a binomial variable by dividing it by 1,000 (or more generally if IMR is expressed as infant deaths per k births, it would be divided by k). In this form, IMR is strictly bound in that it cannot be less than zero nor greater than ($0 \leq IMR \leq 1$). In practice, it is substantially less than one. Once in this form, a Beta model (Binomial) can be fitted to a distribution of IMRs, which when fitted, produces two estimated parameters, α and β. The first

parameter, α, can be interpreted as the count of the event of interest, which in our case is the number of births in which the infant dies before achieving the first year of life. The second parameter, β, can be interpreted as the count of "non-events," which in our case is the number of children born who survive to reach one year of age. Note that "rate" = α/(α + β), which in our case is equivalent to" infant mortality rate" = infant deaths/(infant deaths + survivors to age 1), which reduces to infant deaths/births. Thus, parameter α is the numerator in the expression defining a rate, and when added together, the parameters α and β represent the denominator.

The two parameters estimated by fitting the Beta model to a distribution of IMRs are then used to adjust the reported infant deaths (a) and births (b) for the population in question, even when either one or both is equal to zero. The adjustment is straightforward: adjusted IMR = (a + α)/ ((a + b) + (α + β)). Note, as stated earlier that if a = zero then the adjusted IMR = α/(b + α + β) and that if both a and b are zero, then the adjusted IMR = α/(α + β).

DATA

As discussed in endnote 3, the 2015 county birth data used for Estonia are those reported by Statistics Estonia (2018) while the county infant death data are those reported by the National Institute for Health Development (2018). As the "representative" set of IMRs, 2015 birth and infant death data available from Eurostat (2018) for eight Baltic Sea countries (Denmark, Estonia, Finland, Germany, Latvia, Lithuania, Poland, and Sweden) are used. Table 1 shows the reported births and infant deaths by county for Estonia during calendar year 2015 and Table 2 provides the birth and infant death data for the eight Baltic Sea countries, along with their IMRs.[4]

(Tables 1 and 2 About Here)

RESULTS

The Beta Binomial model procedure found within the "survival/reliability" module of the NCSS statistical analysis package (release 8) was used to obtain the two Beta Model parameters using the infant mortality rates for the eight Baltic Sea countries (Table 2). The major results of interest found in running this procedure with the data are found as Exhibit 2. Note that there are two different estimates of the α and β parameters, one accomplished by the method of moments and the other by Maximum Likelihood Estimation. The parameters of the latter are used here, namely: $\alpha = 12.20081$, and $\beta = 3741.966$.

(Exhibit 2 About Here)

Table 3 shows the estimated 2015 underlying IMRs for the 15 counties found by using the two Beta parameters in conjunction with reported 2015 infant deaths and reported births by county using the averaging formulas described earlier.

(Table 3 About Here)

A Validity Test

Given that the method is producing a revised IMR that is likely to be close to the underlying IMR for a small population and therefore reflective of its intrinsic mortality regime, one would expect the method to do this where one could observe the intrinsic mortality regime. Model stable populations afford this opportunity because they have known intrinsic mortality regimes, the model life tables associated with a given set of model stable populations. To examine how the method works in this environment, I employed the IMR associated with a model stable population found in *Manual IV, Methods of Estimating Basic Demographic Measures from Incomplete Data* (1967). For this purpose, I selected the infant mortality rate associated with

West Level 23 for both sexes, which shows that of 100,000 births, 98,166 are expected to reach the first birthday. This yields an IMR of 0.0184 = 1 -.98166.

Using the IMR of 0.0184 and a seed population of 100,000, a random sample of 5,000 IMRs was generated using the Beta Model simulation provided by the NCSS statistical system (release 8). The sample is sufficiently large to allow the simulation program the opportunity to generate outliers, which it did. As can be seen in Exhibit 3, the mean is 0.01838 with a standard deviation of 0.000423 and a coefficient of variation equal to 0.02305. The minimum IMR is .016849 and the maximum is .020147.

(Exhibit 3 about Here)

From the 5,000 randomly generated observations, I extracted two sets of data. For the first set, I extracted the initial 43 IMR randomly generated observations from the simulation. For the second, I rank-ordered the 5,000 observations: from high to low and then from low to high, and extracted the eight highest IMR and seven lowest IMRs, respectively from them. The idea is that the entire set represents a synthetic population with 58 observations while the second set of 43 simulated IMRs represents the subset of the synthetic population in which IMRs are reported, and the third set of 15 simulated IMRs represents a subset of "small populations" subject to a high level of stochastic uncertainty. These characteristics mimic the 2009-2011 IMRs reported for the 58 counties of state of California, where the results are not reported for 15 counties (due to their small populations).[5] The 43 observations are expected to be closer, on average, to the "underlying" IMR of 0.01838 and have variation, respectively, than that found in the 15 observations. For the set of 43 observations, the mean IMR is 0.01834 and the coefficient of variation is .02305. For the set of 15 observations, the mean IMR is .01855 and the coefficient of variation is .07692. Thus, the set of 43 observations has a mean and a coefficient of variation

closer to the mean and coefficient of variation found in the full set of 5,000 observations than does the set of 15 observations.

A Beta model was fit to the set of 43 observations and its parameters were used to revise the IMRs in the set of 15 observations. The expectation is that the revised IMRs will yield a mean IMR closer to that found for the full 5,000 set of simulated observations and that the variation among these revised means will decline, yielding a smaller coefficient of observation.

The results show that the Beta model moved the initial IMR estimates for the 15 observations closer to the underlying IMR. As such, they are more reflective of the West Level 23 mortality regime that is intrinsic to them: the mean of the original IMRs for the 88 observations is 0.01855 while the mean for the revised IMRs is 0.01839, which is closer to the underlying IMR of 0.01838. In terms of variation, the coefficient of variation for the initial set of 14 IMRs is .07692, while that for the revised set is 0.00338. These results support the argument that the method described in this paper is capable of moving IMRs subject to stochastic uncertainty closer to the underlying IMRs and their respective intrinsic mortality regimes.[6]

DISCUSSION OF RESULTS

The estimated IMRs for the 15 Estonian counties in 2015 all appear reasonable. Supporting this argument is the fact that Harju County has the lowest estimated IMR (2.562) and is distinct in this respect from all of the other 14 counties. This can be seen in Exhibit 4, a scatterplot that displays IMRs by the driving distance from each of the county capitals to Tallinn, the capital of the country, which is located in Harju County. This graph suggests that there is no strong gradient in terms of increasing IMRs as distance from Tallinn increases. Rather, there is a distinct difference between Harju County and the other 14 counties. Looking at this difference in

terms of the variables that affect the occurrence of infant deaths, and in particular, socio-economic status, as well as health care quality and access, it is not surprising to find that Harju County has a wide range of medical services and a per-household 2015 disposable income of €11,743.54, which is well above any other county as well as the country as a whole (€10,102.21).

(Exhibit 4 About Here)

Although the estimated IMRs are subject to errors, the fact that Harju County is estimated to have a distinctly lower IMR than the other 13 counties supports the argument that they are valid, as does the clustering of IMRs found in the other 14 counties. However, it needs to be kept in mind that if the "larger context" selected for this illustration (The IMRs for the eight Baltic Sea countries of Denmark, Estonia, Finland, Germany, Latvia, Lithuania, Poland, and Sweden) was modified, then the estimated IMRs would also be modified. Here, consider the case of the Russian Federation, which borders Estonia, but unlike the eight Baltic Sea countries, is a huge country that encompasses a wide range of demographic characteristics and has a higher (estimated) IMR than any of the eight Baltic Sea countries. With an estimated 2016 IMR of 7.00 (The World Bank, https://data.worldbank.org/indicator/SP.DYN.IMRT.IN), it is well above the average IMR (3.25) for the eight Baltic Sea countries. Adding the Russian Federation to them as part of the "larger context," would change the two Beta model parameters from $\alpha = 12.20081$, and $\beta = 3741.966$ to $\alpha = 7.060866$ and $\beta = 1918.59$. This change does not affect the ranking of the Estonian counties by IMR, but it does affect the IMRs estimated for each of them. For example, the IMRs in all counties but Harju increase. As examples of the changes: (1) the estimated IMR for Harju County would decline from 2.562 to 2.510; (2) the IMR for Rapla County would increase from 3.511 to 4.088; and (3) the IMR for Hiiu County would increase from 3.190 to 3.538.

While the consistency found in the rank ordering suggests that the process is robust, the fact that the individual IMRs change illustrates the sensitivity of the estimated IMRs to the "larger context" selection. In this regard, it is clear that the estimates are subject to judgment. However, the entire process is transparent, which means that the results are not subject to arbitrary and capricious judgments that render them difficult to replication. Moreover, it makes sense that with a different model, one would have different IMR estimates. However, as the validity test indicates, a different model, can be expected to move, on average, the IMRs for the Estonian counties closer to their underlying IMRs, better reflecting their "underlying IMRs." This argument can be generalized to other potential data sets that could be used to build different beta-binomial models. This feature of the beta-binomial approach suggests that while a model built from a given "representational" data set may move the estimated IMRs closer, on average, to their underlying values, than a model built from a different "representational" data set, even a less-than-optimal model should provide reasonable estimates. Couple these features with the fact that the estimates can be efficiently generated by the process described here, suggests that they have the potential to support policy decisions while keeping time and resource requirements low; characteristics that Swanson and Tayman (2012: 304) suggest are important components in deciding what methods to use in developing estimates. Given this, what are the implications for studying and monitoring infant mortality among immigrant populations in Europe?

First, note that while these immigrant populations tend to be clustered and segregated within countries, their aggregate number across all the European Union is not small. Salt (2011: 20) reports that by 2008 nearly 31 million foreign citizens lived in the EU's 27 member countries, up from around 26 million in 2004. He finds that Turks were the largest group at 2.4 million and constituting 7.9% of all non-nationals, followed by Moroccans at 1.7 million, (5.6%)

and Romanians at 1.7 million, constituting 5.4% of all non-nationals (Salt 2011: 22). Eurostat (2017) reported that there were 35.1 million people born outside of the EUs 28 member countries in 2016. With a total EU population of about 511 million in 2016, the foreign-born accounts for about 6.9 percent of the EU's 28 member countries.

With up to 35.1 million of the EU's residents likely to be in spatially segregated clusters and characterized by higher mortality than other populations, it appears that a method for estimating the underlying death rates of small populations and thereby gaining a picture of their intrinsic mortality regimes, would appear to be useful, especially given the call for better health monitoring of these populations (Rechel, Mladovsky and Devillé. 2011). While the method described and evaluated in this paper is aimed at the estimation of infant mortality rates, it can be adapted to the estimation of other mortality rates. Given the small populations of several of the Estonian counties (e.g., Hiiu County's population in 2015 is only 8,582 while that of Lääne county is 24,070 (Statistics Estonia, 2017)), it would appear that even in the absence of data on infant deaths, the method could be used for sub-county estimates, given the availability of birth data. Knowledge of the location of spatially concentrated immigrant populations could be entered into GIS-based data systems in which birth data for same spatial areas could be layered (Chung, Yang, and Bell, 2004). This would set the stage for the estimation of underlying IMRs and other mortality data for these populations and thereby revealing an idea of their intrinsic mortality regimes.[7]

CONCLUDING REMARKS

While the beta-binomial model has been used in medical research (Kim and Lee, 2013; Arostegui, Nuñez-Antón, and Quintana, 2007, and Young-Xu and Chan, 2006), consumer studies (Chatfield and Goodhardt, 1970), bioinfomatics (Pham et al., 2010) and public health

research (Alanko and Lemmons, 1996; Gakidou and King. (2002), it has not found much traction in demographic research. This is surprising on two counts: (1) the components of demographic change, births, deaths, and migration, can all be constructed as rates that are inherently binomial variables; and (2) the method is simple to use, explain, and understand. [8] This paper illustrates one such use with a sub-set of the mortality component, the infant mortality rate. Although the paper focuses on a European application, namely the concentration of ethnic migrant groups, the method can be applied to many other situations where small numbers are present and affected by stochastic uncertainty. As such, it could be used in conjunction not only with other mortality measures such as neo-natality rates, crude death rates, age-specific death rates and cause-specific death rates, but with fertility measures such as crude birth rates and age-specific birth rates. Even more broadly, it could be used with any binomial variable of interest affecting small populations, such as a housing occupancy (or vacancy) rate, employment (or unemployment) rate, cigarette smoking (or non-smoking) rate.

ENDNOTES

1. Murray (1996) has argued that the infant mortality rate is flawed when it is used as an index of overall mortality (i.e., the mortality regime affecting a given population) and that Disability Adjusted life Expectancy (DALE) should be used in its place. However, it has been pointed out by Reidpath and Allotey (2003) that the infant mortality rate and the DALE are so highly correlated that it merely goes to reinforce the intuition that the causes of infant mortality are strongly related to those structural factors like economic development, general living conditions, social well-being, and environmental factors, and, and such, the infant mortality rate remains a useful and comparatively inexpensive indicator of population health. Guillot et al. (2013) also note that infant mortality is very useful because it involves a short lag between the timing of mortality exposures and the timing of corresponding births.

2. I use a classic definition of stochastic uncertainty provided by Doob (1952), namely that it is the manifestation of a process representing numerical values of some system randomly changing over time,

3. The BIRTH data for Estonia are all taken from the online query system available through Statistics Estonia. The entry point for the system is http://pub.stat.ee/px-web.2001/dialog/statfile1.asp. Once there, birth data by county are found by going through the Population database and into the "vital events" folder. In this folder one can obtain births for 2015 by county by selecting the subfolder, "PO11: LIVE BIRTHS BY COUNTY -Modified: 02.06.2017" and following the query system instructions.

   The entry point for obtaining reported infant deaths for 2015 is through the online query system made available by Estonia's National Institute for Health Development, which is similar to the one found at Statistics Estonia: http://pxweb.tai.ee/PXWeb2015/pxweb/en/01Rahvastik/01Rahvastik__04Surmad/SD50.px/?rxid=2cf1fa43-7457-4166-a3c7-e43fdb0a9b97 . This link will take you to "SD50: Infant deaths by sex, county and age at the moment of death." Once there, follow the query system instructions.

   The 2015 per-household disposable income by county was obtained by going through the Social Life database and selecting the "income" folder. Once there, the 2015 disposable income per household can be obtained by selecting the subfolder, "IM15: EQUALISED YEARLY DISPOSABLE INCOME BY COUNTY AND SEX -Modified: 18.12.2017" and following the query system instructions.

   The 2015 population by ethnicity (Estonian, Russian and "other ethnic nationalities") and county is located in the subfolder " PO0222: POPULATION BY SEX, ETHNIC NATIONALITY AND COUNTY, 1 JANUARY." Once there, select 2015 in the first box as the year, both males and females in the second box , all counties in the third box, and Estonians, Russians, and other ethnic nationalities in the fourth and final box.

   The files assembled from these data and all other data files constructed for use in this study are available from the author.

4. The births and infant deaths are for the calendar year 2015, so the denominator (the number of births during calendar year 2015) is temporally consistent with the numerator (the number of infant deaths during calendar year 2015) as a period measure. The population and income data employed in the discussion are for January 1st, 2015 and are consistent with each other. There was no need to center the birth and infant death data on the population data because the latter are not used in constructing the infant mortality rate.

5. Note that as stated in the text, the validity test mimics the fact that for its 58 counties California reports IMRs only for 43 of them for the 2009-11 period, leaving the remaining 15 counties without reported IMRs, As such, the validity test was set up as if there were 43 units for which IMRs were reported and 15 for which they were not, However, all of the data used in the validity test were generated from the synthetic population that is based on Model Life Table, Level 23, as described in the text. The reporting structure as well as the actual data for California can be found through the Open

Portal service provided by the California Health and Human Services Agency via a download of a CVS data set assembled by the California Department of Public Health. This data set can be accessed by going to

https://data.chhs.ca.gov/dataset/infant-mortality-deaths-per-1000-live-births-lghc-indicator-01/resource/ae78da8f-1661-45f6-b2d0-1014857d16e3

and then clicking on the "download" tab, which downloads the file, "Infant Mortality, Deaths Per 1,000 Live Births (LGHC Indicator 01) (CSV)" in CVS form. Once downloaded, it can be saved as an excel file. The data in this file include the infant mortality rates (identified as "rate" in the file) and the infant deaths (identified as "numerator" in the file) and live births (identified as "denominator" in the file) used to calculate the IMRs for all counties and other administrative areas, including the state as a whole. The data represent the period 2009-11. A description of the methods, caveats, and so forth associated with this data set can be found on the ULR shown above.

6. In the validity test, different populations are simulated from a common beta distribution, and the result is that the two sets of populations, large and small, are normally distributed around the intrinsic mean IMR of the "population." The simulation shows that the adjusted IMRs of the small populations move closer the underlying IMR, which indicates that the method works when both the small and large populations represent samples taken from the same underlying population. If the small populations represent a sample from a different population than the sample of large population, then the adjustment may yield a "biased" estimate of the former's underlying IMR. This shows the importance of having a reference set that conceptually represents a sample from the same underlying population as the small population sample. One way to visualize the unbiased and biased outcomes is to picture the case where the method yields: (1) an "unbiased" estimate, which is when the mean IMR of the large populations is between the underlying IMR and the mean IMR of the small populations; and (2) a "biased" estimate when the method does not move the mean IMR for the small population closer to its underlying IMR, which occurs where the mean IMR of the small population is between the underlying IMR and the mean IMR of the large populations.

7. Keep in mind that small populations, however defined, with approximately the same total populations may have different age compositions. For example, one may have a relatively large aged population and another a relatively large young population. This simple example is meant to illustrate the effect of demographic heterogeneity, which can affect measures of mortality (Vaupel and Missov, 2014). In situations where substantial heterogeneity may be present, a model with additional covariates may prove useful because the latter can potentially take into account the effects of demographic heterogeneity.

8. Although Green and Armstrong (2015) discuss simple vs. complex methods in terms of forecasting, their discussion applies here in that the beta-binomial approach falls into the simple methodological category rather than the complex category. Adapting their discussion to methods in general, the work of Green and Armstrong (2015) suggests that while there is no evidence that shows complexity improves accuracy, complexity remains

popular among: (1) researchers, because they are rewarded for publishing in highly ranked journals, which favor complexity; (2) methodologists, because complex methods can be used to provide information that support decision makers' plans; and (3) clients, who may be reassured by incomprehensibility. I believe that the argument by Green and Armstrong (2015) can be applied to Bayesian methods, which represents the "complex" alternative to the "simple" Beta-binomial approach. I prefer the Beta-binomial approach, however, not only because of the argument presented by Green and Armstrong, but also because the application of a Bayesian approach can be difficult, effortful, opaque and even counter-intuitive (Goodwin 2015).

REFERENCES

Alanko, T., and P. Lemmens (1996). Response effects in consumption surveys: An application of the Beta-binomial model to self-reported drinking frequencies. *Journal of Official Statistics* 12 (3): 253-273.

Andersson, R. (2013). Reproducing and reshaping ethnic residential segregation in Stockholm: The role of selective migration moves. *Geografiska Annaler*: Series B, Human Geography, 95(2), 163–187.

Arostegui I., V. Nuńez-Antón, and J. Quintana (2007). Analysis of the Short Form-36 (SF-36): The beta binomial distribution approach. *Statistics in Medicine* 26: 1318-1342

Baker J, A. Alcantara, X. Ruan. (2011). A stochastic version of the Brass PF ratio adjustment of age-Specific fertility schedules. *PLoS One*. 6(8):e23222.

Brass W., A. Coale, P. Demeny, D. Heisel, F. Lorimer, A. Romaniuk, and E. Van de Walle (196×8). *The Demography of Tropical Africa*. Princeton: Princeton University Press.

Bråmå, S. (2008). Dynamics of ethnic residential segregation in Göteborg, Sweden, 1995–2000. *Population, Space and Place* 14(2), 101–117.

Chatfield, C., and G. Goodhardt (1970). The Beta-Binomial Model for Consumer Purchasing Behavior. *Applied Statistics* 19:240–250.

Chen, A., E. Oster, and H. Williams. (2016). Why is infant mortality higher in the United States than in Europe? *American Journal of Economic Policy* 8(2): 89–124.

Chung, K., D. Yang, and R. Bell (2004). Health and GIS: Toward spatial statistical analyses. *Journal of Medical Systems* 28 (4): 349 – 360.

Doob, J. (1953). *Stochastic Processes*. New York. John Wiley & Sons.

Eurostat (2017). Migration and migrant population statistics (available online at http://ec.europa.eu/eurostat/statistics-explained/index.php/Migration_and_migrant_population_statistics

Eurostat (2018). The ratio of the number of deaths of children under one year of age during the year to the number of live births in that year (excel file download, available at http://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&plugin=1&language=en&pcode=tps0002 7).

Gakidou, E., and G. King. (2002). Measuring total health inequality: adding individual variation to group-level differences. *International Journal of Equity in Health* 1:3 doi: 10.1186/1475-9276-1-3

Gardiner, C. (1983*). Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*. New York: Springer.

Goodwin, P. (2015). When simple alternatives to Bayes formula work well: Reducing the cognitive load when updating probability forecasts. Journal of Business Research 68: 1686-1691.

Graham, C., and D. Talay. (2013). *Stochastic Simulation and Monte Carlo Methods*. New York: Springer.

Green, K. and J.S. Armstrong. (2015). Simple versus complex forecasting: The evidence. Journal of Business Research 68: 1678-1685.

Guillot, M., S. Lim, L. Torgasheva, and M. Denisenko. (2013). Infant mortality in Kyrgyzstan before and after the break-up of the Soviet Union. *Population Studies* 67(3): 335-352.

Hummer, R. (2005). Income, Race, and Infant Mortality: Comment on Stockwell et al. *Population Research and Policy Review* 24: 405‑409.

Kalandides, A., & D. Vaiou. (2012). 'Ethnic' neighbourhoods? Practices of belonging and claims to the city. *European Urban and Regional Studies* 19 (3): 254–266.

Kim J., and J. Lee. (2013). Simultaneous confidence intervals for a success probability and intraclass correlation, with an application to screening mammography. *Biometrical Journal* 55 (6):944–954

Kinge, J. and T. Kornstad. (2014). Assimilation effects on infant mortality among immigrants to Norway: Does maternal source country matter? A *Demographic Research* 31 (available online at https://www.demographic-research.org/volumes/vol31/26/default.htm)

Kitagawa, E. and P. Hauser. (1973). *Differential Mortality in the United States: A Study in Socioeconomic Epidemiology*. Cambridge: Harvard University Press.

Kleinman, J. (1996). Underreporting of infant deaths: Then and now. *American Journal of Public Health* 76 (4): 365-366.

Link, B., and J. Phelan. (1995). Social Conditions as Fundamental Causes of Disease. *Journal of Health and Social Behavior* (extra issue): 80‑94.

Link, W. and D. Hahn. (1996). Empirical Bayes Estimation of proportions with application to cowbird parasitism rates. *Ecology* 77 (8): 2528-2537.

Ljunggren, J., and P. Andersen, (2015). Vertical and horizontal segregation: Spatial class divisions in Oslo, 1970–2003. *International Journal of Urban and Regional Research* 39 (2): 305–322.

Misra, D., H. Grason, M. Liao, D. Strobino, K. McDonnell, and A. Allston. (2004). The nationwide evaluation of fetal and infant mortality reviewed (FIMR) programs: development and implementation of recommendations and conduct of essential maternal and child health services by FIMR programs. *Maternal Child Health Journal* 8(4); 217-229.

Moultrie, T., R. Dorrington, A. Hill, K. Hill, I. Rob Dorrington, Allan Hill, I. Timæus and B. Zaba. (2013). *Tools for Demographic Estimation*. Paris: International Union for the Scientific Study of Population.

Murray C. (1996). Rethinking DALYs. pp. 1-98 in C. Murray and A. Lopez (eds.) *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. Cambridge, MA: Harvard School of Public Health

National Institute for Health Development (2018). *SD50: Infant deaths by sex, county and age at the moment of death (*see Endnote 1 for details on the use of the agency's data query system for accessing Estonia's infant death data that are used in this study.

Office for National Statistics. (2015). *Small populations tables from the 2011 Census – user guide*. Newport, South Wales, United Kingdom.

Pham, T., S. Piersma, M. Warmoes, and C. Jimenez (2010). On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. *Bioinformatics* 26 (3): 363–369.

Rechel, B., P. Mladovsky, and W. Devillé. (2011). Monitoring the health of migrants. pp. 82 – 98 in B. Rechel, P. Mladovsky, W. Devillé, B. Rijks, R. Petrova-Benedict, and M. Mckee (Eds.). *Migration and health in the European Union. Berkshire, England*. Open University Press.

Reeske, A., and O. Razum. (2011). Maternal and child health – from conception to first birthday. pp. 139 – 153 in B. Recehl, P. Mladovsky, W. Devillé, B. Rijks, R. Petrova-Benedict, and M. Mckee (Eds.). *Migration and health in the European Union. Berkshire, England*. Open University Press.

Reidpath, D. and P. Allotey (2003). Infant mortality rate as an indicator of population health. *Journal of Epidemiology and Community Health* 57: 344-346.

Salt, J. (2011). Trends in Europe's international migration. pp. 17- 36 in B. Recehl, P. Mladovsky, W. Devillé, B. Rijks, R. Petrova-Benedict, and M. Mckee (Eds.). *Migration and health in the European Union. Berkshire, England*. Open University Press.

Siegel, J. and D. Swanson. (2004). *The methods and materials of demography, 2nd Edition*. Los Angeles: Academic/Elsevier Press.

Statistics Canada. (2017). *Infant and perinatal mortality, by sex, three-year average, Canada, provinces, territories, health regions and peer groups*. (Cansim, Table 102-4319, available at http://www5.statcan.gc.ca/cansim/a26?lang=eng&retrLang=eng&id=1024319&&pattern=&stByVal=1&p1=1&p2=-1&tabMode=dataTable&csid=

Statistics Estonia (2017). See Endnote 1 for details on the use of the agency's data query system for accessing Estonia's birth, income, and population data that are used in this study.

Stockwell, E., F. Goza, and K. Balisteri. (2005). Infant Mortality and Socioeconomic Status: New bottle, Same Old Wine." *Population Research and Policy Review* 24: 387‑39

Stockwell, E., M. Bedard, D. A. Swanson, and J. Wicks. (1987). Public Policy and the Socioeconomic Mortality Differential in Infancy. *Population Research and Policy Review* 6 (Fall):105-121.

Swanson, D. and J. Tayman. (2012). *Subnational Population Estimates*. Dordrecht, The Netherlands: Springer.

United Nations. (1967). *Manual IV, Methods of Estimating Basic Demographic Measures from Incomplete Data.* New York, NY: United Nations.

US National Center for Health Statistics (no date). Data Use Restrictions (https://wonder.cdc.gov/datause.html)

US National Center for Health Statistics (2018). *Linked Birth / Infant Death Records 2007-2015, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program, on CDC WONDER On-line Database,* http://wonder.cdc.gov/lbd-current.html.

VanEenwyk, J. and S. Macdonald. (2012). *Guidelines for Working with Small Numbers.* Olympia, WA: Washington Department of Health, Environmental Public Health Division. (available at https://www.doh.wa.gov/Portals/1/Documents/1500/SmallNumbers.pdf

Vaupel, J, and T. Missov. Unobserved population heterogeneity: A review of formal relationships. *Demographic Research* 31 (22): 659-686.

Voss, P. R., C. Palit, B. Kale, and H. Krebs. (1995). Censal ratio methods. pp.70–89 in N. W. Rives, W. J. Serow, A. S. Lee, H. F. Goldsmith, and P. R. Voss (Eds.) *Basic methods for preparing small-area estimates.* Madison: Applied Population Laboratory, University of Wisconsin.

Young-Xu, Y. and K. Chan. (2008). Pooling overdispersed binomial data to estimate event rate BMC *Medical Research Methodology* 8:58

Table 1. Infant Deaths and Births by County, Estonia, 2015

Area	2015 Total Infant Deaths	2015 LIVE BIRTHS
Harju county	15	13,907
Hiiu county	0	6,864
Ida-Viru county	3	70
Jõgeva county	1	1,222
Järva county	0	289
Lääne county	0	263
Lääne-Viru county	3	218
Põlva county	0	594
Pärnu county	3	206
Rapla county	2	807
Saare county	0	291
Tartu county	6	306
Valga county	0	1,747
Viljandi county	0	278
Võru county	2	453
Estonia	35	298
Unknown	N/A	1

Sources: the births are taken from Statistics Estonia (2017) and the deaths from Estonia's National Institute of Health Development (2018)

Table 2. Infant Mortality, Infant Deaths and Live Births in 2015 for Eight Baltic sea countries.

COUNTRY/REGION	2015 Infant mortality rate (PER 1000 LIVE BIRTHS)	2015 Infant mortality rate/1000	2015 INFANT DEATHS (=IMR/1000*BIRTHS)	2015 Number of live births
Denmark	3.7	0.0037	215	58,205
Germany	3.3	0.0033	2,434	737,575
Estonia	2.5	0.0025	35	13,907
Latvia	4.1	0.0041	90	21,979
Lithuania	4.2	0.0042	132	31,475
Poland	4	0.004	1,477	369,308
Finland	1.7	0.0017	94	55,472
Sweden	2.5	0.0025	287	114,870

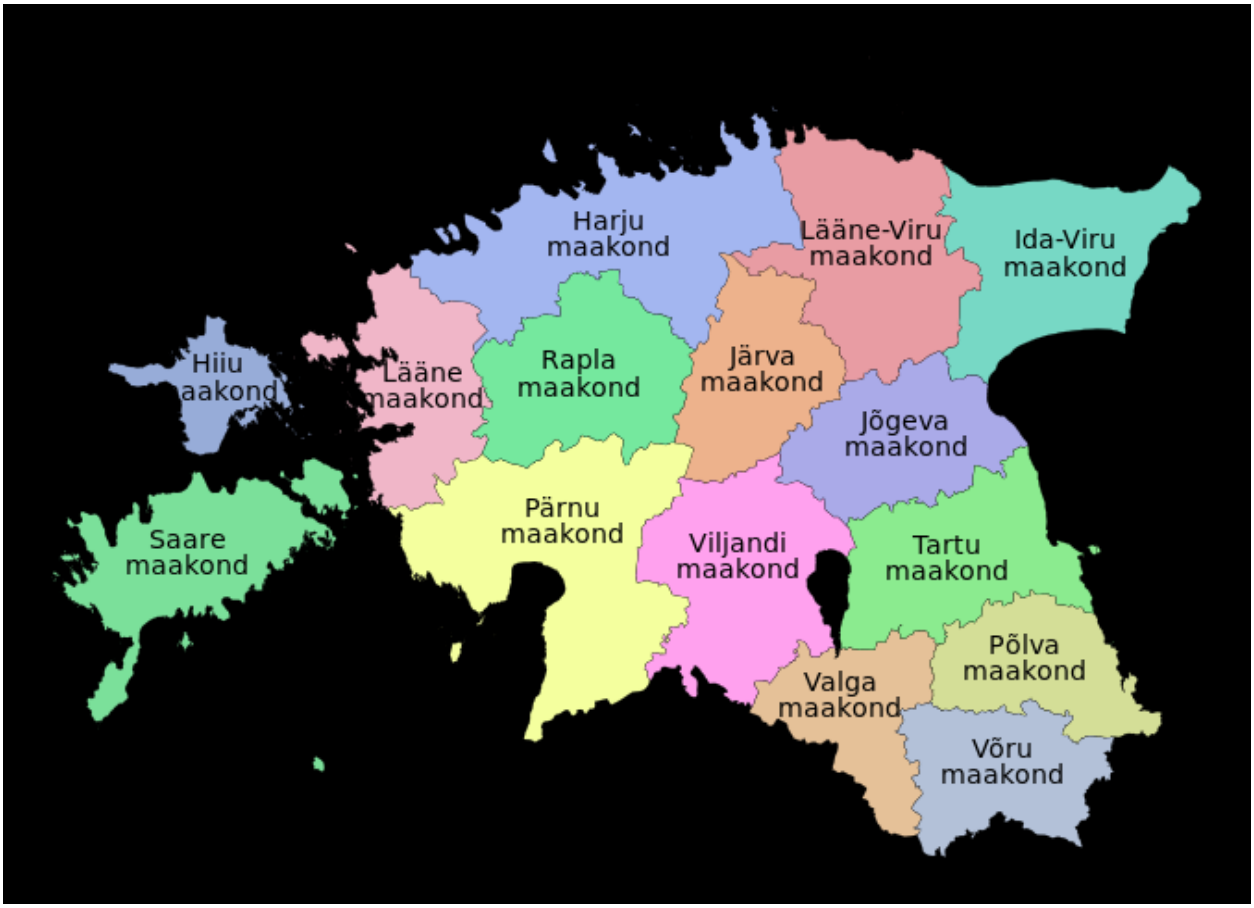
Source: Eurostat (2018)

Table 3. Reported IMR and the Estimated Underlying IMR by County in Estonia for 2015

AREA	IMR (PER 1000) BASED ON REPORTED INFANT DEATHS & BIRTHS	ESTIMATED UNDERLYING IMR (PER 1000)
ESTONIA	2.5167	2.673
Harju county	2.1853	2.562
Hiiu county	0.0000	3.190
Ida-Viru county	2.4550	3.055
Jõgeva county	3.4602	3.265
Järva county	0.0000	3.037
Lääne county	0.0000	3.072
Lääne-Viru county	5.0505	3.496
Põlva county	0.0000	3.081
Pärnu county	3.7175	3.333
Rapla county	6.8729	3.511
Saare county	0.0000	3.005
Tartu county	3.4345	3.309
Valga county	0.0000	3.026
Viljandi county	0.0000	2.900
Võru county	6.7114	3.504

Sources cited elsewhere. Calculations are by the author.

Figure 1. Map of Estonia by County



Source: ww.worldgenweb.org

Figure 2. NCSS Report on the Fit of the Beta Model to the IMRs of Eight Baltic Sea Countries

Parameter Estimation Section

Parameter	Method of Moments Estimate	Maximum Likelihood Estimate	MLE Standard Error	MLE 95% Lower Conf. Limit	MLE 95% Upper Conf. Limit
Minimum (A)	0	0			
Maximum (B)	1	1			
α	12.4456	12.20081	6.018849	0.4040834	23.99754
β	3816.97	3741.966	1884.298	48.80914	7435.123
Log Likelihood		-44.7225			
Mean	0.00325	0.003249938			
Median	0.003163931	0.003162153			
Mode	0.002990425	0.002985158			
Sigma	0.0009196273	0.0009287869			

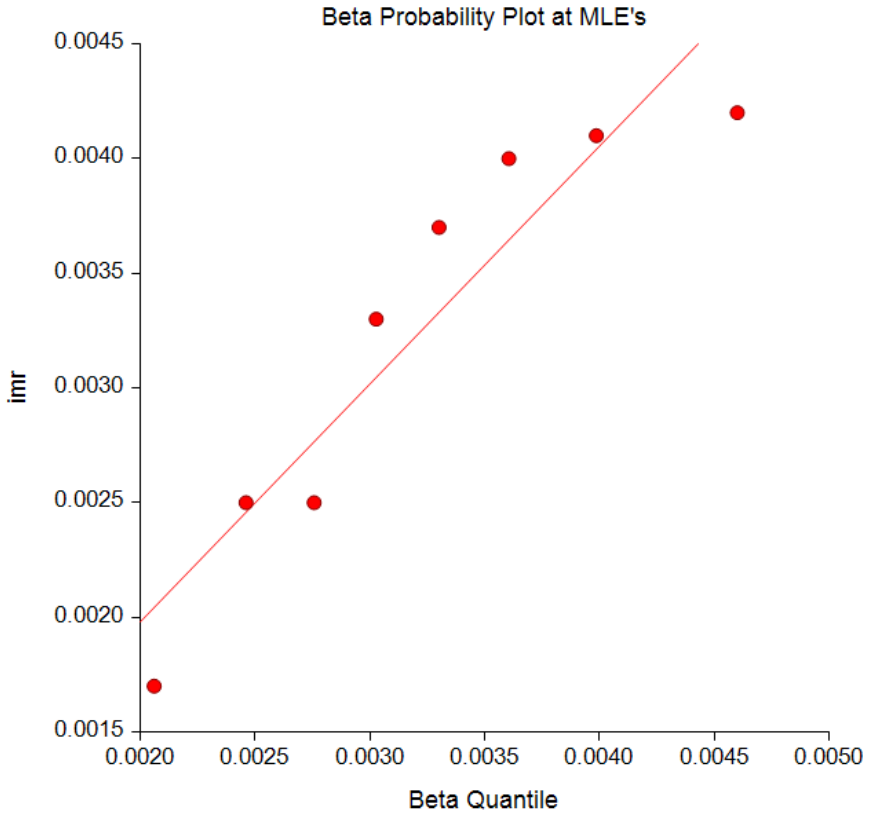
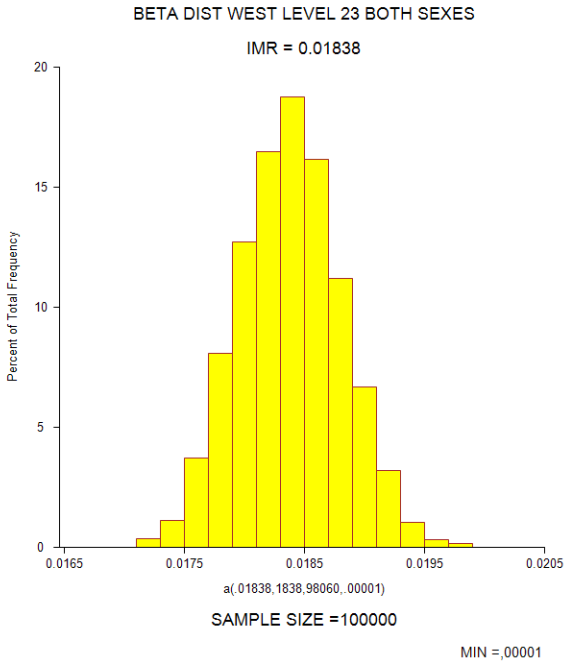


Figure 3. Characteristics of the Synthetic Population used in the Validity Test

Data Simulation Report
Histogram Section of Simulated Data



Descriptive Statistics of Simulated Data

Statistic	Value	Statistic	Value
Mean	0.01838248	Minimum	0.01684878
Standard Deviation	0.0004237251	1st Percentile	0.01744547
Skewness	0.07195781	5th Percentile	0.01769227
Kurtosis	2.934381	10th Percentile	0.0178332
Coefficient of Variation	0.02305049	25th Percentile	0.01808544
Count	5000	Median	0.01838394
		75th Percentile	0.01866444
		90th Percentile	0.0189272
		95th Percentile	0.01908542
		99th Percentile	0.01937772
		Maximum	0.02014658

Figure 4. Scatterplot of IMRs for Counties by distance from County Capital to Tallinn, the National Capital