CS-BIGS 7(1):14-25 http://www.csbigs.fr

Shrinkage estimation of rate statistics

Einar Holsbø
Department of Computer Science, UiT — The Arctic University of Norway
Vittorio Perduca
Laboratory of Applied Mathematics MAP5, Université Paris Descartes

This paper presents a simple shrinkage estimator of rates based on Bayesian methods. Our focus is on crime rates as a motivating example. The estimator shrinks each town's observed crime rate toward the country-wide average crime rate according to town size. By realistic simulations we confirm that the proposed estimator outperforms the maximum likelihood estimator in terms of global risk. We also show that it has better coverage properties.

Keywords : *Official statistics, crime rates, inference, Bayes, shrinkage, James-Stein estimator, Monte-Carlo simulations.*

1. Introduction

1.1. Two counterintuitive random phenomena

It is a classic result in statistics that the smaller the sample, the more variable the sample mean. The result is due to Abraham de Moivre and it tells us that the standard deviation of the mean is $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, where n is the sample size and σ the standard deviation of the random variable of interest. Although the equation is very simple, its practical implications are not intuitive. *People have erroneous intuitions about the laws of chance*, argue Tversky and Kahneman in their famous paper about the law of small numbers (Tversky and Kahneman, 1971).

Serious consequences can follow from small-sample inference ignoring deMoivre's equation. Wainer (2007) provides a notorious example: in the late 1990s and early 2000s private and public institutions provided massive funding to small schools. This was due to the observation that most of the best schools—according to a variety of performance measures—were small. As it turns out, there is nothing special about small schools except that they are small: their over-representation among the best schools is a consequence of their more variable performance, which is counterbalanced by their over-representation among the worst schools. The observed superiority of small schools was simply a statistical fluke.

Galton (1886) first described another stochastic mechanism that is dangerous to ignore. Galton observed that children of tall (or short) parents usually grow up to be not quite as tall (or short), i.e. closer to average height. Today we know this phenomenon as regression to the mean, and we will find it wherever we find variation. Imagine a coach who berates a runner who had an unusually slow lap time and finds that, indeed, the next lap is faster. The coach, who always berates slow runners, has not had the opportunity to realize that the next lap is very likely to be faster no matter

what. As long as there is variability in lap time we will some times see unusually slow laps that we can do nothing about and make no inference from. In this case too do our intuitions about the laws of chance fail us. People, including scientists, make the mistake of ignoring regression all the time. Mathematically regression to the mean is as simple as imperfect correlation between instances.

1.2. *These phenomena in official statistics*

The small-schools example is egregious because it led to wasteful public spending. The statistics themselves were probably fine, but their interpretation was not careful enough. Such summary statistics are often presented without regard for uncertainty. For instance, every year Statistics Norway (ssb.no), the central bureau of statistics in Norway, presents crime report counts. The media usually reports these numbers as rates and inform us that some small town that few people know about is the *most criminal* in the country. Often the focus is on violent crimes. Figure 1 below shows these rates for 2016. Not knowing de Moivre's result it might be striking to observe that many of the towns with the highest rates are small towns. Similarly, not knowing regression it might be striking to observe that, on average, towns with a high rate in one year will have a lower one in any other year, see Figure 2 below. These are unavoidable stochastic phenomena. Thus there is reason to believe that we should somehow adjust our expectations about these numbers. We will see below that such an adjustment also makes statistical sense.

1.3. *Shrinkage estimation*

There is an astonishing decision-theoretic result due to Charles Stein: suppose that we wish to estimate $k \geq 3$ parameters $\theta_1, \ldots, \theta_k$ and observe k independent measurements, $x_1 \ldots x_k$, such that $x_i \sim N(\theta_i, 1)$. There is an estimator of θ_i that has uniformly lower risk, in terms of total quadratic loss, than the obvious candidate x_i (Stein, 1956). In other words, the maximum likelihood estimate is inadmissible. Stein showed this by introducing a lower-risk estimator that biases or *shrinks*, the x_is toward zero. James and Stein (1961) introduced an improved shrinkage estimator, which we will see below. Efron and Morris (1973) show a similar result and a similar estimator for shrinking toward the pooled mean. There are many successful applications of shrinkage estimation, see for instance the examples from Morris (1983). The common theme is a setting where the statistician wants to estimate many similar variable quantities.

1.4. *An almost-Bayesian estimator*

In this case study we consider the official Norwegian crime report counts. We assume that in a given year the number of crimes reported in town i, denoted k_i, corresponds to the number of criminal events in this town. We further assume that each inhabitant can at most be reported for one crime a year. Our goal is to estimate the *crime probability* θ_i: probability that a person will commit a crime in this town. The obvious estimator is the maximum likelihood estimate (MLE) for a binomial proportion $\hat{\theta}_i = {}^{k_i}/_{n_i}$, where n_i is the population of town i.

The MLE binomial model rests on an assumption that inhabitants commit crimes independently according to an identical crime probability. There are reasons to believe that this is not the case. The desperately poor might be more prone to stealing than the middle class professional. There is a weaker assumption called *exchangeability* that says that individuals are similar but not identical. More precisely we assume that their *joint* criminal behavior (some number of zeros and ones) does not depend on knowing who the individuals are (the order of the zeros and ones). It is an important theorem in Bayesian inference, due to De Finetti, that a sequence of exchangeable variables are independent and identically distributed conditional on an unknown parameter θ_i that is distributed according to an a priori (or prior) distribution $f(\theta_i)$ (Spiegelhalter et al., 2004). In the binomial sense, θ_i has the remarkable property that it is the long-run frequency with which crimes

occur regardless of the i.i.d. assumption; the prior precisely reflects our opinion about this limit. By virtue of De Finetti's theorem, the exchangeability assumption justifies the introduction of the unknown parameter θ_i in a binomial model for k_i, so long as we take the prior into account.

To make an argument with priors is to make a Bayesian argument. Shrinkage is implicit in Bayesian inference: observed data gets pulled toward the prior (and indeed the prior is pulled toward the data likelihood). We propose an almost Bayesian shrinkage estimator, $\hat{\theta}_i^s$, that accounts for the variability due to population size. Our estimator is *almost* Bayesian because we do not treat the prior very formally, as will be clear below.

In a Bayesian argument we treat θ_i as random. The statistician specifies a prior distribution $f(\theta_i)$ for the parameter that reflects her knowledge (and uncertainty) about θ_i. As in the frequentist setting, she then selects a parametric model for the data given the parameters, which allows her to compute the likelihood $f(x|\theta_i)$. Inference about θ_i consists of computing its posterior distribution by Bayes' theorem:

$$f(\theta_i|x) = \frac{f(x|\theta_i)f(\theta_i)}{\int f(x|\theta_i)f(\theta_i)\,\mathrm{d}\theta_i}.$$

There are various assessments we could make about the collection of θ_i. If we assume they are identical we can pool them and use a single prior. If we assume they are independent we specify one prior for each and keep them separate. If we assume they are exchangeable—similar but not identical—it follows from De Finetti that there is a common prior distribution conditional on which the $\theta_1, \ldots, \theta_m$ are i.i.d. (Spiegelhalter et al., 2004).

We make this latter judgment and take a beta distribution common to all crime probabilities as prior. Our likelihood for an observed number of crime reports follows a binomial distribution. It is a classic exercise to show that the posterior distribution of θ_i is then also a beta distribution. The problem remains how to choose the parameters for the prior. On the idea that a given town is probably not that different from all the other towns, we will simply pool the observed crime rates for all towns and fit a beta distribution to this ensemble by the method of moments.

Under squared error loss, the posterior mean as point estimate minimizes Bayes risk. The posterior mean serves as our shrinkage estimate, $\hat{\theta}_i^s$, for θ_i. We will see that $\hat{\theta}_i^s$ in effect shrinks the observed crime rate $\hat{\theta}_i$ toward the country-wide mean $\bar{\theta} = \sum \frac{1}{m}\hat{\theta}_i$ by taking into account the size of town i.

Bayesian inference allows for intuitive uncertainty intervals. In contrast to a classical frequentist confidence interval, which can be tricky to interpret, we can say that θ_i lies within the Bayesian credible interval with a certain probability. This probability is necessarily subjective, as the prior distribution is subjective. We will conduct simulations to compare the coverage properties of our estimator to the classical asymptotic confidence interval.

1.5. Resources

This case-study is written with a pedagogical purpose in mind, and can be used by advanced undergraduate and beginning graduate students in statistics as a tutorial around shrinkage estimation and Bayesian methods. We will mention some possible extensions in the conclusion that could be the basis for student projects. Data and code for all our analyses, figures, and simulations are available at https://github.com/3inar/crime_rates

2. Data

We will work with the official crime report statistics released by Statistics Norway (SSB) every year. These data contain the number of crime reports in a given Norwegian town in a given year. The counts are stratified by crime type, e.g. violent crimes, traffic violations, etc.

We will focus on violent crimes. SSB separately provides yearly population statistics for each town. Figure 1 shows the 2016 crime rates (i.e. counts per population) for all towns in Norway against their respective populations. This is some times called a funnel plot for the funnel-like tapering along the horizontal axis: a shape that signals higher variance among the smaller towns.

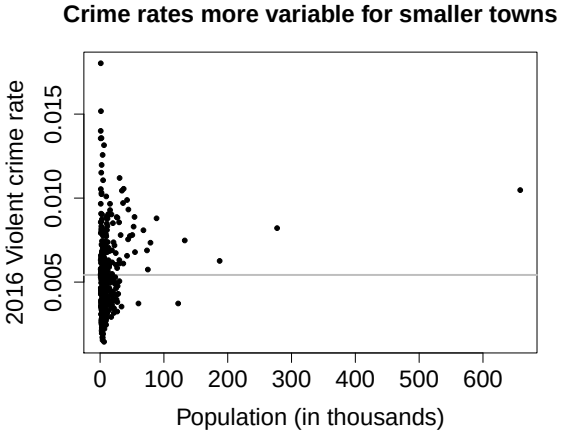

Figure 1: Rates of violent crime vs population in 2016 for all towns in Norway. The grey line shows the country-wide mean.

Figure 2 compares the crime rates in 2015 with those in 2016 and shows that the more (or less) violent towns in 2015 were on average less (or more) violent in 2016. The solid black line regresses 2016 rates on 2015 rates. The dashed grey line is what to expect if there were no regression toward the mean. It has an intercept of zero and a slope of unity. The solid grey line is the overall mean in 2016. The most extreme town in 2015, past .025 on the x-axis, is much closer to the mean in 2016. The solid black regression line shows that this is true for all towns on average. The fact that 2015 and 2016 are consecutive years is immaterial; regression to the mean will be present between any two years.

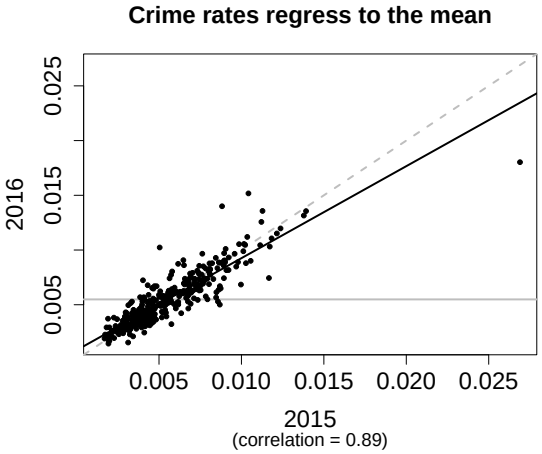

Figure 2: Regression to the mean from year to year. The plot compares 2016 and 2015; the black regression line shows that towns with high crime rates in 2015 tend to have lower crime rates in 2016, and vice versa for low crime rates. The grey dashed line shows what perfect correlation between 2015 and 2016 would look like.

Figure 3 shows the distribution of the pooled violent crime rates for 2016. The solid black line is a beta distribution fit to these data.

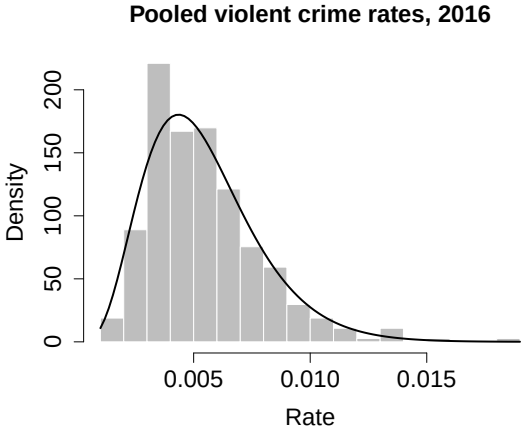

Figure 3: The distribution of violent crime rates in Norway, 2016. The black line describes the method-of-moments fit of a beta distribution to these data.

2.1. *Simulation study*

We run a simulation study for validation. If we assume that the crime probability in town i is stationary we can pool the observed crime rates of all years and use their average, $\bar{\theta}_i$, as a reasonable "truth." This allows us to assess the performance of our estimator against known, realistic crime probabilities, which of course is impossible in the real data. The simulated crime report count in town i is $k_i \sim \text{Binomial}(\bar{\theta}_i, n_i)$, where n_i is the 2016 population of town i. Figure 4 shows a realization of this procedure. Although not a perfect replica of Figure 1—the real data do not have any rates below .0017—it looks fairly realistic.

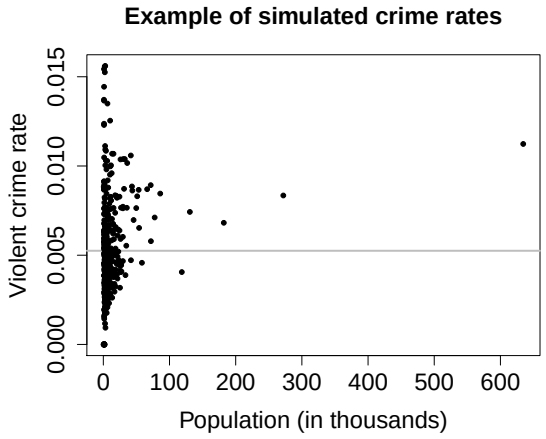

Figure 4: Funnel plot of a set of simulated crime rates

3. Methods

3.1. *Shrinkage estimates*

We treat θ_i as the probability for a person to commit a crime in a given period. We model the total number of crime reports in the i-th town, k_i, as the number of successful Bernoulli trials among n_i, where n_i is the population of this town. As explained in the introduction, this suggests the following simple Bayesian model, also shown in Figure 5:

$$\theta_i|\alpha,\beta \sim \text{Beta}(\alpha,\beta),$$
$$k_i|\theta_i \sim \text{Binomial}(n_i,\theta_i).$$

As mentioned the assumption of town exchangeability leads to this hierarchical model. This assumption might not be appropriate if we had reasons to think, for instance, that some regions are more prone to crime than others. In this case, region-specific priors might be better.

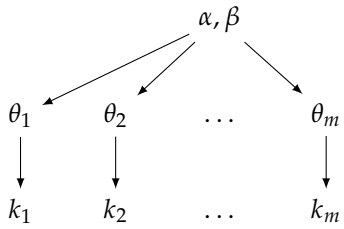

Figure 5: A graph describing our model. Crime counts, k_i, are (conditionally) i.i.d. binomials whose respective parameters, θ_i, are (conditionally) i.i.d. according to a common prior.

The posterior follows from the fact that the beta distribution is conjugate to itself with respect to the binomial likelihood. Generally, conjugacy means that the prior and posterior distributions belong to the same distributional family and usually entails that there is a simple closed-form way of computing the parameters of the posterior. Wasserman (2010, p. 178) shows a derivation of the posterior in the beta–binomial model:

$$\theta_i|k_i \sim \text{Beta}(\alpha + k_i, \beta + n_i - k_i).$$

We will look into the relation between the parameters of the posterior to those of the prior in terms of successes and failures in the results section.

The shrinkage estimate for the crime probability in town i is the posterior mean

$$\hat{\theta}_i^s = \frac{\alpha + k_i}{\alpha + \beta + n_i}.$$

The maximum likelihood estimate for θ_i is the observed crime rate $\hat{\theta}_i = {}^{k_i}/_{n_i}$. In order to fix values of α and β, we pool the MLEs for all towns $\hat{\theta}_1, \ldots, \hat{\theta}_m$ and fit a beta distribution to these data by the method of moments. We show the resulting fit in Figure 3. Because the expectation and variance of a Beta(α, β) are $\frac{\alpha}{\alpha+\beta}$ and $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, respectively, the parameter estimates for the prior are

$$\beta = \frac{\alpha(1-\bar{\theta})}{\bar{\theta}}, \text{ and}$$
$$\alpha = \left(\frac{1-\bar{\theta}}{S^2} - \frac{1}{\bar{\theta}}\right)\bar{\theta}^2.$$

Here $\bar{\theta} = \frac{\sum_i \hat{\theta}_i}{m}$ and $S^2 = \frac{\sum_i (\hat{\theta}_i - \bar{\theta})^2}{m-1}$ are the sample mean and variance of the pooled MLEs.

Instead of estimating α and β from the data like this, which ignores any randomness in these parameters, we could have a prior distribution for the parameters themselves. This would yield a typical Bayesian hierarchical model. Note also that in forming the estimate for town i, we end up using its information twice: once in eliciting our prior and once in the likelihood. This is convenient because we need only to find one prior rather than one for each town where we exclude the ith town from the ith prior. This bit of trickery does not make much difference: we have several hundreds of towns and hence removing a single town does not affect the shape of the prior much.

The estimate $\hat{\theta}_i^s = \frac{\alpha+k_i}{\alpha+\beta+n_i}$ shrinks the observed, or MLE, crime rate toward the prior mean $\bar{\theta}$. We can rewrite so that $\hat{\theta}_i^s = \delta_i\bar{\theta} + (1-\delta_i)\hat{\theta}_i$, with $\delta_i = \frac{\alpha+\beta}{\alpha+\beta+n_i}$. Here δ_i directly reflects the prior's influence on $\hat{\theta}_i^s$, and we see that this influence grows as the town size, n_i, shrinks.

3.2. *James-Stein estimates*

For completeness we demonstrate empirically that the James–Stein estimator is superior to the MLE in terms of risk. If town i has a large enough population, we can consider the normal approximation to the binomial distribution and assume

$$\hat{\theta}_i = \frac{k_i}{n_i} \sim \mathcal{N}\left(\theta_i, \sigma_i^2\right),$$

where $\sigma_i^2 = \frac{\theta_i(1-\theta_i)}{n_i}$ is unknown. If we assume that towns are similar in terms of variance we can consider the pooled variance estimate

$$\sigma_P^2 = \frac{\sum_{i=1}^m (n_i - 1)\hat{\sigma}_i^2}{\sum_{i=1}^m (n_i - 1)},$$

where $\hat{\sigma}_i^2 = \frac{\hat{\theta}_i(1-\hat{\theta}_i)}{n_i} = \frac{k_i(n_i-k_i)}{n_i^3}$. The James-Stein estimator of crime probability for town i is then

$$\hat{\theta}_i^{JS} = \left(1 - \frac{(m-2)\hat{\sigma}_P^2}{\sum_{i=1}^m \hat{\theta}_i^2}\right)\hat{\theta}_i.$$

This is a shrinkage toward zero. It assumes that crime rates are probably not as high as they appear. This is different from our assumption that crime rates are probably not as far away from the average as they appear. It is simple to modify the above to shrink toward any origin. The Efron-Morris variant (Efron and Morris, 1973) shrinks toward the average:

$$\hat{\theta}_i^{JS} = \bar{\theta} + \left(1 - \frac{(m-2)\hat{\sigma}_P^2}{\sum_{i=1}^m (\hat{\theta}_i - \bar{\theta})^2}\right)(\hat{\theta}_i - \bar{\theta}).$$

We will use this variant so that the two methods shrink toward the same point.

3.3. *Uncertainty intervals*

We construct credible intervals from the posterior. A 95% credible interval contains .95 of the posterior density, and the simplest way to construct one is to place it between the .025 and .975 quantiles of the posterior. For the MLE we use the typical normal approximation (or Wald) confidence interval. There is to our knowledge no straight-forward way to construct confidence intervals for the JS estimator, so we will leave this as an exercise for the reader.

3.4. Global risk estimates

We use the total squared-error loss function,

$$L(\theta, \hat{\theta}^s) = \sum_{i=1}^{m} (\theta_i - \hat{\theta}_i^s)^2,$$

to measure the global discrepancy between the true rates $\theta = (\theta_i)_{i=1,\dots,m}$ and estimates $\hat{\theta}^s = (\hat{\theta}_i^s)_{i=1,\dots,m}$. We do the same for the maximum likelihood and James-Stein estimates $\hat{\theta} = (\hat{\theta}_i)_{i=1,\dots,m}$ and $\hat{\theta}^{JS} = (\hat{\theta}_i^{JS})_{i=1,\dots,m}$, respectively.

We will compare the expected loss, or risk, of the three estimators $R(\cdot) = E[L(\cdot)]$, confirming the well-known property that shrinkage estimators dominate the MLE. We obtain Monte Carlo estimates of risk by averaging $L(\cdot)$ across repeated simulations.

3.5. Coverage properties

For the credible interval $C^s = (a, b)$, we want to assess the coverage probability $\mathbb{P}(\theta \in C^s)$ and compare with $\mathbb{P}(\theta \in C^W)$ for the classical Wald confidence interval. We will not assess the James–Stein estimator in terms of coverage. Let $I(C_i)$, where $C_i = C_i^s$ or C_i^W, be the indicator function that is equal to unity if $\theta_i \in C_i$, and zero otherwise. We obtain MC estimates of coverage probability by averaging the mean internal coverage, $\frac{1}{m}\sum_{i=1}^{m} I(C_i)$, across repeated simulations. An uncertainty interval should be well-calibrated: if the size of the interval is 95% it should trap the true parameter .95 of the time.

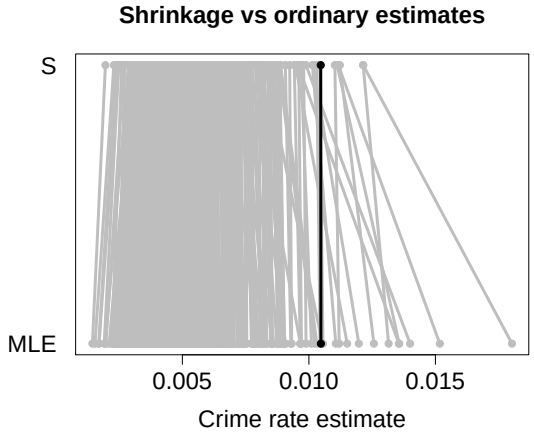

Figure 6: Comparing shrinkage and maximum likelihood estimates. Oslo, in black, is both close enough to the grand mean and large enough in size that the estimate does not change.

4. Results

4.1. Official SSB data

We focus on violent crimes in the year 2016. Figure 6 shows the effect of shrinking the observed crime rates toward the prior mean. We see that the more extreme estimates shrink toward the center. The city with highest crime rate according to the maximum likelihood estimate is Havsik ($\hat{\theta} = 0.018$), a small town with slightly more than 1000 inhabitants ($n = 1054$). After shrinkage, Havsik still ranks first, but the shrinkage estimate is much lower ($\hat{\theta}^s = 0.012$). Similarly the town with the lowest crime rate is Selbu ($\hat{\theta} = 0.0017$), another small town ($n = 4132$). Selbu's shrinkage estimate is higher than the MLE by more than 40% ($\hat{\theta}^s = 0.0024$). Oslo, shown in black, is a big city ($n = 658390$) and the difference between the two estimates is null ($\hat{\theta} - \hat{\theta}^s = 7 \times 10^{-6}$).

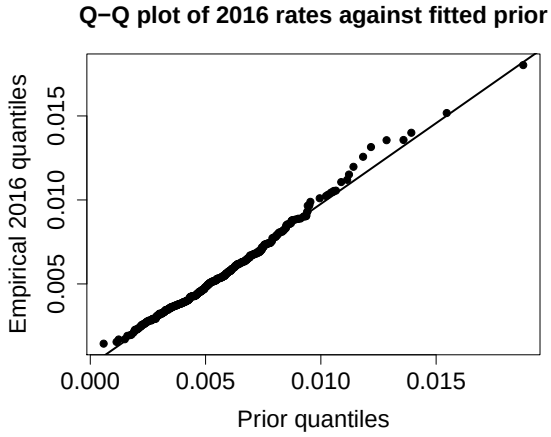

Figure 7: Quantile–quantile plot of 2016 crime rates against the fitted prior. The solid line describes a perfect fit.

Figure 7 is a quantile–quantile plot of the 2016 violent crime rates against the fitted prior. There is some very slight deviation around the tails, but overall it looks like a nice fit.

By shrinking toward the ensemble we add some information—we use the term informally—to the observed rate. We can quantify this by looking at the form of the beta distribution, so far taken for granted in this treatment. Its density function is

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)},$$

where the beta function in the denominator is simply the normalizing constant

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t.$$

A natural interpretation is that this is a distribution over the probability of success, i.e. crime, in a sequence of Bernoulli trials with $\alpha - 1$ successes and $\beta - 1$ failures (cf. the binomial distribution). Hence we can interpret the posterior for town i as a distribution over the probability of success in a series of Bernoulli trials with $\alpha' = \alpha + k_i$ successes and $\beta' = \beta + n_i$ failures (ignoring the -1 for convenience). In our data we have that $\alpha \approx 5$ and $\beta \approx 917$; it is as though we add the information of 922 extra trials in the binomial sense. In other words we add a priori 922 inhabitants, including five criminals, to each town.

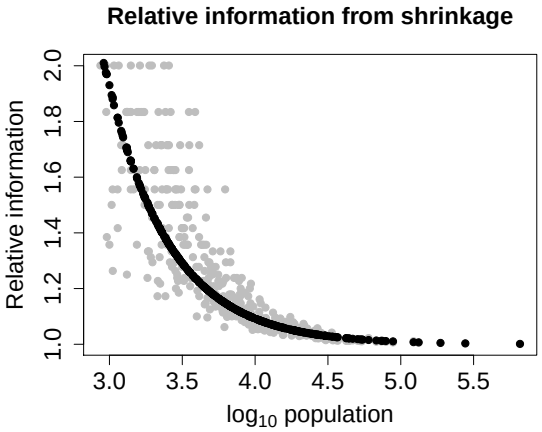

Figure 8: Relative information in the posterior mean compared to the MLE. The figure shows $(\alpha + k_i)/k_i$ in grey and $(\beta + n_i - k_i)/(n_i - k_i)$ in gray. These represent the added information in terms of number of successes and number of failures added to the MLE to form the shrinkage estimate. For the smallest towns, we practically double the information.

Figure 8 shows α' and β' (gray and black) relative to the number of successes (k_i) and failures (n_i - k_i) for each town in the 2016 data. For the smaller towns, there is double the information in the shrinkage estimate, while for larger towns there is no practical increase. Naturally the value of this extra information depends on the degree to which the prior is relevant.

Figure 9 shows the ten most violent towns according to shrinkage estimate along with their 95% credible intervals. The official, or MLE, crime rate is shown as a red point. We see some change in ordering. For Hasvik—a small and presumably quiet village in northern Norway—the MLE is so implausible that it is outside the credible interval. For Oslo—the biggest city in Norway—the estimate doesn't change.

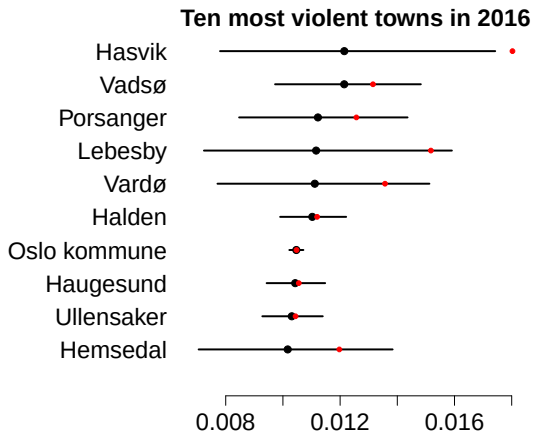

Figure 9: The ten towns with the highest crime rate, ordered by shrinkage estimate. The bars are 95% credible intervals. MLEs shown in red.

Figure 10 shows historical data for the three most violent and the three least violent towns in 2016, according to official crime rate. We show shrinkage estimates in red and official statistics in black. The vertical bars are 95% uncertainty intervals. The shrinkage estimate is usually more conservative, at least for the more violent towns, but the trends remain similar for both estimates. The credible intervals are shorter than the classical confidence intervals. We will see that in spite of this their coverage is better under simulation. It is interesting that the three most violent towns are all in Finnmark: Norway's largest and most sparsely populated county.

4.2. *Simulated data*

To obtain MC estimates of risks we run 100 000 simulations for each of our two experiments. Figure 11 shows kernel density estimates of the distributions of global loss. Our shrinkage estimates show lower global risk than maximum likelihood: $\hat{R}(\theta, \hat{\theta}^s) = 0.00054$ versus $\hat{R}(\theta, \hat{\theta}) = 0.00066$. The James–Stein estimates fall almost exactly between the two with $\hat{R}(\theta, \hat{\theta}^{JS}) = 0.00059$. We might have observed better results for JS had we used a variant of JS that allows unequal variances. Note that we fixed θ_i for this experiment, so we are only assessing the risk function in a single point.

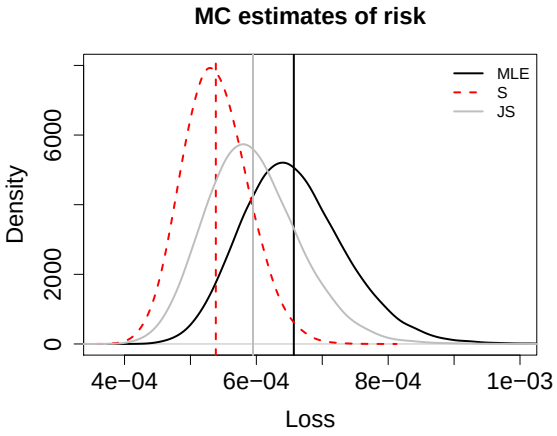

Figure 11: Distributions of $L(\theta, \hat{\theta})$ (solid black), $L(\theta, \hat{\theta}^s)$ (dashed red), and $L(\theta, \hat{\theta}^{JS})$ (solid grey). Vertical lines estimate the risk.

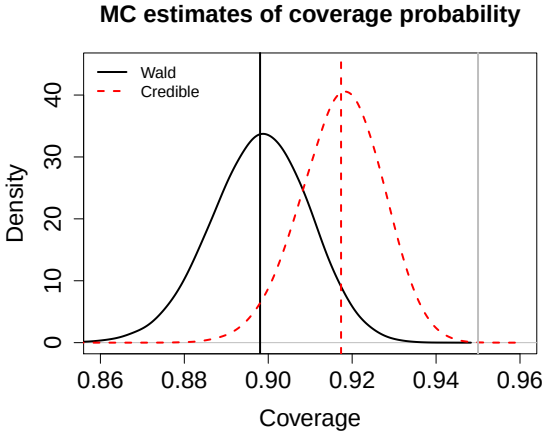

Figure 12: Distributions of the internal coverage $\frac{1}{m}\sum_{i=1}^{m} I(C_i^W)$ (solid black) and $\frac{1}{m}\sum_{i=1}^{m} I(C_i^s)$(dashed red). Vertical lines estimate coverage probability. The grey line shows the nominal coverage of .95.

Figure 12 presents estimated coverage probabilities in the same manner as Figure 11. The grey line shows the nominal coverage of .95. The coverage probability of the credible interval for

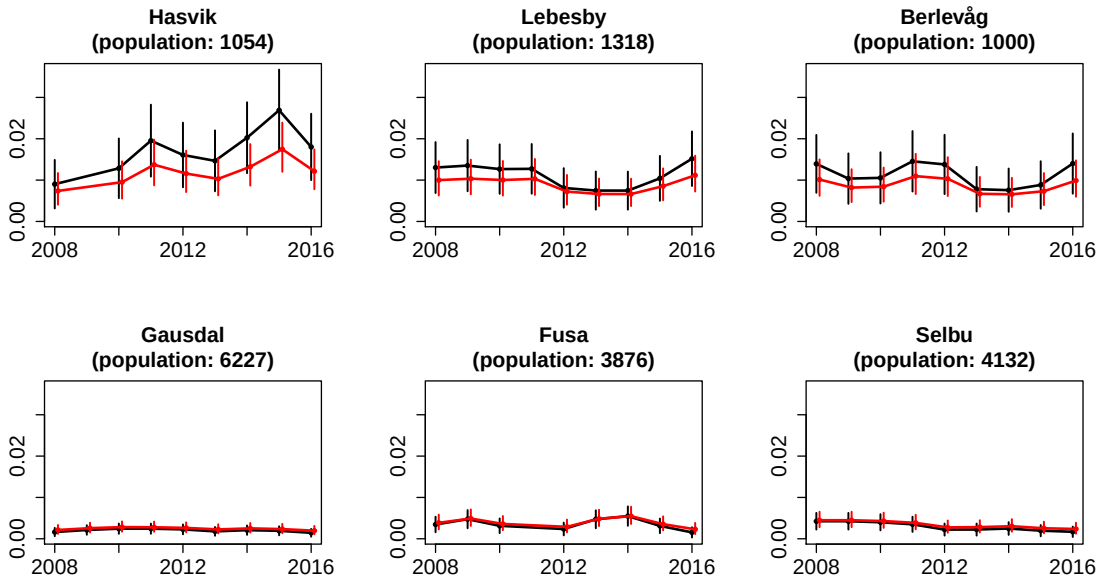

Figure 10: Historical data for the three most violent and the three least violent towns in 2016, ordered by official crime rate (MLE). The official statistics are drawn in black, and shrinkage estimates in red. The vertical bars indicate confidence and credible intervals, respectively.

the shrinkage estimator, $\hat{\mathbb{P}}(\theta \in C^s) = 0.917$, is closer to the nominal value than that of the the standard interval, $\hat{\mathbb{P}}(\theta \in C^W) = 0.898$. There is however still room for improvement.

5. Conclusion

This case study shows a simple method for simultaneous estimation of all town-specific crime rates in a country. The method is Bayesian in spirit, although we take some shortcuts with our prior. It is known that under squared-error loss the posterior mean is the optimal decision w.r.t. a given prior. In other words it minimizes Bayes risk, and is called the Bayes estimate. The theory gives us that Bayes estimates are admissible (Wald, 1947), and thus cannot be dominated. The risk estimates of our simulation agree with this. Our analysis provides an estimate of the crime probability with favorable frequency properties in terms of mean squared error and coverage.

Our simulations show that the Bayesian credible intervals from this treatment are narrower and have better coverage than the standard Wald confidence interval. Hence we get better information about the location of θ_i. Brown et al. (2001) show extensively that the Wald confidence interval for the binomial proportion behaves erratically for extreme values of p, for varying values of n, and for (un)lucky combinations of the two. Our result is interesting but quite narrow. Generalizing it requires more work.

Smaller towns are over-represented among the most and least violent towns in the official Norwegian data. Mathematically this has to be the case. Applying shrinkage methods to these data we get more conservative estimates for these variable and often extreme quantities. At the same time it seems that variance is not the only factor that places some of these small towns among the most violent. As Figure 9 shows, the top and bottom three in 2016 show a certain stability year by year. Hasvik in Finnmark has never ranked especially low

since 2008. Small towns in the north are often ranked high for violence. There could be many reasons for this and we leave further analysis to the criminologists.

These simple and useful estimation methods are best understood by practical examples. We encourage readers and students to actively follow this tutorial by playing with the available code and data. We used a single prior for all towns. It would be an interesting extension to use a mixture of beta distributions to account for any heterogeneity due to different latent rate levels. In this case, an EM algorithm could be used to assign each town to a class. Or, since Finnmark seems to be a special case, we might estimate per-county priors. It is also possible to include Bayesian multiple testing procedures to infer a list of cities likely to have true crimes rate above some given threshold. There is a temporal aspect to these data that we have not looked into. It would be possible to start out with a country-wide prior, but after this let the prior for one year be the posterior from the previous. Interested readers can find other ideas for further development in Robinson (2017). Gelman and Nolan (2017) also discuss a similar project to this one in their manual for statistics teachers.

In this treatment we have moved from descriptive figures typical of official statistics to model-based inferential statistics, estimating a crime probability rather than reporting a crime count. This allows us to account for variance and perhaps avoid over-interpreting noise, and hence avoid small-schools-type mistakes. We believe that probabilistic thinking can enrich descriptive statistics and aid in their interpretation.

Acknowledgements

We would like to thank our anonymous reviewer for the very thorough and very useful comments and suggestions.

References

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statist. Sci.*, 16(2):101–133.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors—an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

Gelman, A. and Nolan, D. (2017). *Teaching statistics: A bag of tricks*. Oxford University Press.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

Robinson, D. (2017). *Introduction to Empirical Bayes: Examples from Baseball Statistics*.

Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*, volume 13. John Wiley & Sons.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 197–206, Berkeley, Calif. University of California Press.

Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, 76(2):105.

Wainer, H. (2007). The most dangerous equation. *American Scientist*, 95(3):249.

Wald, A. (1947). An essentially complete class of admissible decision functions. *Ann. Math. Statist.*, 18(4):549–555.

Wasserman, L. (2010). *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated.

Correspondence: einar@cs.uit.no