



Shrinkage estimation for fun and profit

Einar Holsbø

February, 19th, 2018.



To appear in csbiggs.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.

Q: why do students in small schools perform better?

$$\text{Var}(\bar{x}) = \frac{\sigma}{n}$$

Q: how tall are the children of
tall parents?

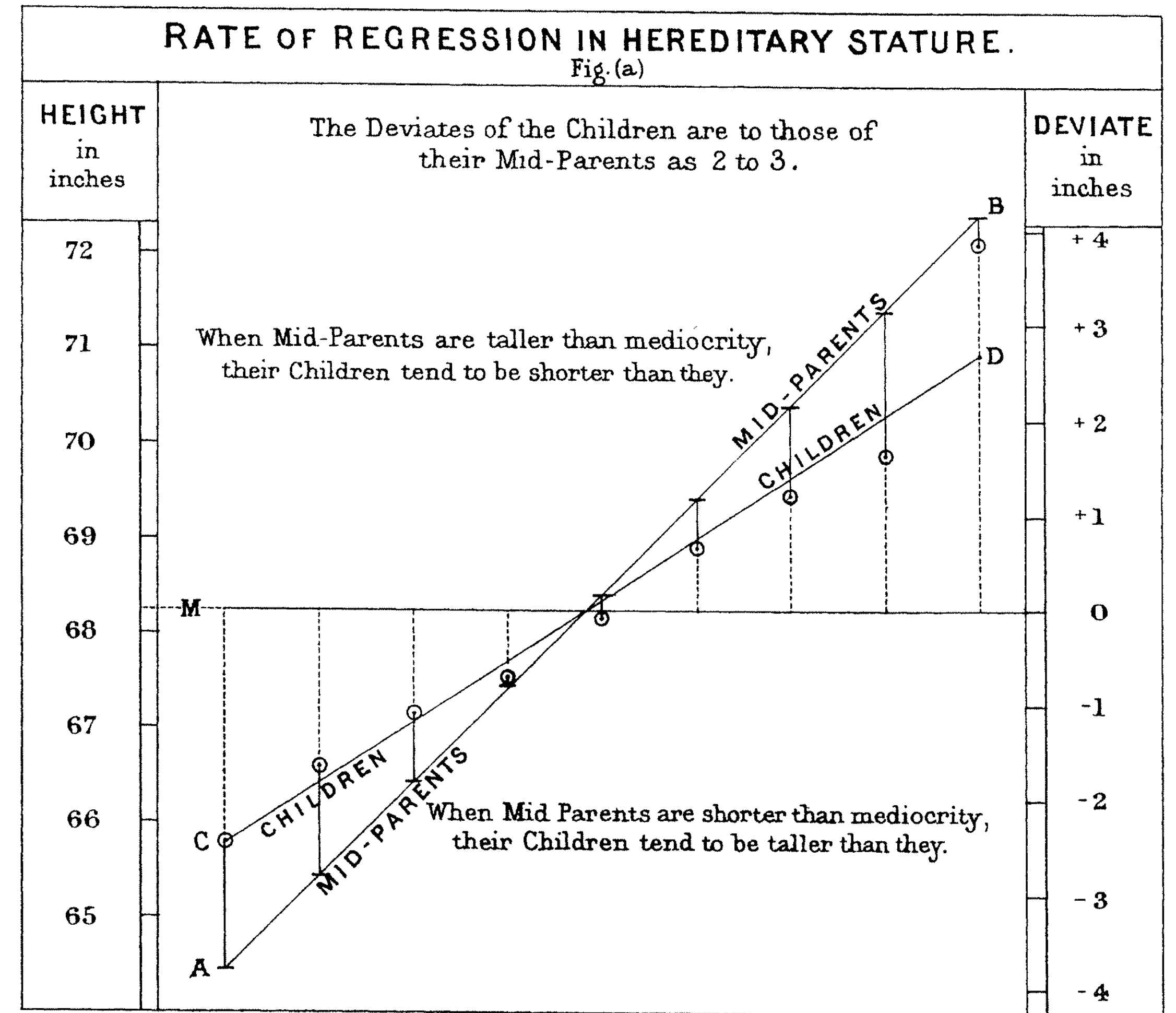
ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.

By FRANCIS GALTON, F.R.S., &c.

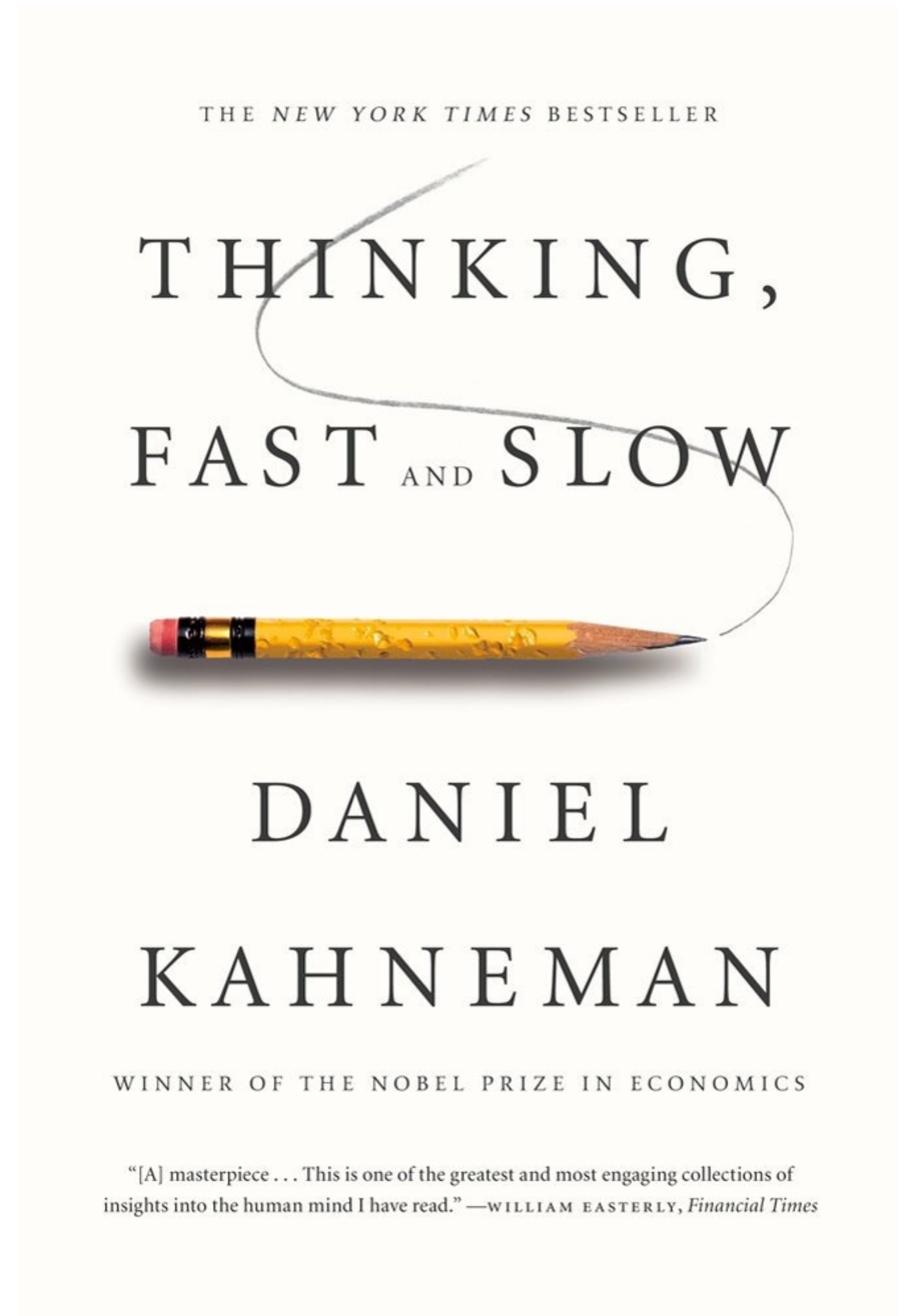
[WITH PLATES IX AND X.]

THIS memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.



‘On many occasions I have praised cadets... the next time they usually do worse. On the other hand, I have often screamed into a cadet’s earphone for bad execution, and in general he does better on his next try.’

<https://www.spectator.co.uk/2011/12/he-knew-he-was-wrong/>

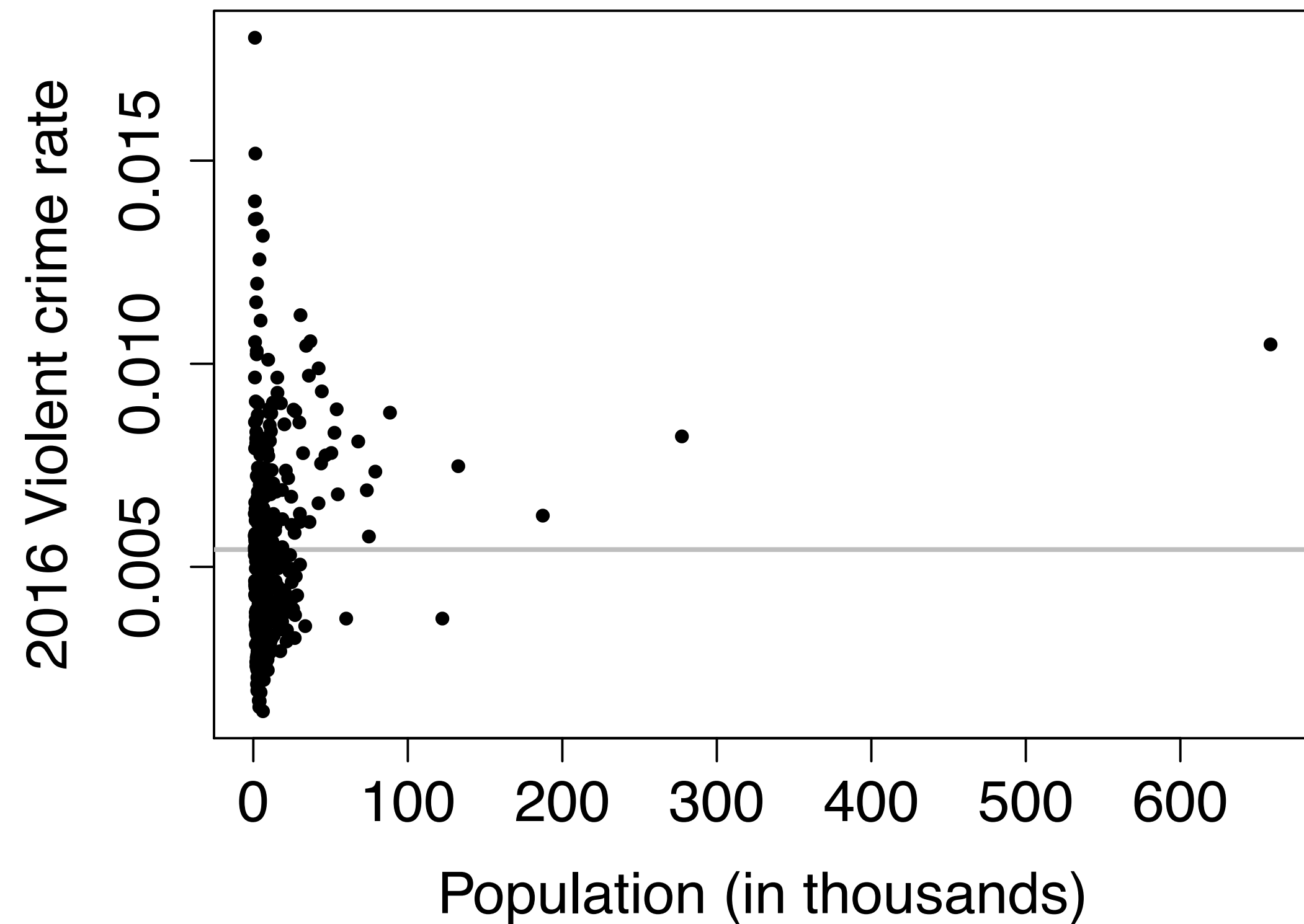


These phenomena are **everywhere.**

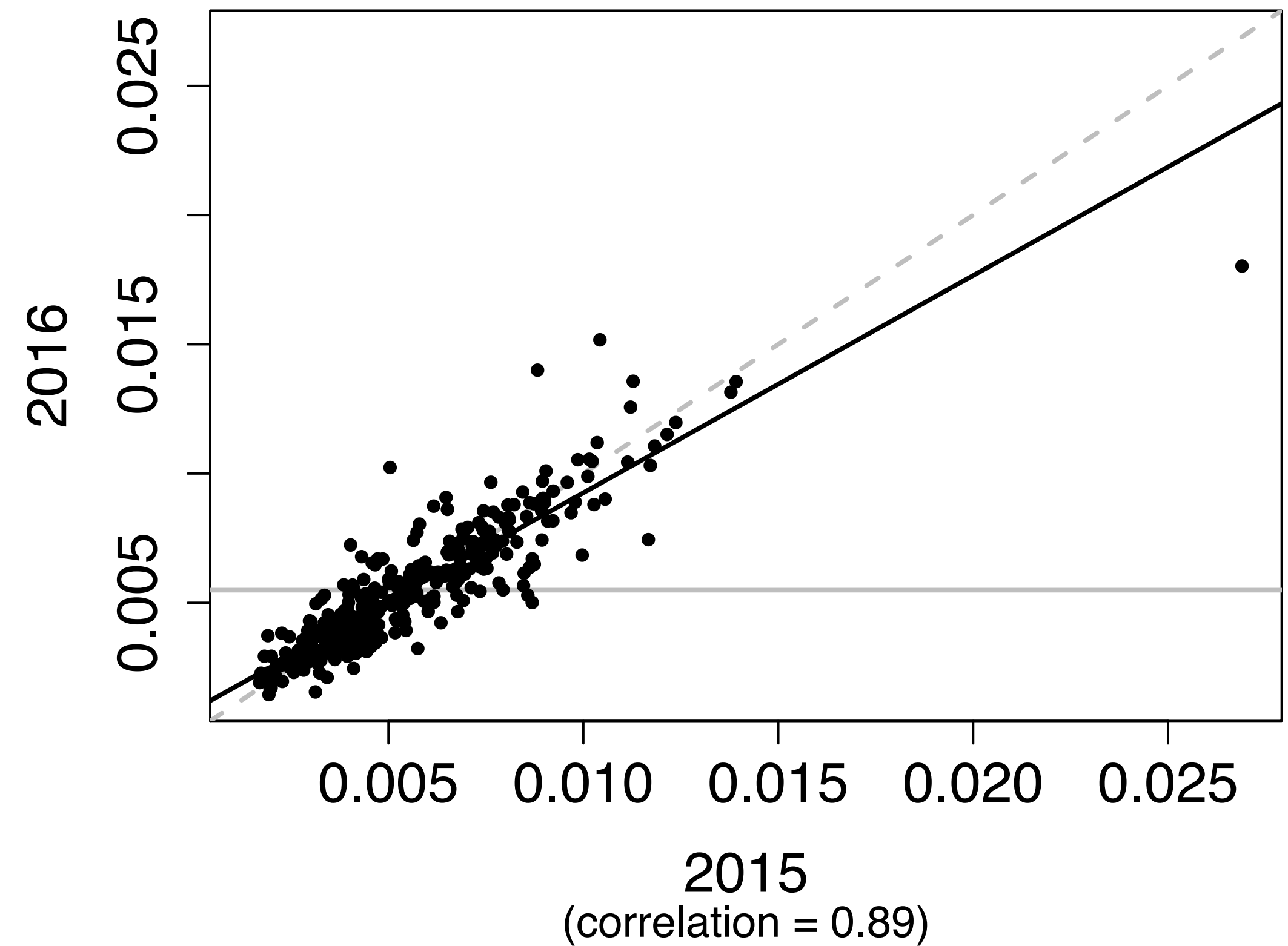
Beware unusual results.

Alive and well in official ssb.no data

Crime rates more variable for smaller towns



Crime rates regress to the mean





VOLDELIG: Ifølge Finnmark Dagblad er det i løpet av ett år i gjenn... [Vis mer](#)

Dette er Norges mest voldelige by

Her politianmeldes 80 voldshandlinger årlig.

Welcome to Vadsø Population: Einar

VOLDELIG: Ifølge Finnmark Dagblad er det i løpet av ett år i gjenn... [Vis mer](#)

Dette er Norges mest voldelige by

Her politianmeldes 80 voldshandlinger årlig.

Stein's Paradox in Statistics

The best guess about the future is usually obtained by computing the average of past events. Stein's paradox defines circumstances in which there are estimators better than the arithmetic average

by Bradley Efron and Carl Morris

Sometimes a mathematical result is strikingly contrary to generally held belief even though an obviously valid proof is given. Charles Stein of Stanford University discovered such a paradox in statistics in 1955. His result

was applied to Major League players as they were recorded after their first 45 times at bat in the 1970 season. These were all the players who happened to have batted exactly 45 times the day the data were tabulated. A batting average is defined, of course,

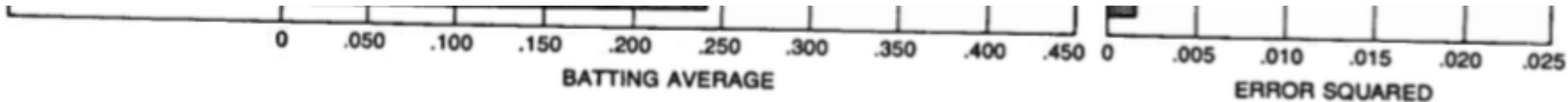
as the total number of hits divided by the total number of times at bat. The grand average factor c is .212. Substituting these values in the equation, we find that for each player z equals $.265 + .212(\bar{y} - .265)$. Because c is about .2, each average will shrink about 80 percent of the distance to the grand average, and the total




In estimating > 2 averages simultaneously, there is a better* estimator than the observed average.

* in terms of expected squared-error loss



■ ■ ■



-  INITIAL AVERAGE
-  SEASON AVERAGE
-  JAMES-STEIN ESTIMATOR

BATTING ABILITIES of 18 major-league baseball players are estimated more accurately by the method of Charles Stein and W. James than they are by the individual batting averages. The averages employed as estimators are those calculated after each player had had 45 times at bat in the 1970 season. The true batting ability of a player is an unobservable quantity, but it is closely approximated by his long-term average performance. Here the true ability is represented by the batting average maintained during the remainder of the 1970 season. For 16 of the players the initial average is inferior to another number, the James-Stein estimator, as a predictor of batting ability. The James-Stein estimators, considered as a group, also have the smaller total squared error.

Better results by **shrinking**
toward some common value

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

Better results by **shrinking**
toward some common value

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

All other averages used to adjust a single avg.

“Borrowing strength from the ensemble”

or

“Partial pooling”

Bayesian modelling

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

Bayesian modelling

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

Bayesian modelling

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

posterior \propto prior \times likelihood

Bayesian modelling

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

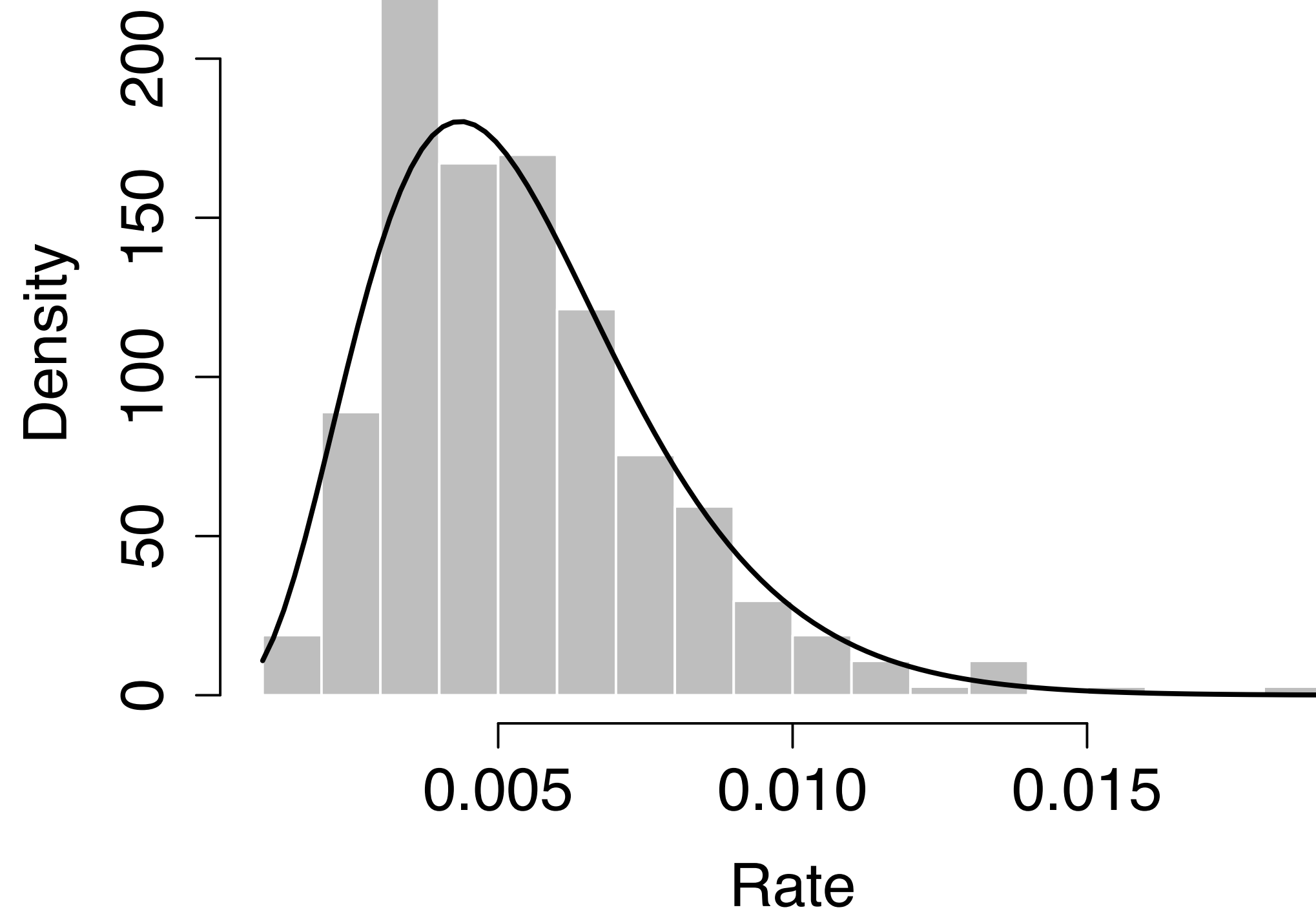
posterior \propto prior \times likelihood

Parameter of interest (θ) treated as random

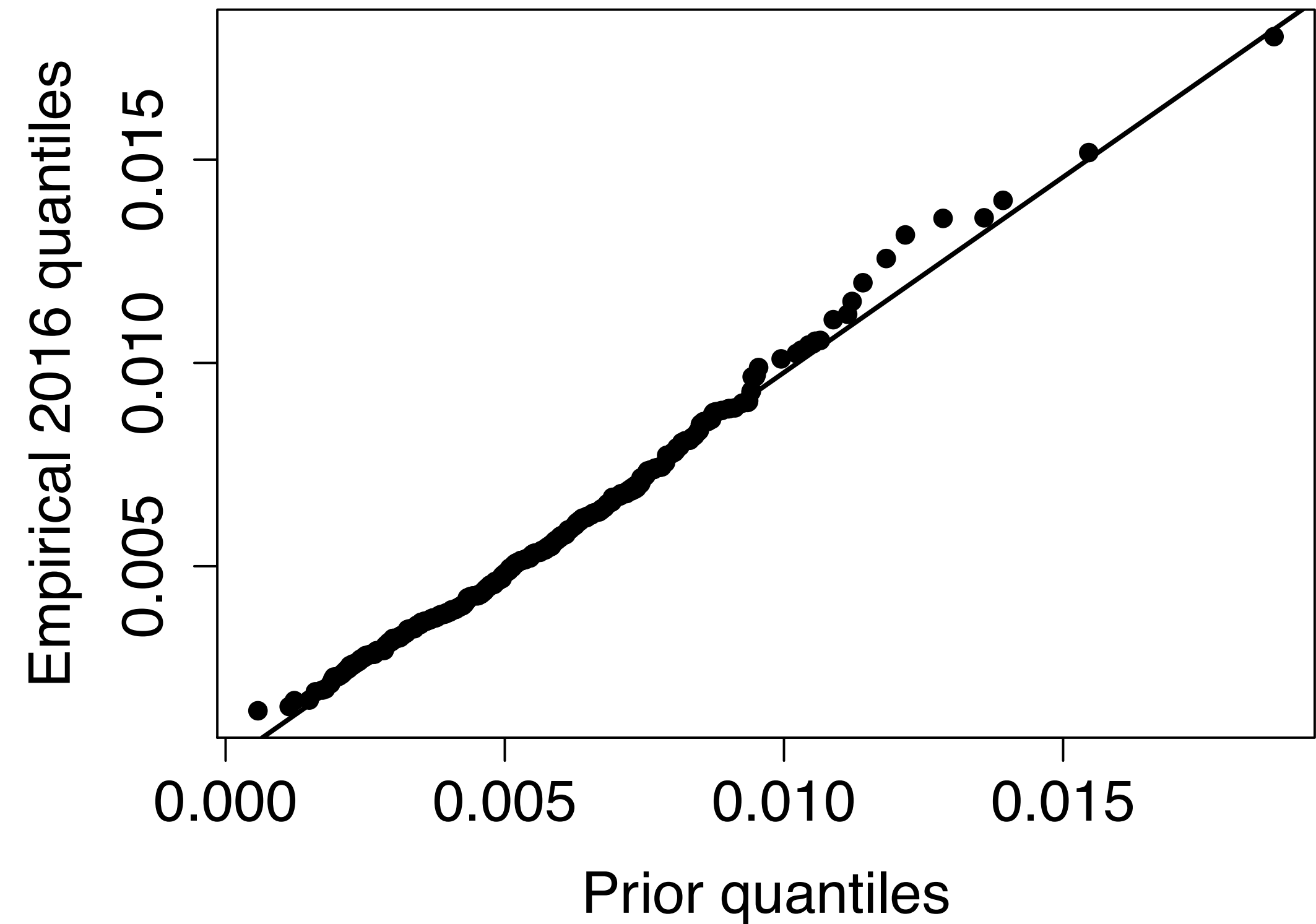
The prior represents our knowledge/uncertainty about the quantity of interest: **what would you say about crime rate if you knew nothing about a town?**

Assume rates are quite similar to one another

Pooled violent crime rates, 2016



Q-Q plot of 2016 rates against fitted prior



Likelihood: what is the evidence in the data?

**Binomial: number of successes (crimes)
in some number of trials (inhabitants)**

Posterior: Taking prior beliefs and evidence into account. **How should I adjust my prior belief about crime rates once I see the evidence?**

Hierarchical model for crime rates

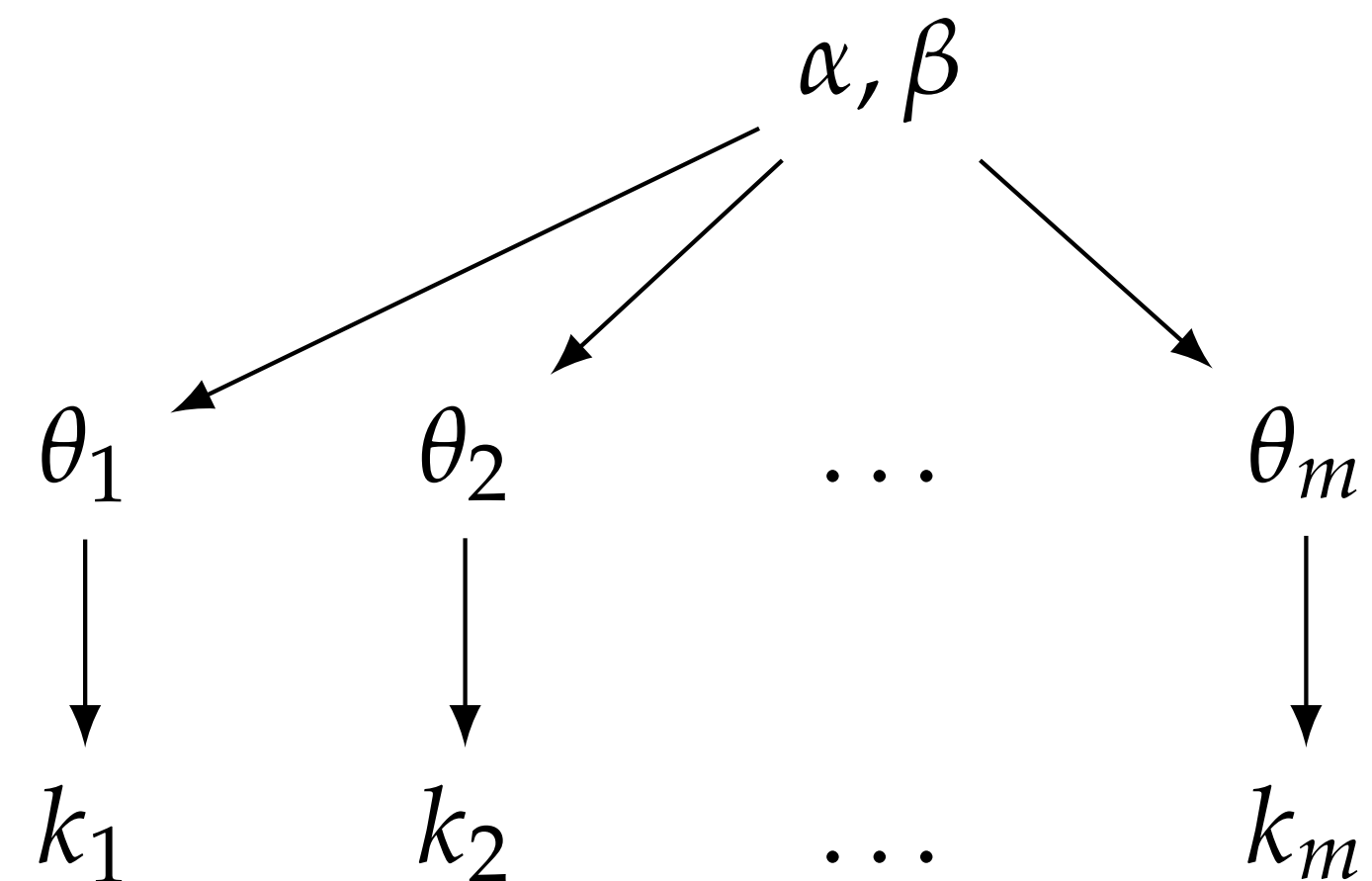
Crime rate = #crimes/#inhabitants

$$\theta = k/n$$

$$\theta_i | \alpha, \beta \sim \text{Beta}(\alpha, \beta),$$

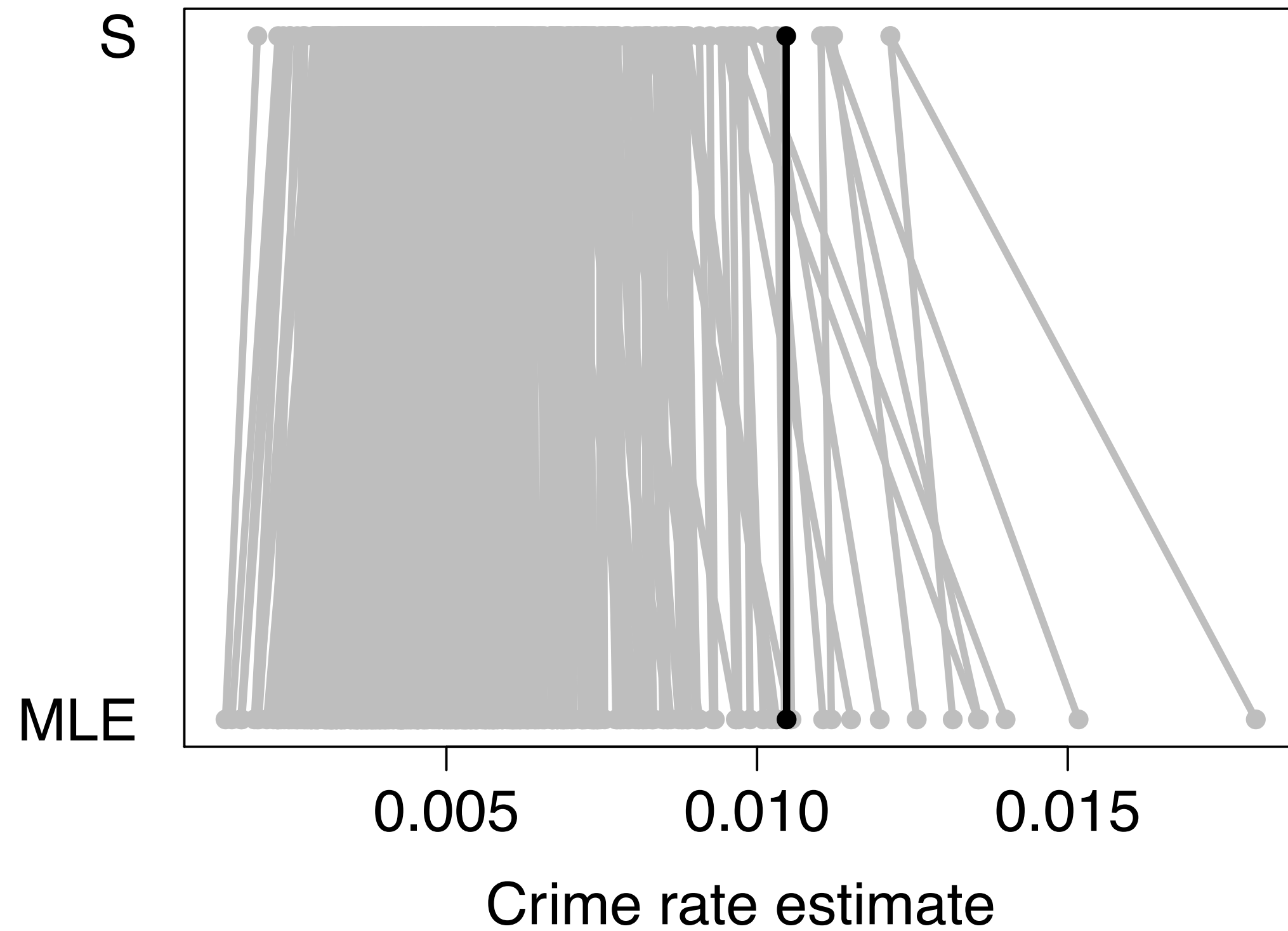
$$k_i | \theta_i \sim \text{Binomial}(n_i, \theta_i).$$

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

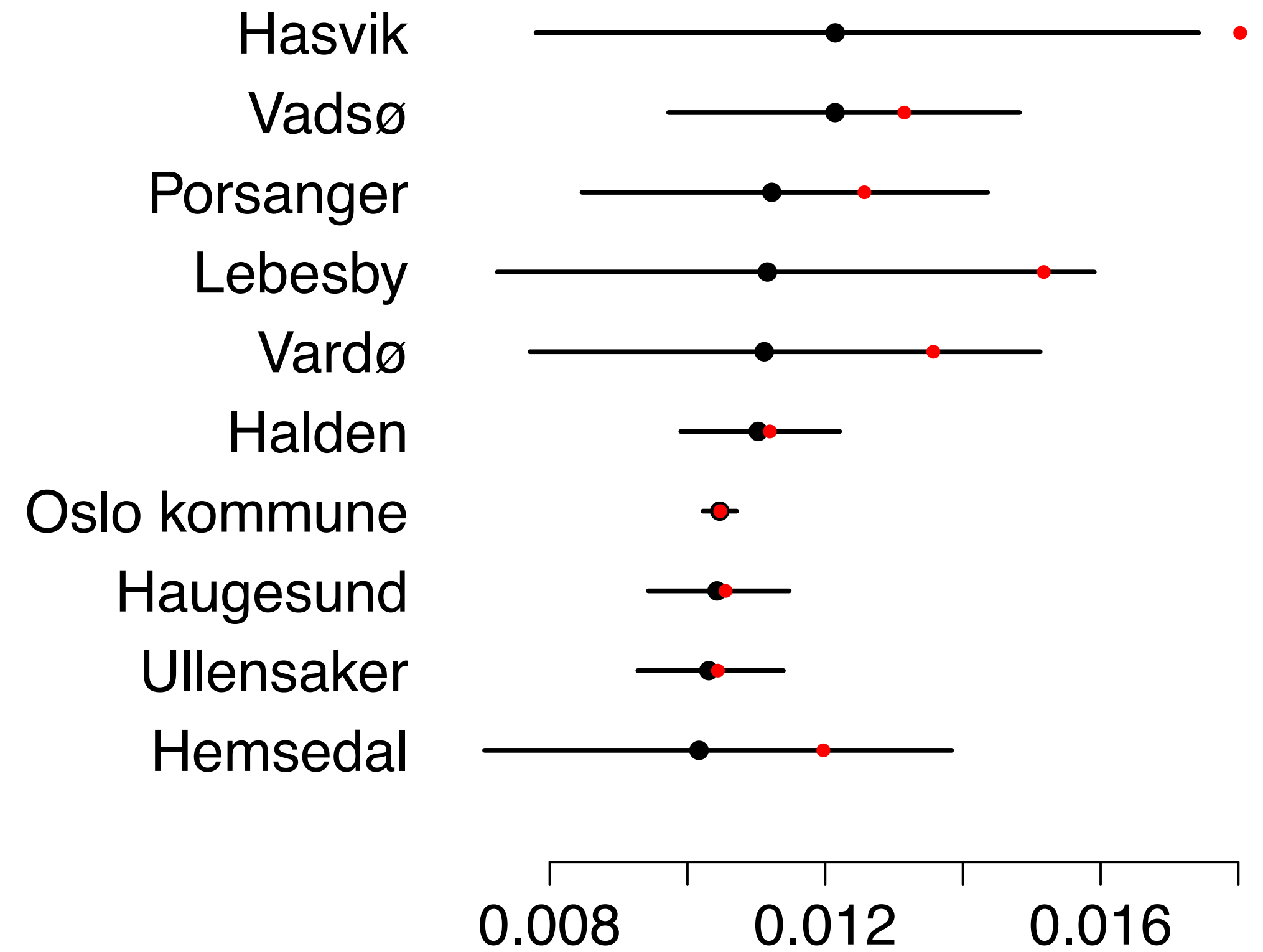


Posterior also beta: its mean gives us a **shrinkage** estimate

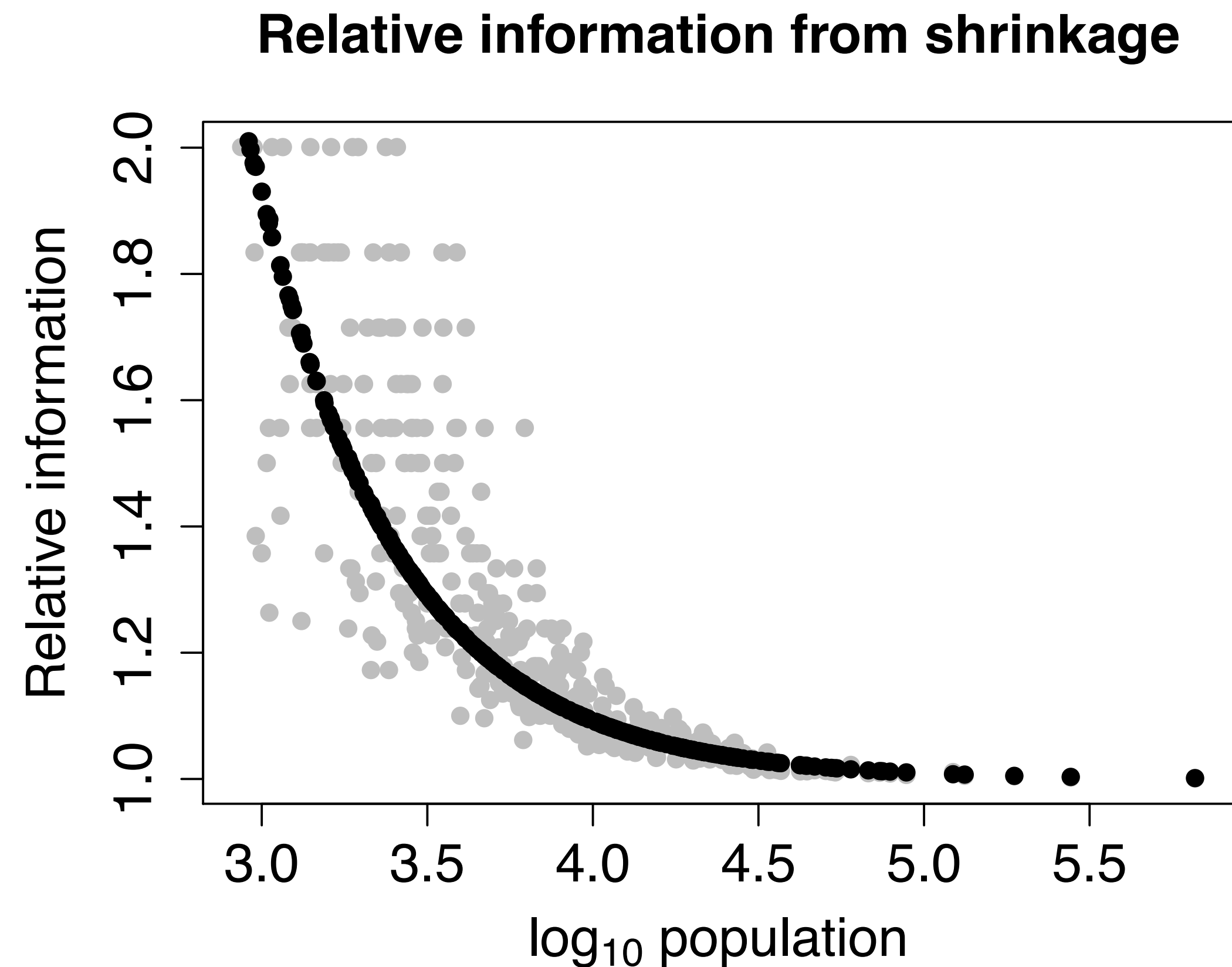
Shrinkage vs ordinary estimates



Ten most violent towns in 2016



Posterior also beta: its mean gives us a **shrinkage** estimate

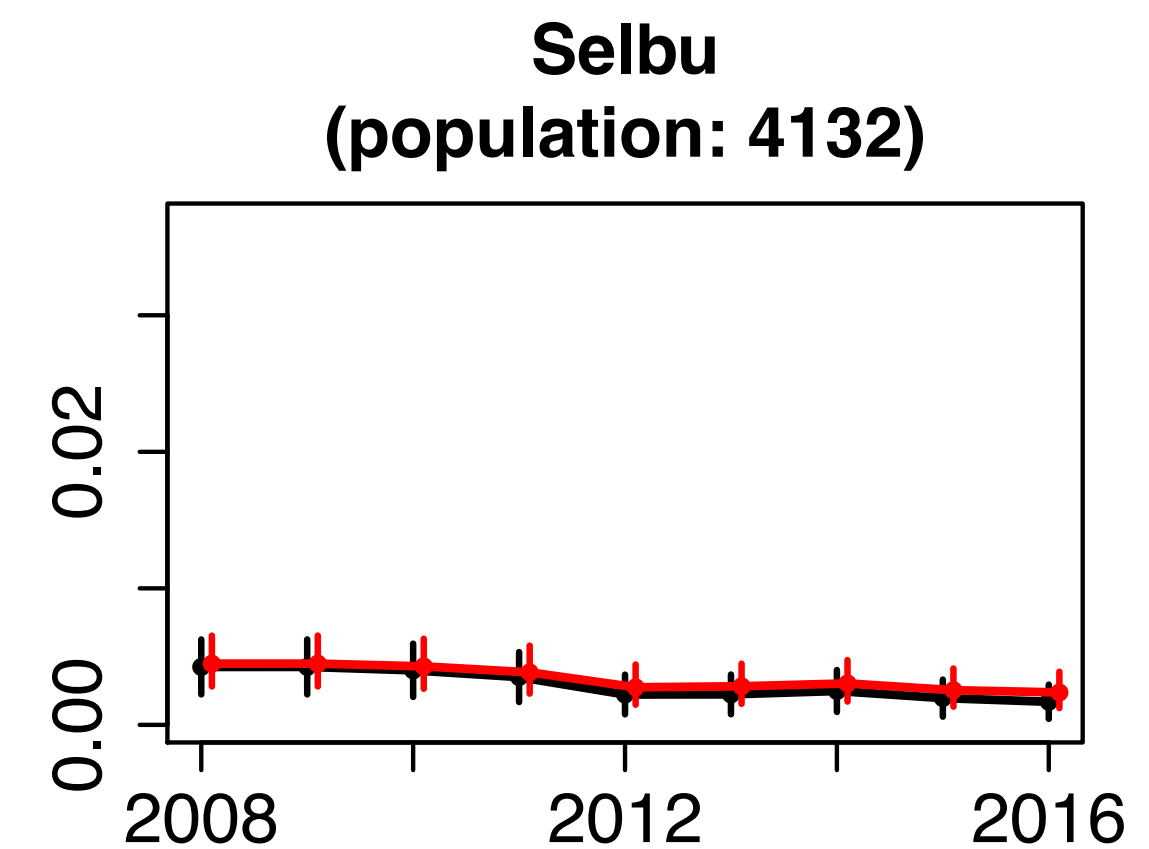
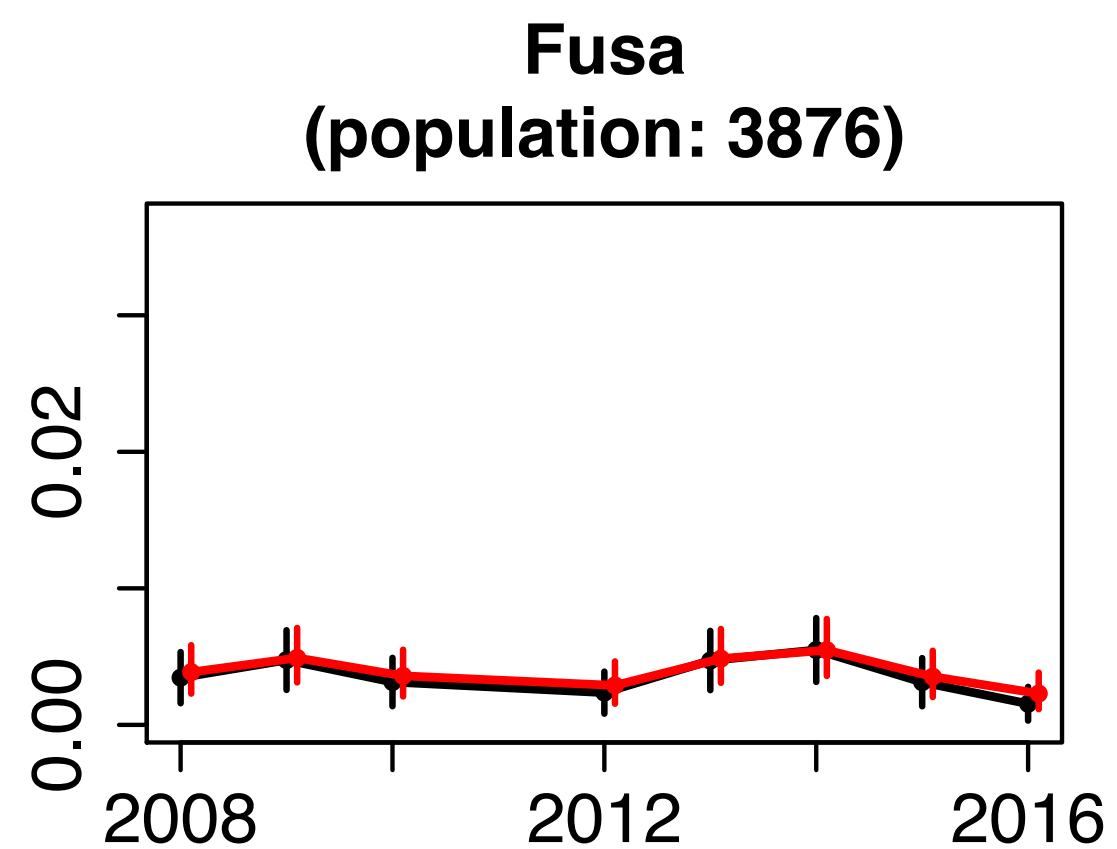
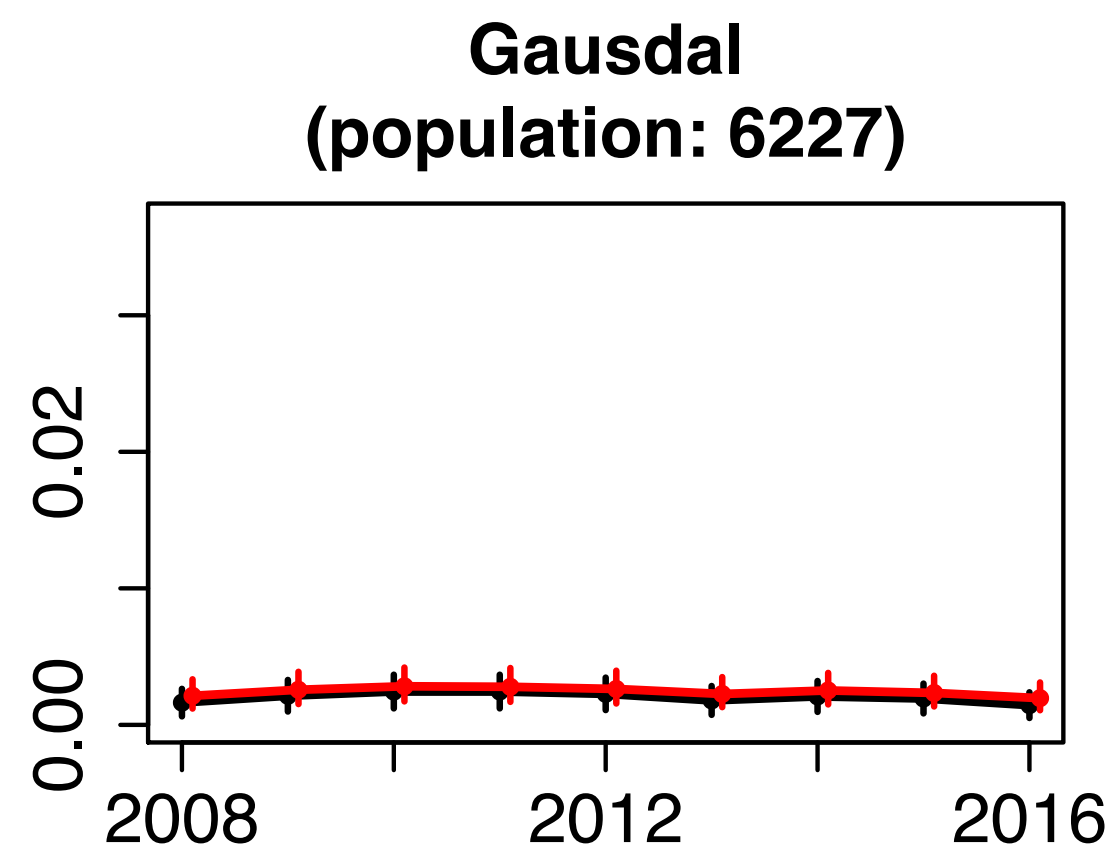
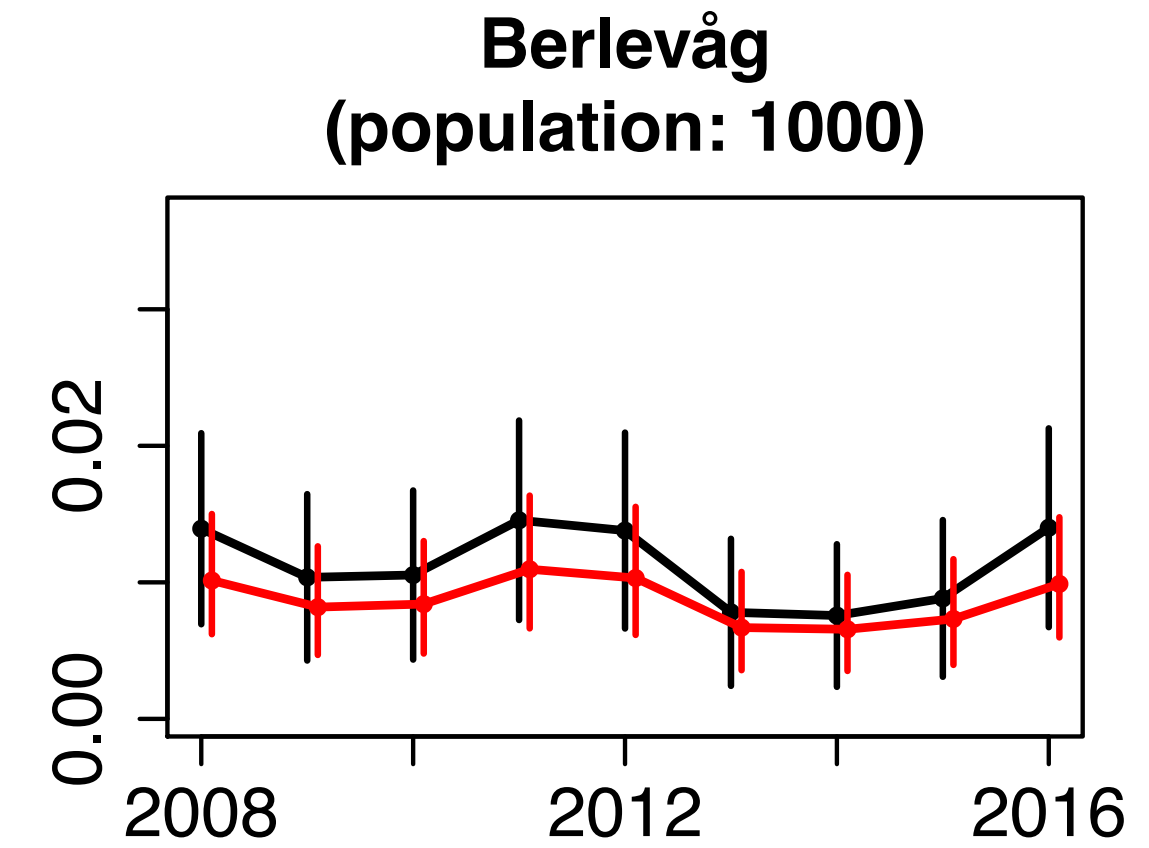
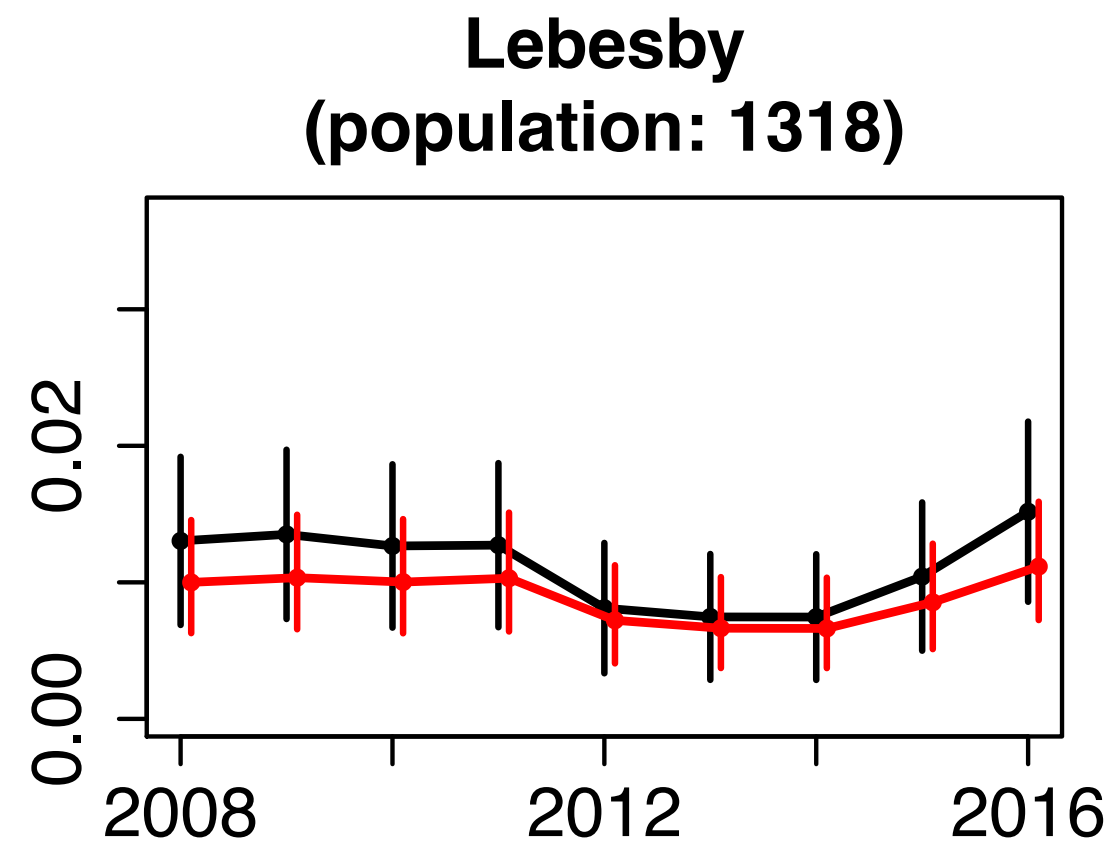
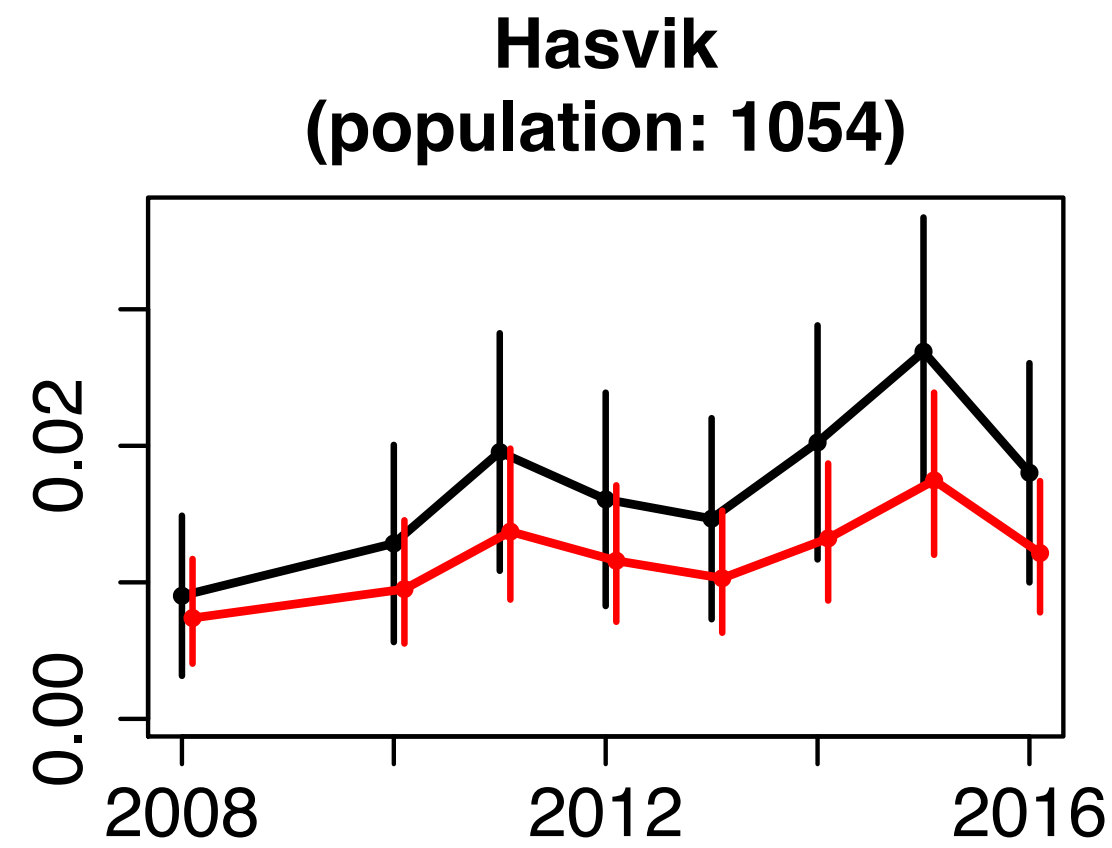


**Interpretation of the Beta(α , β):
distribution over probability when
observed $\alpha - 1$ successes, $\beta - 1$
failures.**

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

Our prior is equivalent to adding 922 inhabitants to each town, 5 of whom are criminals.

Posterior also beta: its mean gives us a **shrinkage** estimate



There aren't 80 violent crimes reported in Vadsø every year.

That happened **one year**.

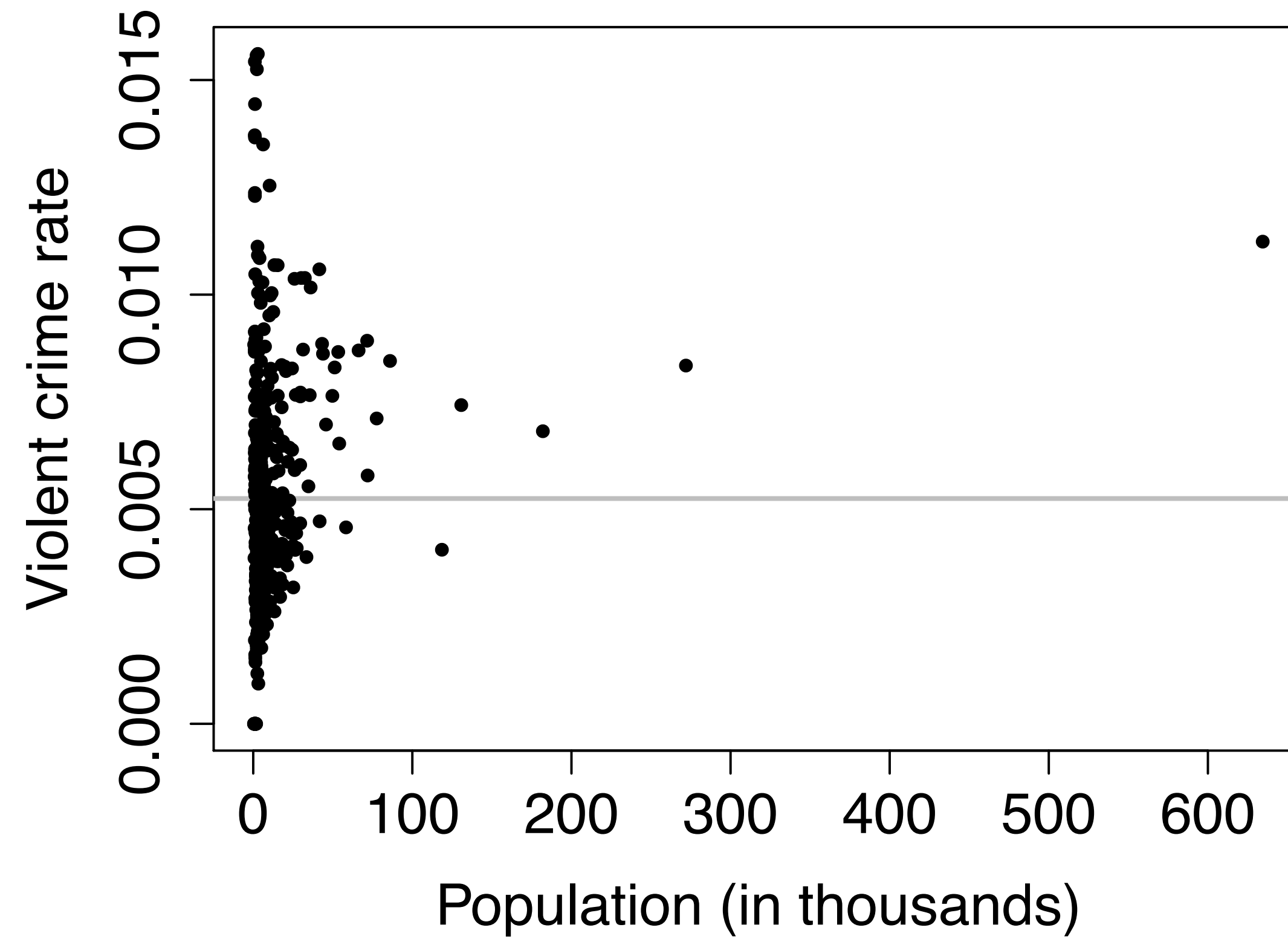
There aren't 80 violent crimes reported in Vadsø every year.

That happened **one year**.

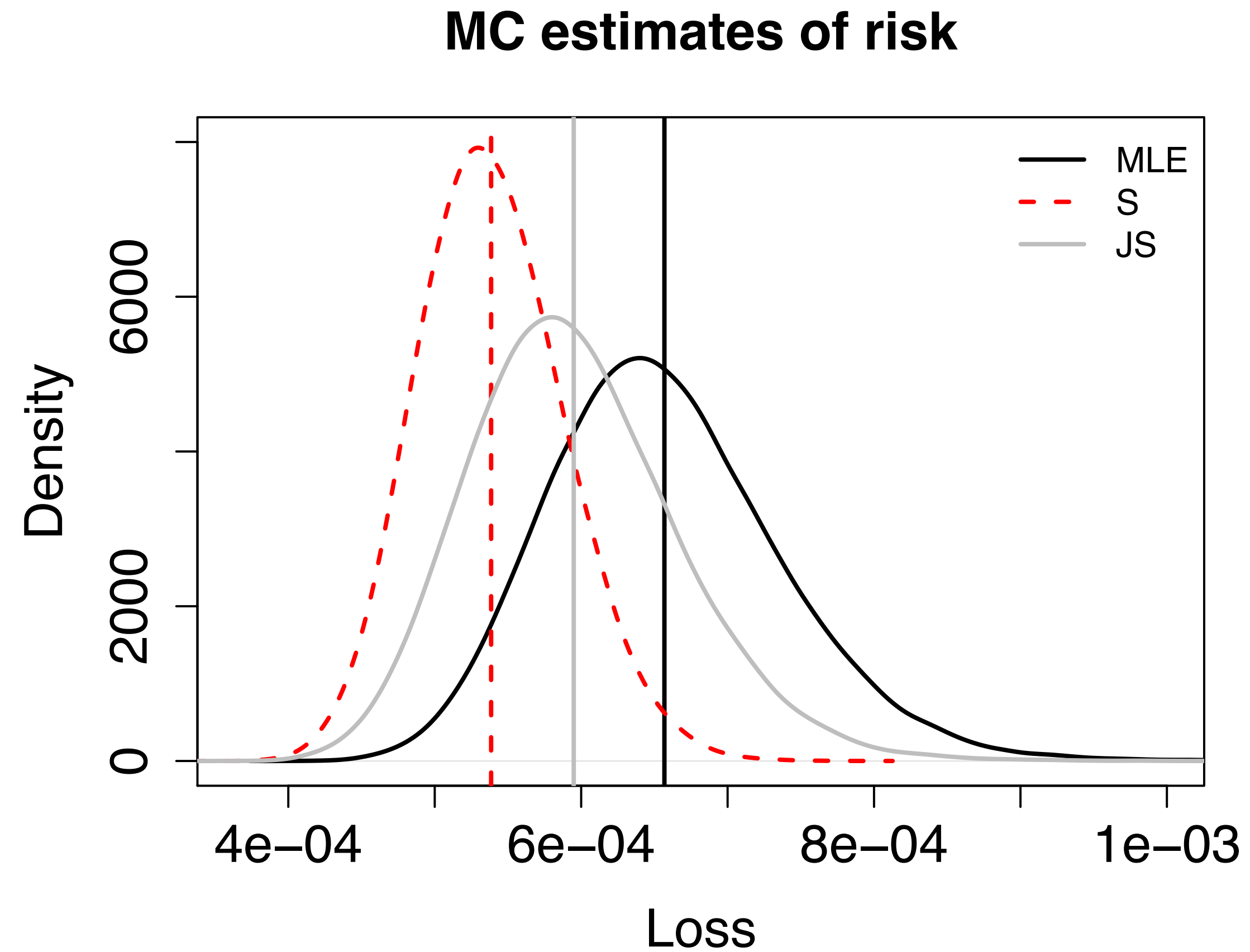
In fact there are **about 89 reports a year**, averaged across 2008—2016.

Some simulation results

Example of simulated crime rates

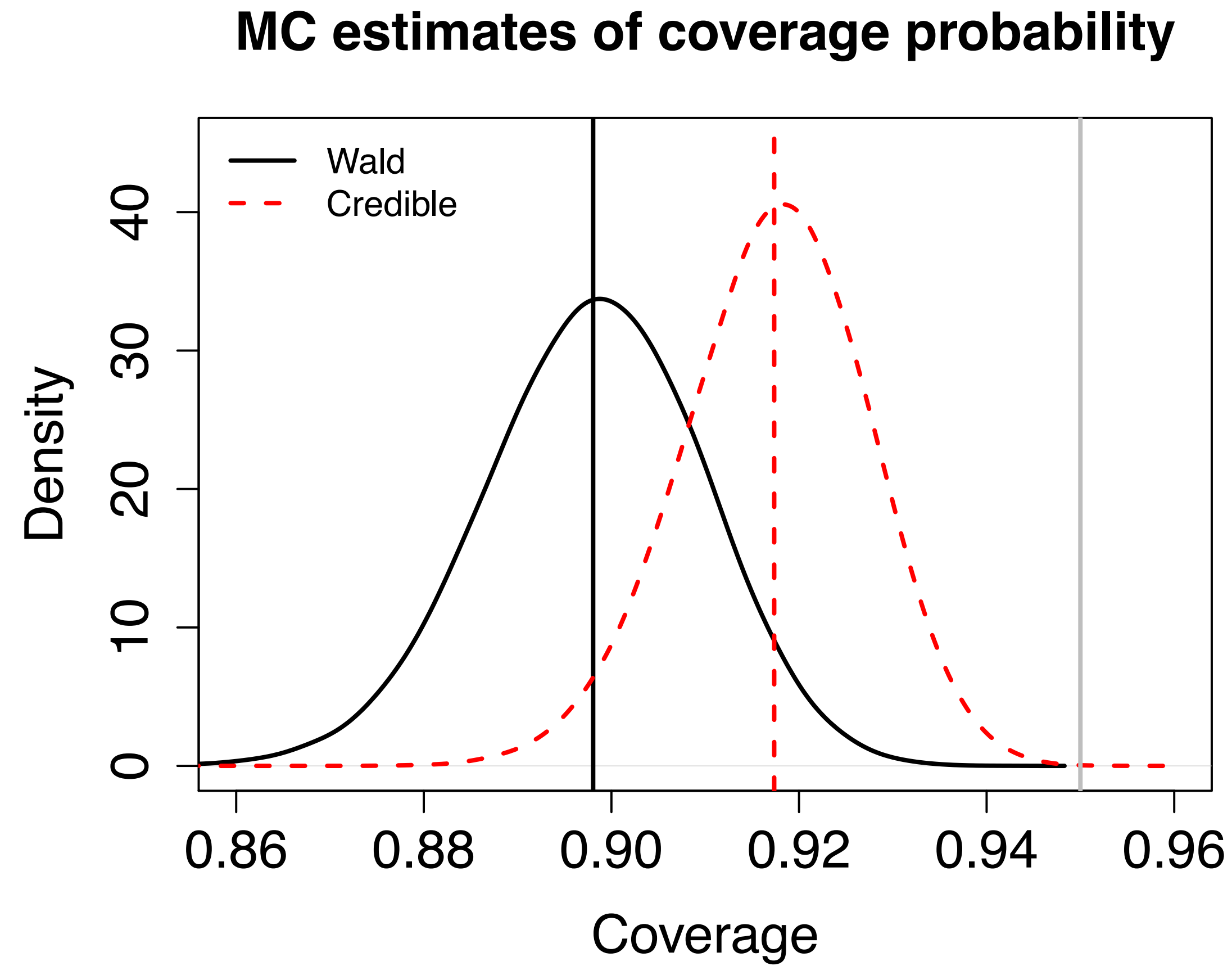


Some simulation results



Lower risk: risk is the expected loss/error in estimates

Some simulation results



Better coverage: a 95% interval should trap the truth 95% of the time.

Summary

- Don't get too excited about values far from the average
- Borrowing strength can reduce risk, improve interval calibration
- Fun (playing w data) can lead to profit (publication)



Thank you!

Slides available at 3inar.github.io

