UNIVERSITÀ DEGLI STUDI DI TRIESTE

Dipartimento di scienze economiche, aziendali, matematiche e statistiche "Bruno de Finetti"

Bayesian Statistics

Multiple parameter models

Francesco Pauli

A.A. 2019/20

Example: Kidney cancer death rates

For the 3130 counties of the US we know the death counts y_j in 1980-89 and populations n_j, death rates are computed as $r_j = y_j / n_j$.

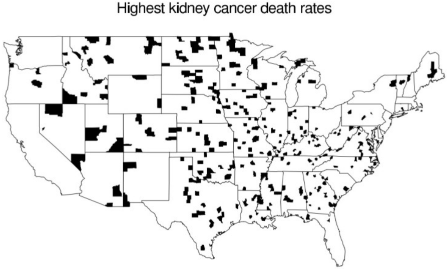


Figure 2.6 *The counties of the United States with the highest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989. Why are most of the shaded counties in the middle of the country? See Section 2.7 for discussion.*

Example: Kidney cancer death rates

For the 3130 counties of the US we know the death counts y_j in 1980-89 and populations n_j, death rates are computed as $r_j = y_j / n_j$.

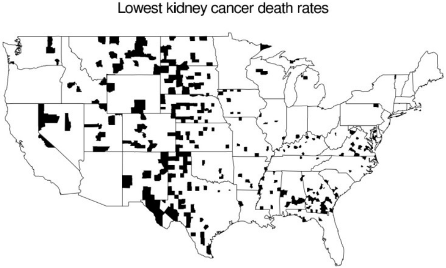


Figure 2.7 *The counties of the United States with the lowest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989. Surprisingly, the pattern is somewhat similar to the map of the highest rates, shown in Figure 2.6.*

Pattern on the map

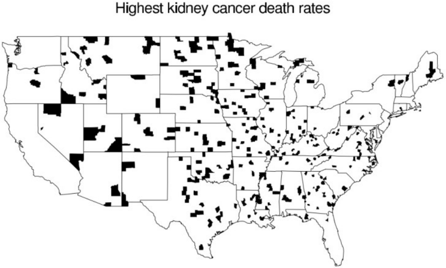


Figure 2.6 *The counties of the United States with the highest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989. Why are most of the shaded counties in the middle of the country? See Section 2.7 for discussion.*

Many of the counties in the Great Plains in the middle of the country, but relatively few counties near the coasts, are shaded.

Pattern on the map

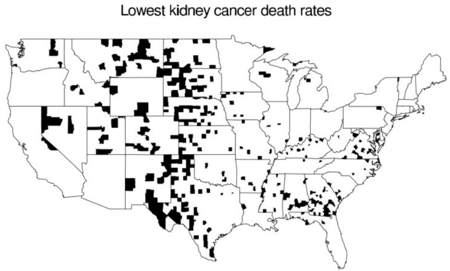


Figure 2.7 *The counties of the United States with the lowest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989. Surprisingly, the pattern is somewhat similar to the map of the highest rates, shown in Figure 2.6.*

The pattern of the counties with the lowest rate is similar, the point is that this pattern is due to the size of the counties.

Pattern on the map, explained

Kidney cancer is rare, the overall death rate is $4.65 \times 10^{?5}$.

Population of counties can be as low as 1000 up to 100s of thousand, in a scarsely populated county, let $n_j = 1000$ one has, in 10 years, $E(Y_j) = 0.4$ and if we assume a Poisson distribution for Y_j

- $P(Y_j = 0) = 0.63$
  - these will have a rate of 0 and so will be among the lowest
- $P(Y_j = 1) = 0.29$
  - these will have a rate of $1/10000$ and so will be among the highest

More populated countries are unlikely to have such extreme outcomes.

County Model

We assume

$$y_j \sim \text{Poisson}(10 n_j \theta_j)$$

we also assume

$$\theta_j \sim \text{Gamma}(20, 430\,000)$$

(implying $E(\theta_j) = 4.65 \times 10^{-5}$ and $\sqrt{V(\theta_j)} = 1.04 \times 10^{-5}$ so the posterior distribution is

$$\theta_j | y_j \sim \text{Gamma}(\alpha = 20 + y_j, \beta = 430\,000 + 10 n_j)$$

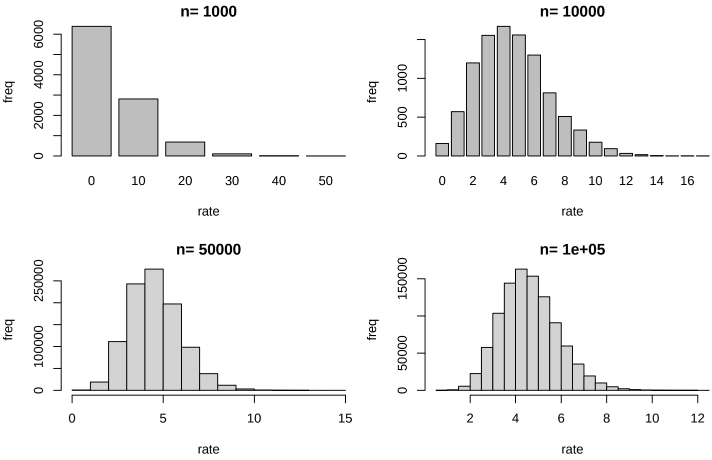
with mean and variance

$$E(\theta_j | y_j) = \frac{20 + y_j}{430\,000 + 10 n_j}; \quad \sqrt{V(\theta_j | y_j)} = \sqrt{\frac{20 + y_j}{(430\,000 + 10 n_j)^2}}$$

Results for some hypothetical countries

y	n	rate ($\times 10^{-5}$)	α	β	mean ($\times 10^{-5}$)	sd ($\times 10^{-5}$)
0	1000	0.00	20	440000	4.55	1.02
1	1000	10.00	21	440000	4.77	1.04
2	1000	20.00	22	440000	5.00	1.07
0	50000	0.00	20	930000	2.15	0.48
1	50000	0.20	21	930000	2.26	0.49
2	50000	0.40	22	930000	2.37	0.50
50	50000	10.00	70	930000	7.53	0.90
100	50000	20.00	120	930000	12.90	1.18
0	1000000	0.00	20	10430000	0.19	0.04
1	1000000	0.01	21	10430000	0.20	0.04
2	1000000	0.02	22	10430000	0.21	0.04
1000	1000000	10.00	1020	10430000	9.78	0.31
2000	1000000	20.00	2020	10430000	19.37	0.43

Predictive distribution of the observed rate



Predictive distribution of the observed rate

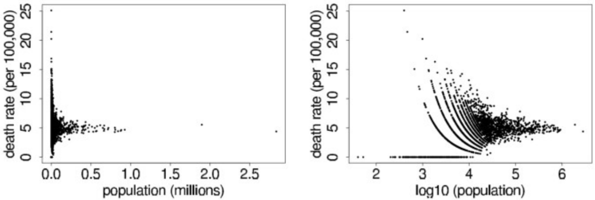


Figure 2.8 *(a) Kidney cancer death rates* $y_j/(10n_j)$ *vs. population size* n_j. *(b) Replotted on the scale of* $\log_{10}$ *population to see the data more clearly. The patterns come from the discreteness of the data* $(n_j = 0, 1, 2, \ldots)$.

Predictive distribution of the observed rate

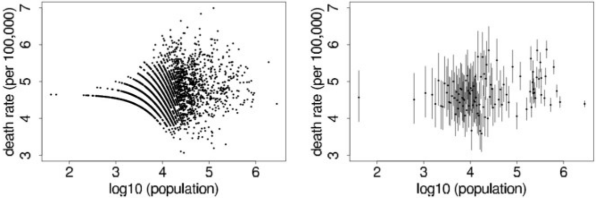


Figure 2.9 *(a) Bayes-estimated posterior mean kidney cancer death rates,* $E(\theta_j|y_j) = \frac{20+y_j}{430000+10n_j}$ *vs. logarithm of population size* n_j*, the 3071 counties in the U.S. (b) Posterior medians and 50% intervals for* θ_j *for a sample of 100 counties* j*. The scales on the y-axes differ from the plots in Figure 2.8b.*