Spatial Statistics for Epidemiologists

Part 2: Methods for Areal Data

Leonhard Held and Birgit Schrödle

27.-29. October 2010

Standardization

- Mortality and incidence rates depend on the age-sex composition of the population
- Use of standardization to produce comparable measures using a standard population
  - internal standard population – consists of age-specific rates from a superpopulation containing all study regions
  - external standard population – based on individuals from an entirely separate population
- Two methods of standardization:
  - direct - when the study population is large enough that age-specific rates are stable
  - indirect – when a population is small, the number of events are also small. Standardized mortality rates (SMRs) or standardized incidence rates (SIRs) are produced.

Direct standardization

Age standardized rate

$$\text{ASR} = \frac{\sum w_i \lambda_i}{\sum w_i}$$

where

- w_i are the person-years at risk in age group i in the standard population
- $\lambda_i = d_i/p_i$ is the rate in age group i in the study population in the time period of interest
- d_i is the observed number of deaths in age group i
- p_i is the person-years at risk in age group i in the study population

Indirect standardization: Some motivation

- y_{ij}: Number of cases
- n_{ij}: Number of individuals/person-years at risk

  in district $i = 1, \dots, I$ and age group $j = 1, \dots, J$

$$\boxed{\text{Assume } y_{ij} \sim \text{Binomial}(n_{ij}, \pi_{ij})}$$

- π_{ij}: Disease risk

Poisson approximation

$$\boxed{\text{Binomial}(n_{ij}, \pi_{ij}) \quad \approx \quad \text{Poisson}(n_{ij} \cdot \pi_{ij})}$$

Now assume proportionality: $\pi_{ij} = p_j \cdot \theta_i$

- p_j: Reference probability for age group j
- θ_i: Relative risk in district i

$\rightarrow$ Age-standardization: Suitable choice of p_j and aggregation over age group:

$$y_i = \sum_j y_{ij} \sim \text{Poisson}(\overbrace{\sum_j n_{ij} \cdot p_j}^{e_i} \cdot \theta_i)$$

- e_i: Expected number of cases in district i

Standardized mortality ratio

$$\boxed{y_i \sim \text{Poisson}(e_i \cdot \theta_i)}$$

ML-estimate of the relative risk θ_i: Standardized mortality ratio

$$\boxed{\text{SMR}_i = y_i / e_i}$$

Problems:

- Small number problem: $\text{Var}(\text{SMR}_i) = \frac{\theta_i}{e_i} \propto 1/e_i$
- The spatial structure in the data is ignored

German cancer mortality data, 1986-1990

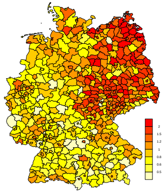
$I = 544$ districts

SMRs (internal standardization with $\sum_i y_i = \sum_i e_i$) for 7 different types of cancer:

Type	$\sum_i y_i$
Brain (males)	9,348
Cervix (females)	13,455
Larynx (males)	7,283
Lung (males)	134,820
Oesophagus (males)	12,835
Oral cavity (males)	15,466
Thyroid (males)	1,502

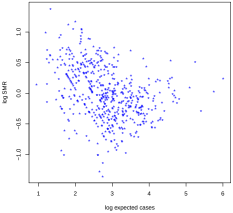
SMRs of cervix cancer

```
> summary(cervixSMRs)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2583  0.7386  0.9651  1.1000  1.3600  3.9720
> map.germany(cervixSMRs)
```



Variability of SMR estimates

- Variability of SMR estimates decreases with increasing number of expected cases



Discussion

1. Age-standardization is based on proportionality assumption
2. Statistical variation of SMR's depends on number of population at risk
3. SMR's totally ignore the spatial structure of the data
4. The same problems occur if analysis is based on rates rather than ratios

Overdispersion

$$\boxed{\text{Assume constant risk } \lambda \text{ and } y_i \sim \text{Poisson}(e_i\lambda)}$$

- $\rightarrow \hat{\lambda} = \sum y_i / \sum e_i$
- Calculate Poisson deviance D
- Compare with degrees of freedom (DF): $I - 1$
- Overdispersion estimate: $D/(I-1)$
- If constant risk assumption is correct then $D \sim \chi^2_{I-1} \rightarrow$ compute p-value
- Beware: formal test is based on asymptotics, i.e. relatively large numbers

Example: German cancer mortality data, 1986-1990

	type	deviance	df	overdispersion	pvalue
1	brain	651.4	543	1.2	0.001
2	cervix	1826.5	543	3.36	0.000
3	larnyx	790.7	543	1.46	0.000
4	lung	6100.2	543	11.23	0.000
5	oesophagus	1004	543	1.85	0.000
6	oral cavity	1264.3	543	2.33	0.000
7	thyroid	1292.6	543	2.38	0.000

But for another dataset from Sardinia …

	type	deviance	df	overdispersion	pvalue
8	IDDM	344.9	365	0.94	0.768

Discussion

1. What does lack of overdispersion indicate?
   - Not much evidence in the data for variation of disease risk
   - The variation of the SMR's can be explained as Poisson noise
   - Often due to (extremely) small numbers
   - → A spatial analysis will typically be sensitive to model/prior assumptions
2. What does overdispersion indicate?
   - There is evidence for extra-Poisson variation, which may be spatially structured or unstructured
   - There is need for further statistical analysis to filter out sample variation and possibly account for spatial correlation

The Clayton & Kaldor model

$\boxed{\text{Assume } y_i \sim \text{Poisson}(e_i\theta_i) \text{ with relative risk } \theta_i \sim \text{Gamma}(\alpha, \nu)}$

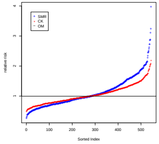
- Hyperparameters α and ν unknown
- Estimate α and ν by Empirical Bayes method: $\rightarrow \hat{\alpha}$, $\hat{\nu}$
- $\rightarrow$ Shrinkage estimate: $\theta_i | y \sim \text{Gamma}(y_i + \hat{\alpha}, e_i + \hat{\nu})$ with relative risk estimate

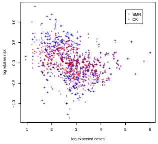
$$\hat{\theta}_i = \frac{y_i + \hat{\alpha}}{e_i + \hat{\nu}} = \frac{1}{e_i + \hat{\nu}} \left(e_i \cdot \frac{y_i}{e_i} + \hat{\nu} \cdot \frac{\hat{\alpha}}{\hat{\nu}} \right)$$

- $\hat{\theta}_i$ is weighted average of SMR y_i/e_i and "prior" mean $\hat{\alpha}/\hat{\nu}$

Shrinkage

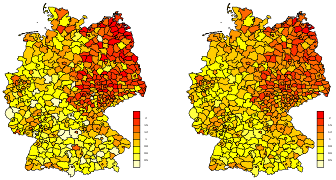
- Clayton & Kaldor estimates are shrinked towards the overall mean $\hat{\alpha}/\hat{\nu}$
- Estimates based on large numbers of observed cases are close to the corresponding SMR; those based on small numbers are close to the overall risk





Maps of SMR and CK estimates

- Cancer type: Cervix



Example: German cancer mortality data, 1986-1990

- For assessment of overall shrinkage compare coefficient of variation (CV) of SMRs with CV of estimated Gamma distribution ($1/\sqrt{\hat{\nu}}$):

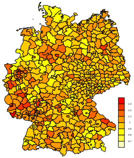
	type	cvsmr	cvck	shrinkage
1	brain	0.35	0.10	71 %
2	cervix	0.46	0.37	19 %
3	larnyx	0.42	0.18	57 %
4	lung	0.21	0.19	8 %
5	oesophagus	0.37	0.19	48 %
6	oral cavity	0.36	0.23	36 %
7	thyroid	0.83	0.56	32 %

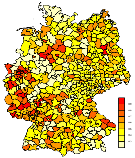
But for Sardinia ...

	type	cvsmr	cvck	shrinkage
8	IDDM	1.47	0.15	90 %

"Bayesian p-values"

- Posterior probability that the relative risk θ_i in area i is above the overall risk $\sum y_i / \sum e_i$ (here: 1) or any other reference value
- Example larynx cancer: Plots show CK estimate for each region and the respective posterior probability





Discussion

The CK model for estimating disease risk is an improvement over SMRs because

1. the method filters out "Poisson noise" and hence visualizes extra-Poisson variation
2. artefacts in SMR maps due to heterogeneity in population size are deleted

However, the spatial structure is ignored