Biostatistics

Applications in Serological Data Analysis

Nuno Sepúlveda, 19.01.2026

Syllabus

1. General review
    a. Population/Sample/Sample size
    b. Type of Data – quantitative and qualitative variables
    c. Common probability distributions/popular tests
2. Applications in Medicine
    a. Construction and analysis of diagnostic tools – Binomial distribution, ROC curve, sensitivity, specificity, Rogal-Gladen estimator
    b. Estimation of treatment effects - generalized linear models
    c. Survival analysis - Kaplan-Meier curve, log-rank test, Cox's proportional hazards model
3. Applications in Genetic and Epigenetic Data
    a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, multiple testing correction
    b. Methylation association studies – M versus beta values
4. Applications in Serological Data Analysis
    a. Determination of seropositivity using Gaussian mixture models
    b. Reversible catalytic models for estimating seroconversion rate
    c. Sample size calculation for estimating seroconversion rate

Exercise: data_serology.csv

Multiplex assays for the identification of serological signatures of SARS-CoV-2 infection: an antibody-based diagnostic and machine learning study

Jason Rosado, Stéphane Pelleau, Charlotte Cockram, Sarah Hélène Merkling, Narimane Nekkab, Caroline Demeret, Annalisa Meola, Solen Kerneis, Benjamin Terrier, Samira Fafi-Kremer, Jerome de Seze, Timothée Bruel, François Dejardin, Stéphane Petres, Rhea Longley, Arnaud Fontanet, Marija Backovic, Ivo Mueller, Michael T White

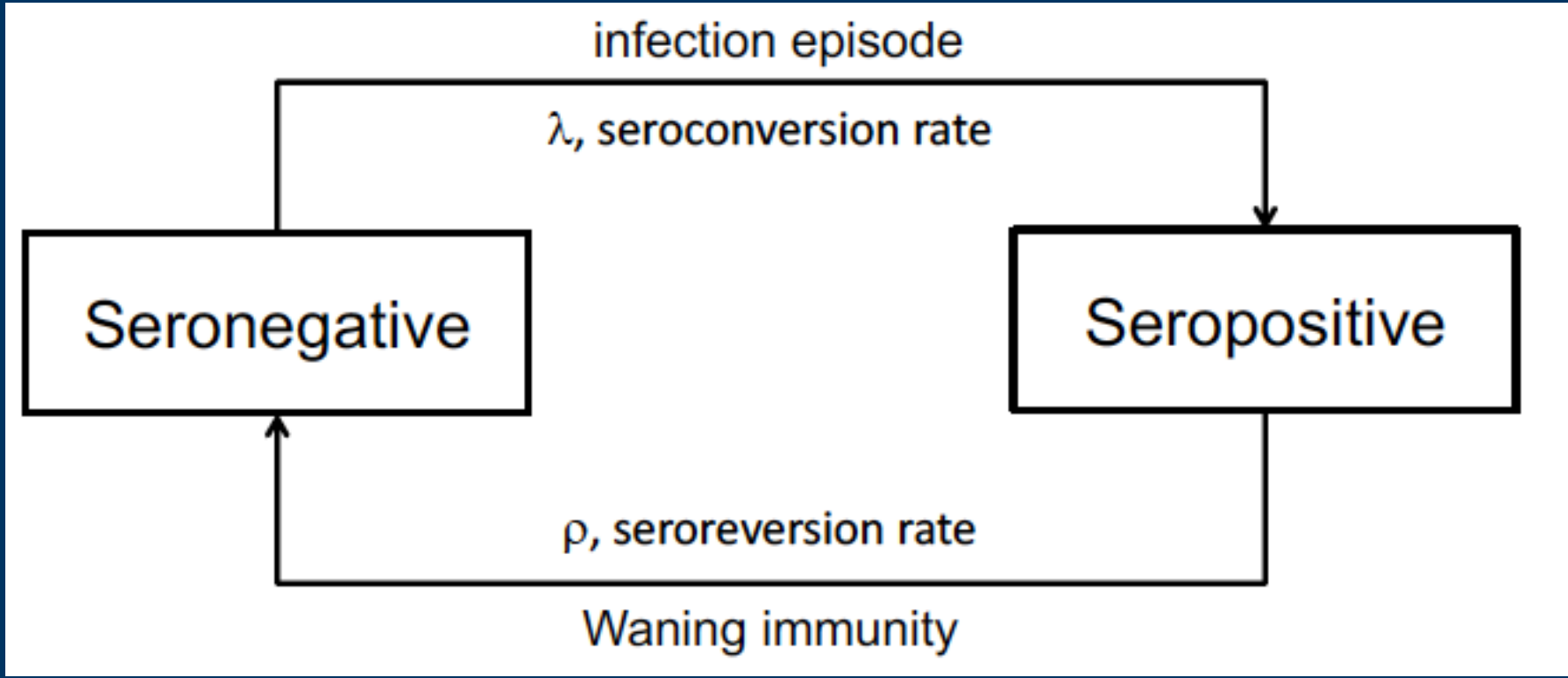
Focus on data from participants with status = positive (samples collected during the pandemic)

Estimate a two-Gaussian mixture model for S1RBD_NA_IgG_dil or S1RBD_NA_IgG_MFI using package mixtools. Estimate the 3-sigma cutoff and the respective sensitivity and specificity. Estimate the raw seroprevalence and corrected seroprevalence using the Rogan-Gladen Estimator.

Do you think the two-Gaussian mixture model is adequate for the data?

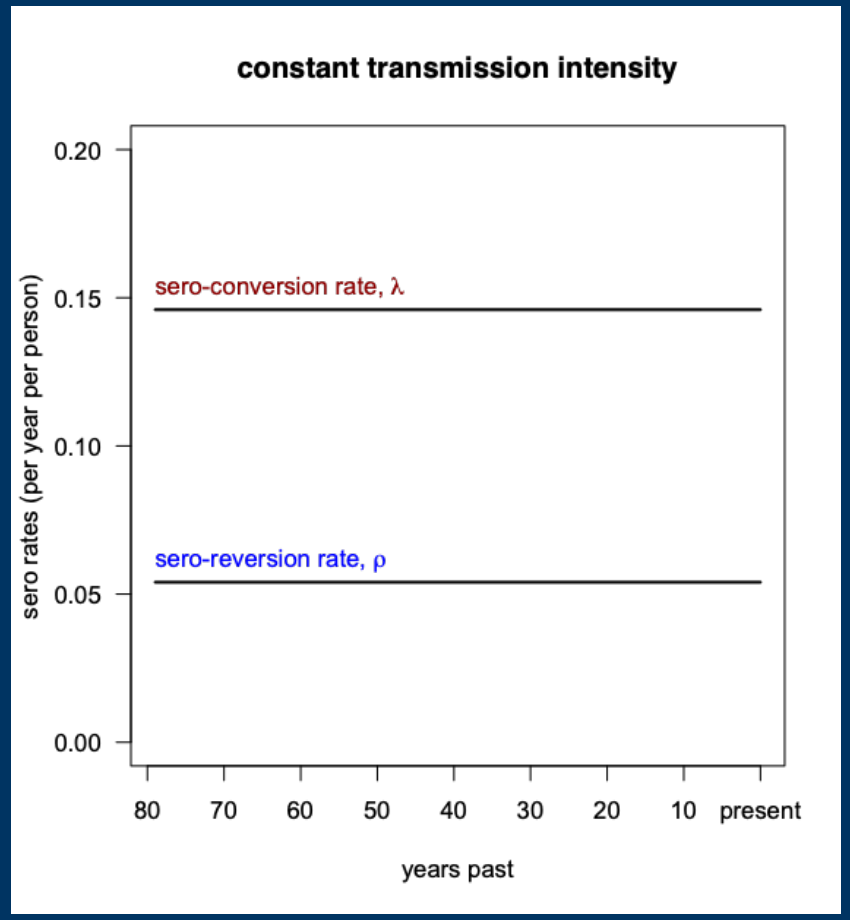
2.
Estimating seroconversion rate (using reversible catalytic models)

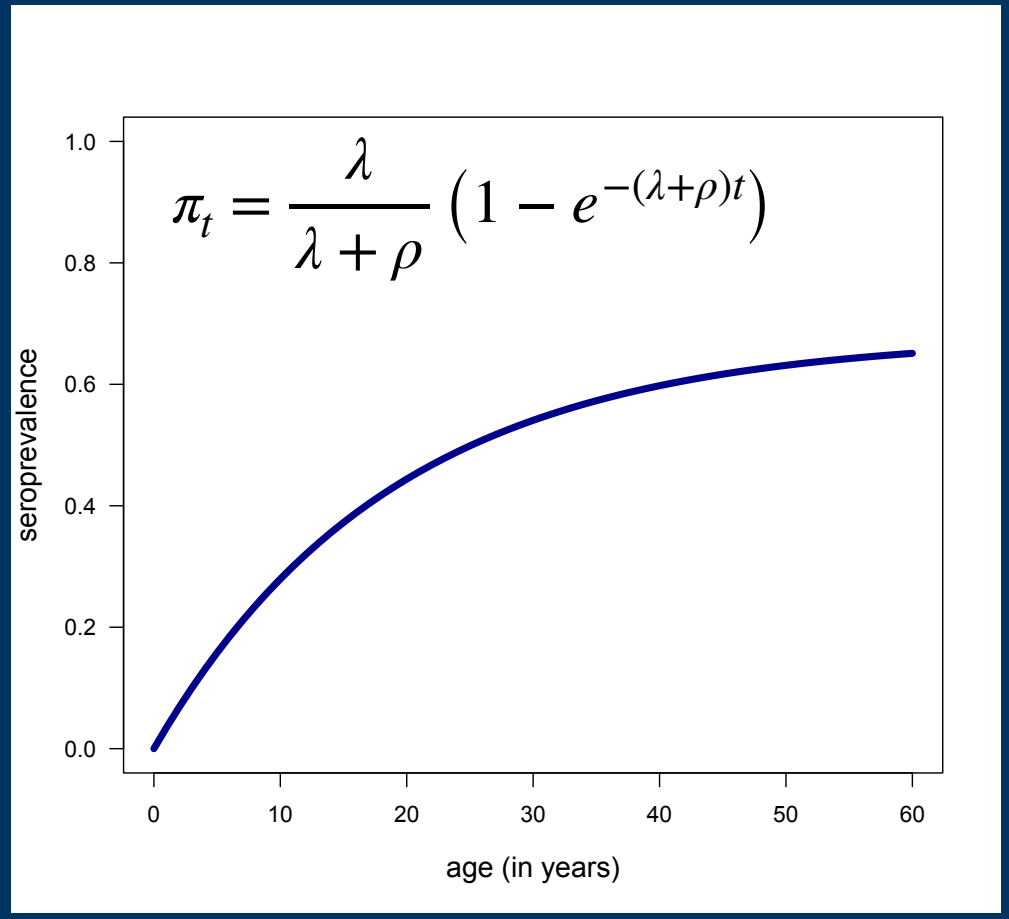
Reversible catalytic models



How can you model this stochastic process?

Constant transmission intensity



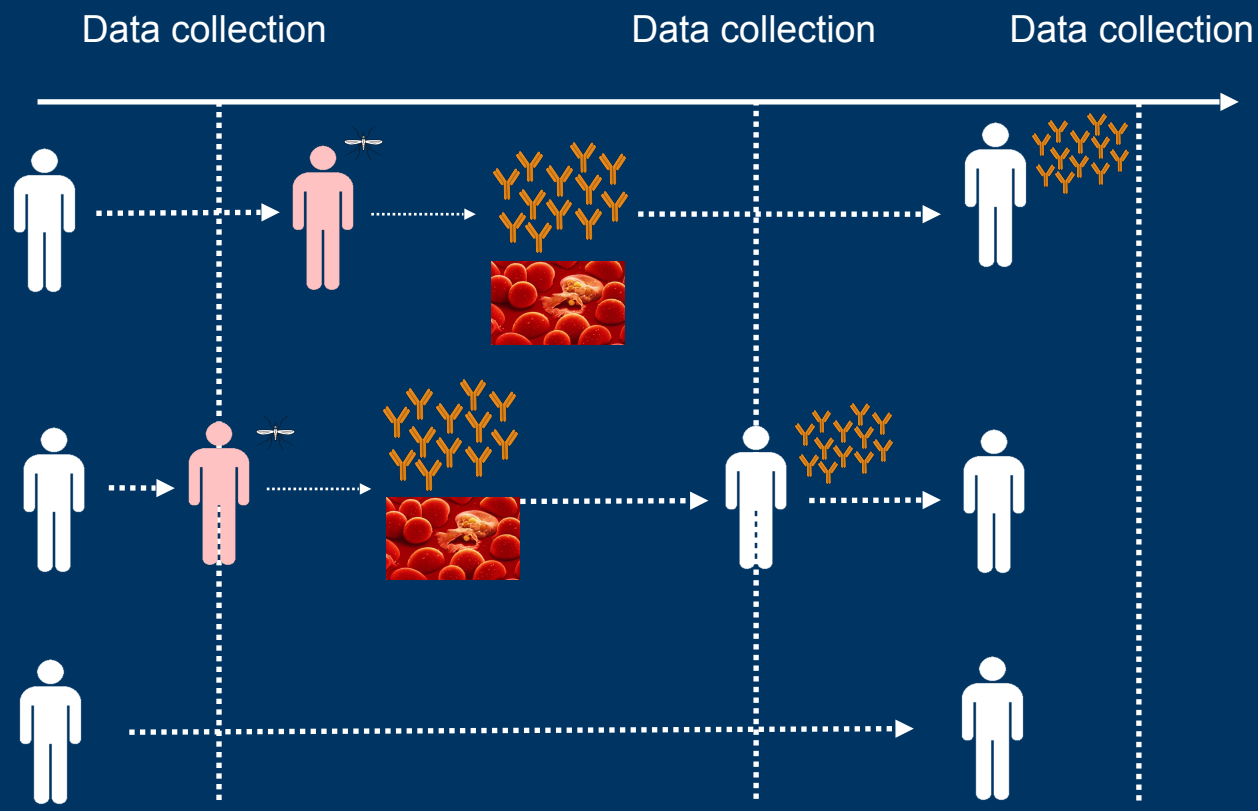


Markov chain formulation

$$P_t = P_0' e^{Qt} \qquad P_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad Q = \begin{pmatrix} -\lambda & \lambda \\ \rho & -\rho \end{pmatrix}$$

$$e^{Qt} = \begin{pmatrix} \frac{\rho}{\lambda+\rho} + \frac{\lambda}{\lambda+\rho} e^{-(\lambda+\rho)t} & \frac{\lambda}{\lambda+\rho}(1 - e^{-(\lambda+\rho)t}) \\ \frac{\rho}{\lambda+\rho}(1 - e^{-(\lambda+\rho)t}) & \frac{\lambda}{\lambda+\rho} + \frac{\rho}{\lambda+\rho} e^{-(\lambda+\rho)t} \end{pmatrix}$$

Longitudinal surveys



Statistical information: ++++

Direct observation of serological transitions

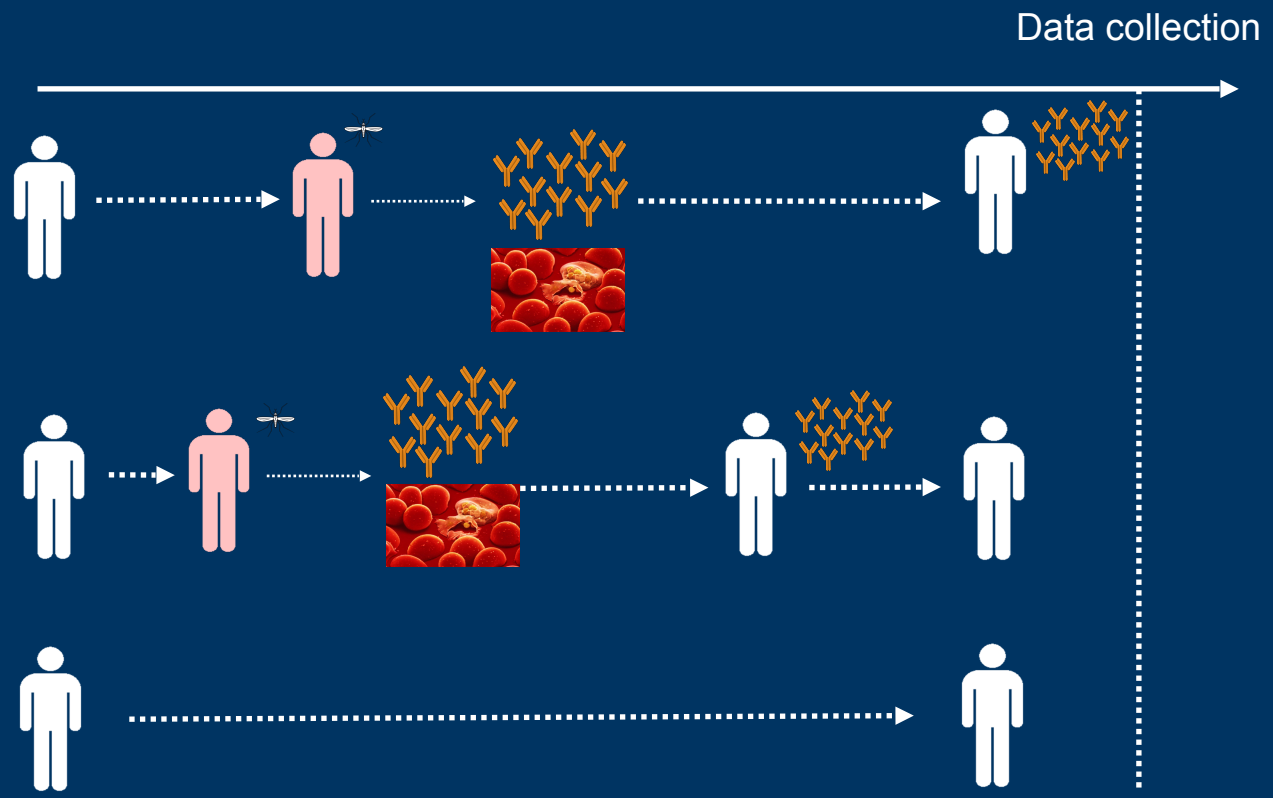
Execution difficulty: ++++

Time consuming

Sampling intensive

Participation adherence/drop-outs

Cross-sectional surveys



Statistical information: ++

No direct observation of serological transitions

Age as proxy of time

Execution difficulty: ++

Easy to engage participation

Quick sampling

What is the sampling model?

Longitudinal versus cross-sectional surveys

Type of Study	Seroconversion rate	Seroreversion rate
Longitudinal	**0.021** (0.001-0096)	**0.163** (0.001,0.729)
Cross-sectional	**0.023** (0.001,0.052)	**0.0001** (0.001,0.255)

Why is the seroreversion rate estimated differently using these two types of surveys?

Arnold et al (2014). PLoS One

Fixed seroreversion rate at 0

$$\rho = 0 \Rightarrow \pi_t = 1 - e^{-\lambda t}$$

$$\Rightarrow \log(1 - \pi_t) = -\lambda t$$

Do you know this model?

$$\Rightarrow \log(-\log(1 - \pi_t)) = \log \lambda + \log t$$

Do you know this model?

What are the practical implications in terms of model fitting?

Exercise: data_bioko.csv

OPEN ACCESS Freely available online

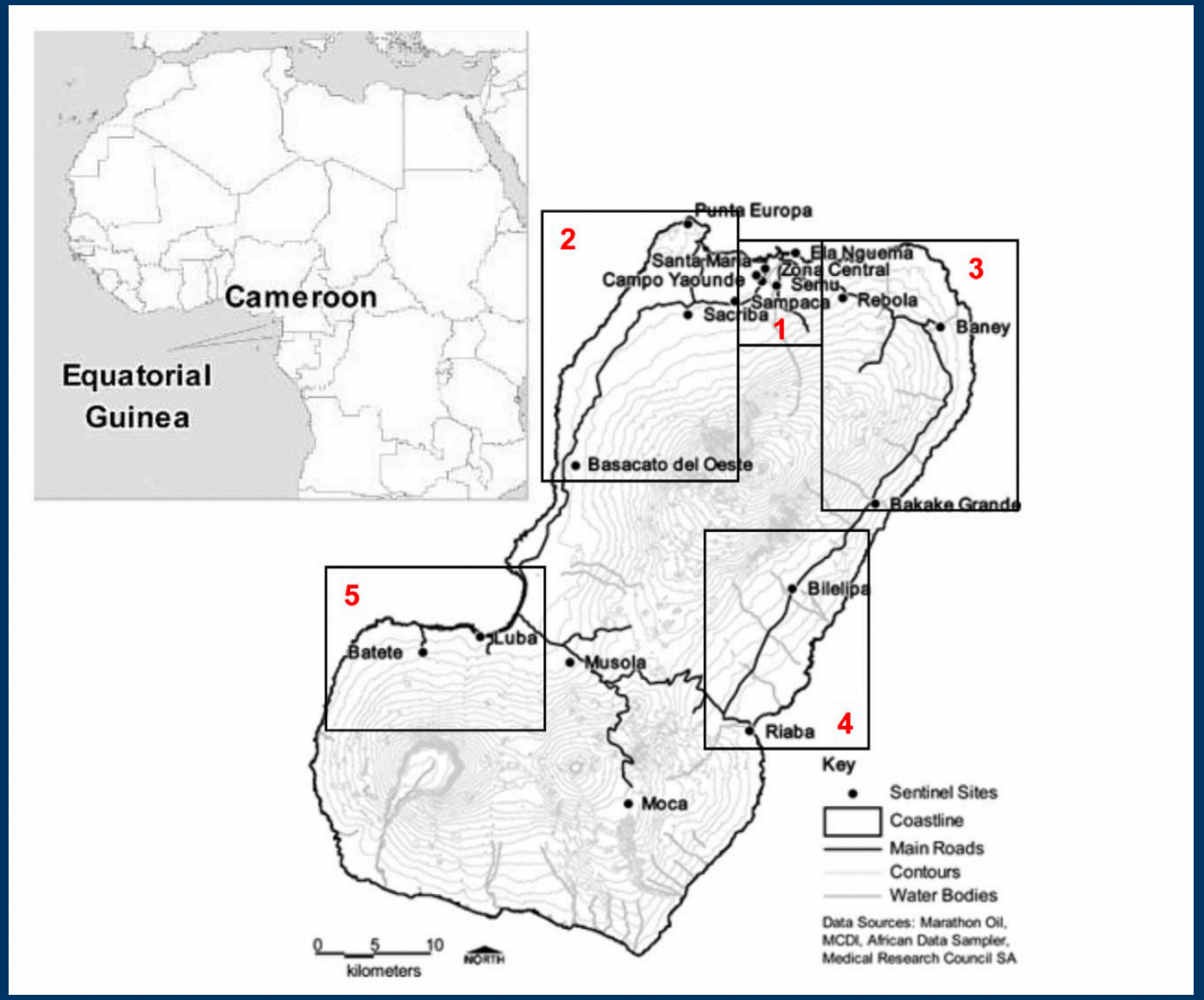
PLoS one

Serological Markers Suggest Heterogeneity of Effectiveness of Malaria Control Interventions on Bioko Island, Equatorial Guinea

Jackie Cook[1], Immo Kleinschmidt[2], Christopher Schwabe[3], Gloria Nseng[4], Teun Bousema[1], Patrick H. Corran[1], Eleanor M. Riley[1], Chris J. Drakeley[1]*

1 Department of Immunology and Infection, London School of Hygiene and Tropical Medicine, London, United Kingdom, **2** Department of Infectious Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, United Kingdom, **3** Medicinal Care Development International, Silver Spring, Maryland, United States of America, **4** Ministry of Health and Social Welfare, Malabo, Equatorial Guinea

Exercise: data_bioko.csv



Exercise: data_bioko.csv

Table 1. Demographic characteristics of the study population.

		% [n]						
		Malabo N = 2328	**North West N = 1749**	**North East N = 1323**	**South East N = 700**	**South West N = 588**	**Other** N = 699**	**Total N = 7387**
Age (years)	0–1	14.1 [324]	10.5 [182]	10.4 [137]	12.2 [85]	10.0 [58]	7.5 [52]	11.4 [838]
	1–5	21.1 [458]	18.0 [312]	19.8 [261]	14.6 [102]	16.8 [97]	15.2 [106]	18.6 [1363]
	5–15	26.3 [605]	30.6 [531]	30.1 [396]	21.0 [146]	24.0 [139]	28.7 [200]	27.5 [2017]
	15–90	38.6 [890]	41.0 [712]	39.8 [524]	52.2 [364]	49.2 [285]	48.6 [338]	42.5 [3113]
Sex	Female	61.2 [1410]	54.2 [932]	61.1 [805]	55.8 [389]	58.4 [338]	54.8 [382]	58.2 [4256]
House recently sprayed[1]	Yes	74.2 [1580]	81.2 [1306]	85.6 [1076]	81.7 [519]	89.5 [477]	87.9 [574]	81.2 [5532]
Slept under ITN[2]	Yes	82.6 [1629]	68.0 [988]	65.8 [797]	63.3 [404]	73.1 [385]	71.4 [449]	72.4 [4652]
Parasite positive	Yes	14.8 [300]	27.0 [374]	7.9 [94]	21.7 [135]	18.6 [97]	12.1 [75]	16.9 [1075]

[1]- within the previous 6 months.
[2]- on the night before the survey.
**Moca and Musola kept separate due to their high altitude.
doi:10.1371/journal.pone.0025137.t001

Exercise: data_bioko.csv

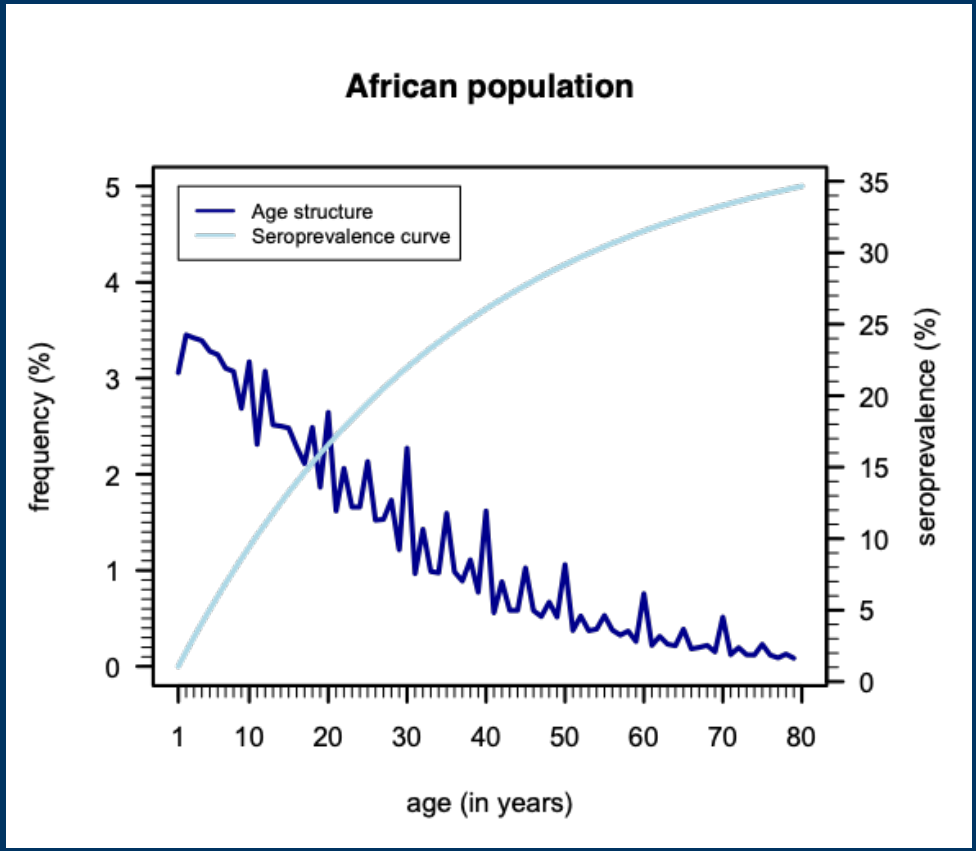
Estimate a two-Gaussian mixture model to serological data related to antibody against PfAMA1 protein (Ab_pfama1_titers) using mixtools package (normalmixEM command). Calculate the 3σ-rule cutoff and apply it to determine the serological status of each individual (seronegative/seropositive).

Estimate a reversible catalytic model in the seropositive data from the North West district with a seroreversion rate=0 using glm function and offset of the covariate log(Age). What is the estimate of the seroconversion rate? Does this model fit the data well?

Estimate a reversible catalytic model with constant seroconversion and seroreversion rates using seroaid source. Compare the seroconversion rate estimate with the previous one.

3.
Calculating sample size for controlling precision of seroconversion rate estimate

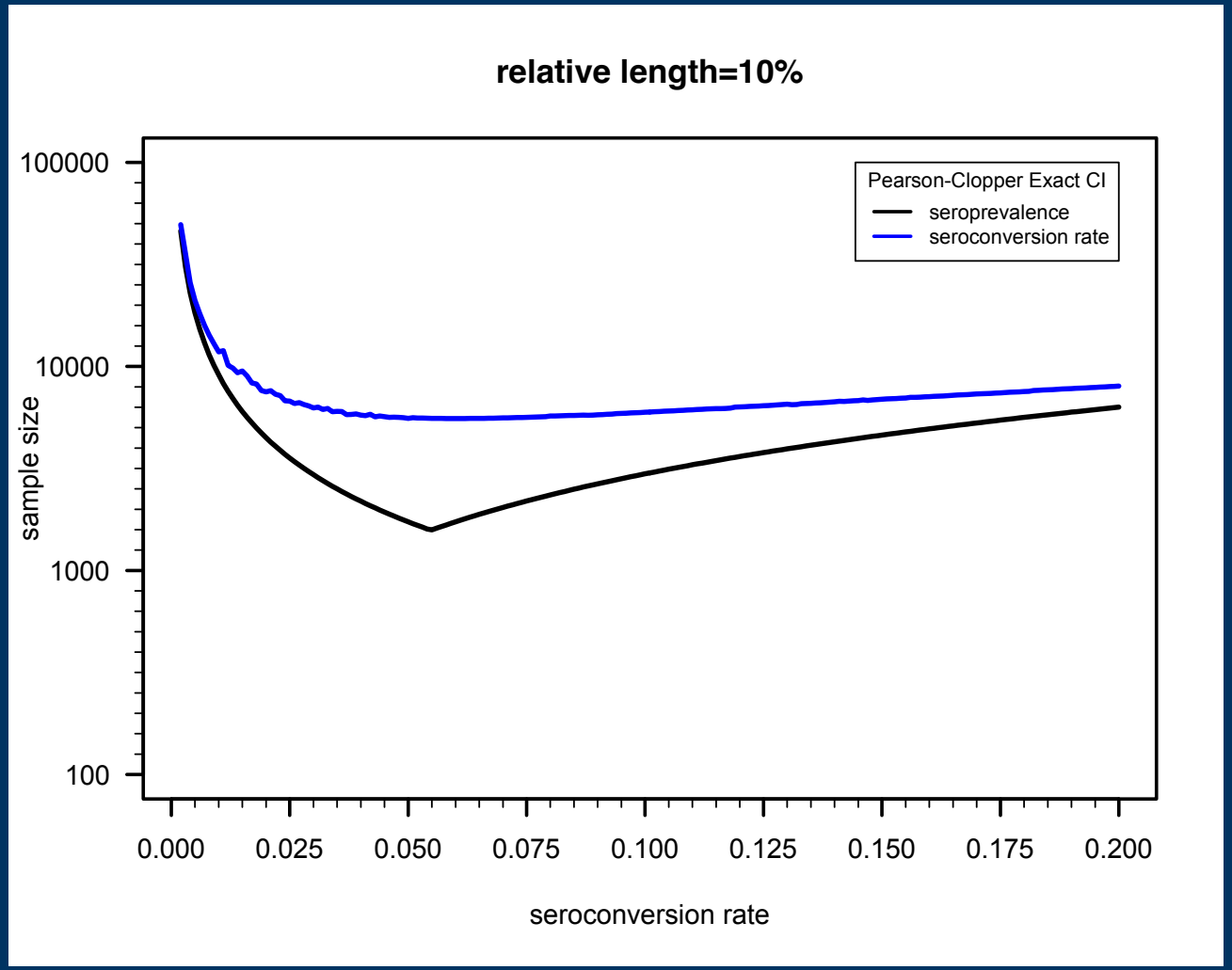
Sample size calculation for seroconversion rate



Sepúlveda et al (2015). Malaria Journal

Wald’s confidence interval

Sample size calculation for seroconversion rate



Sample size calculation in practice

1. Desired precision
2. Antibody with known seroreversion rate
3. Transmission intensity of the population
4. Age structure associated with sampling scheme
5. Type of confidence interval to be used

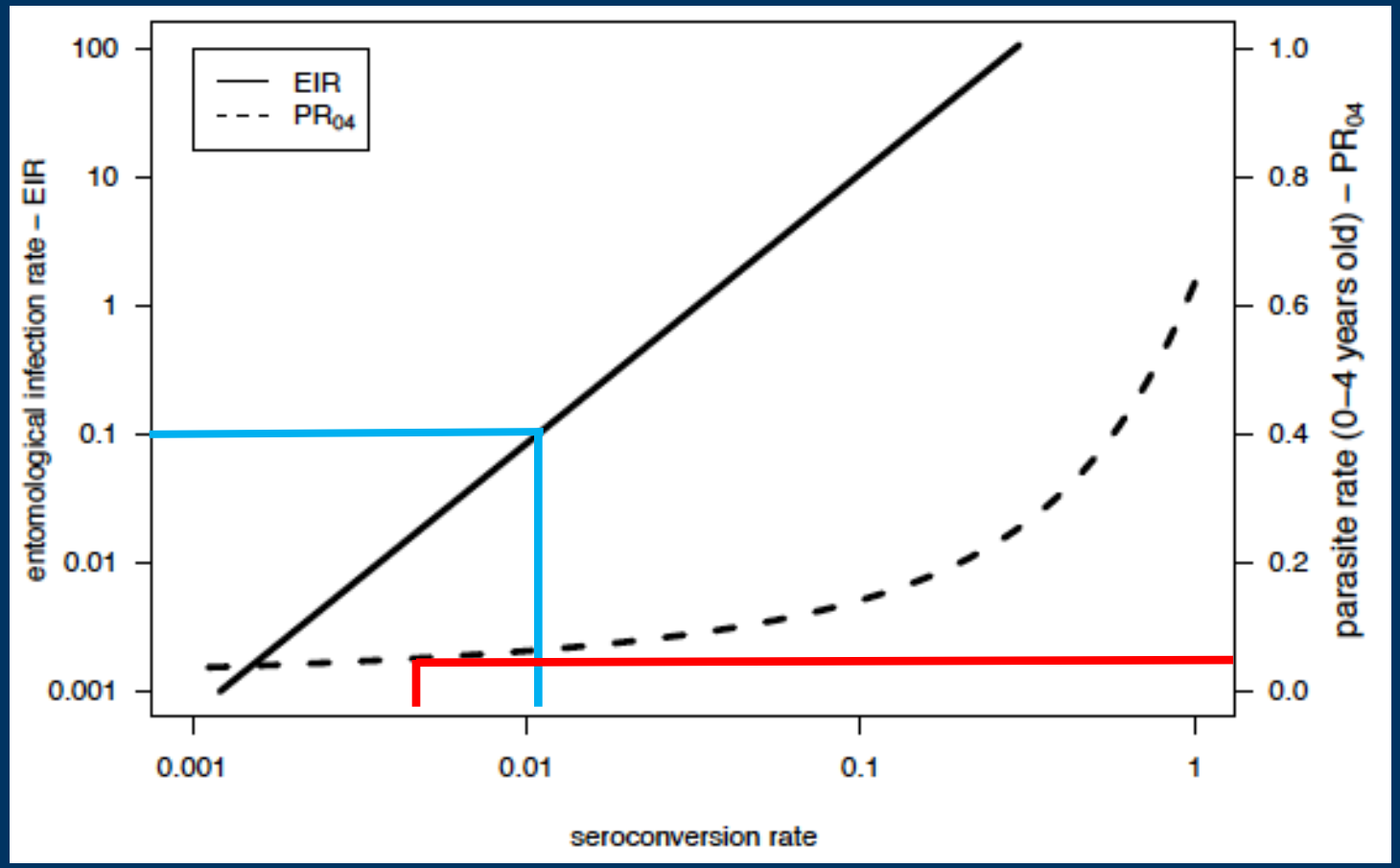
Identification of transmission intensity

Table 1 Expected relationship between EIR, PR_{04}, SCR and SP in African (AFR), Southeast Asian and South American (SEA + SA) populations where seroreversion rate was fixed at 0.017

			Seroprevalence	
EIR	PR_{04}	SCR	AFR	SEA + SA
0.01	0.050	0.0036	0.057	0.073
0.10	0.073	0.0108	0.156	0.195
1.00	0.119	0.0324	0.365	0.437
10.0	0.231	0.0969	0.647	0.720
100.0	0.625	0.2900	0.860	0.896

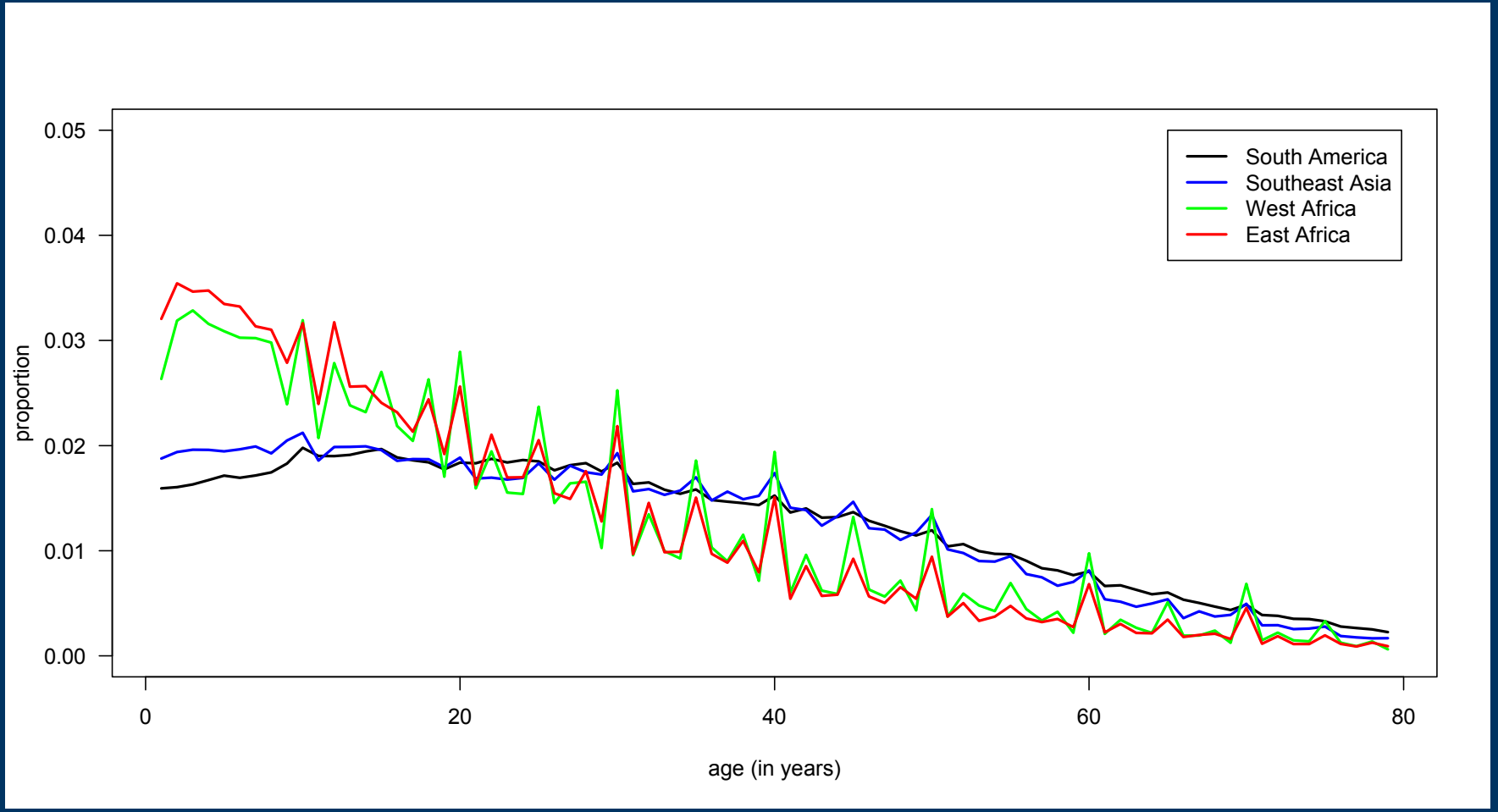
Sepúlveda & Drakeley (2015). Malaria J.

Identification of transmission intensity



Sepúlveda & Drakeley (2015). Malaria J.

Identification of age structure



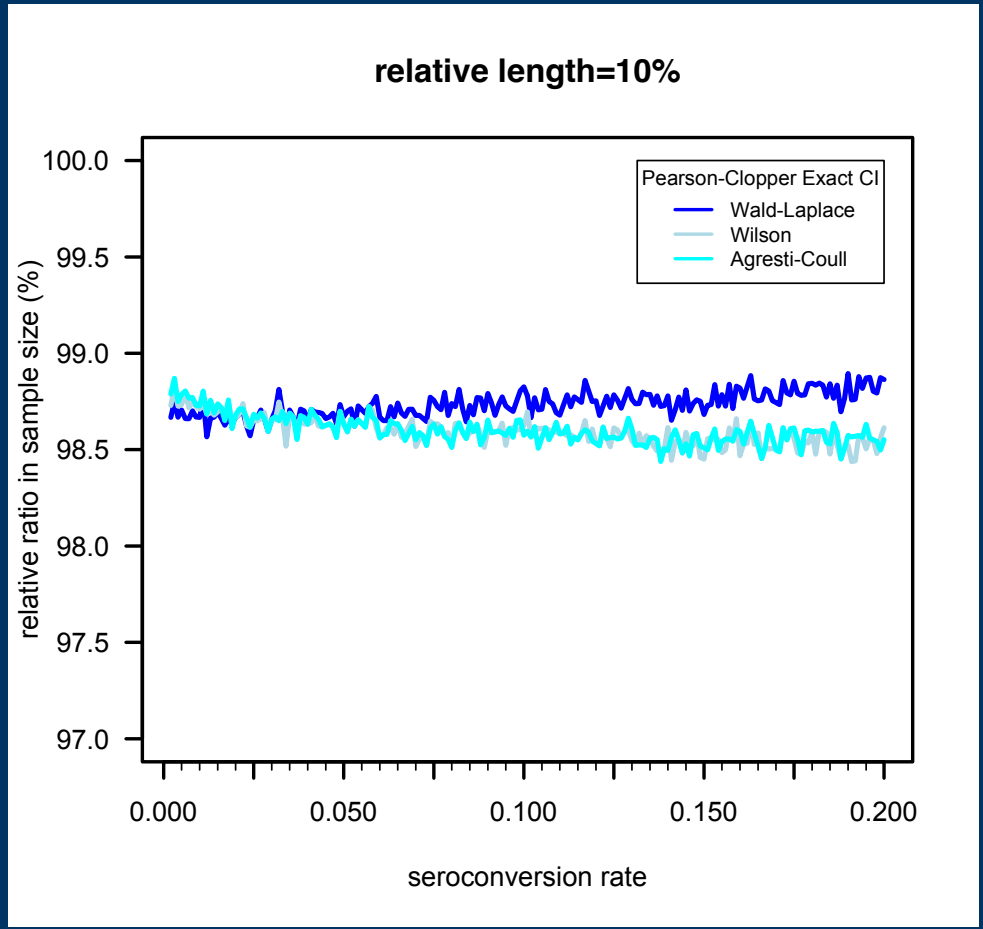
Type of confidence interval

Pearson-Clopper exact
Coverage higher than nominal confidence level

Wald
Degenerate when x=0 or x=n
Overshooting

Wilson
Good coverage in extreme probabilities

Agresti-Coull
Overshooting
Better coverage than the exact CI



Sepúlveda & Drakeley (2016). Proceedings of Portuguese Society for Statistics

Exercise:

Estimate the sample size for estimating the estimated seroconversion rate in the previous exercise assuming the age distribution of north West region and a fixed seroreversion rate of 0.017.

Aim to estimate the seroconversion with a relative length of 0.1, 0.25, and 0.5 of 95% confidence interval.

Use package RCMsize and command sample_s.