Biostatistics

Other topics

Nuno Sepúlveda, 22.01.2025

Syllabus

1. General review
    a. What is Biostatistics?
    b. Population/Sample/Sample size
    c. Type of Data – quantitative and qualitative variables
    d. Common probability distributions
    e. Work example – Malaria in Tanzania
2. Applications in Medicine
    a. Construction and analysis of diagnostic tools – Binomial distribution, sensitivity, specificity, ROC curve,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 Genetics, Genomics, and other 'omics 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, estimation of biological age
    c. Gene expression studies based on RNA-seq experiments – Tests based on Poisson and Negative-Binomial
4. Other Topics
    a. Estimation of Species diversity – Diversity indexes, Poisson mixture models
    b. Serological data analysis – Gaussian (skew-normal) mixture models

How to estimate species richness?

Abundance	Number of Species
0	D-k
1	m_1
2	m_2
3	m_3
....	
l	m_l

Augmented Species-Abundance distribution

Second solution (profile likelihood)

$$f(\{m_i\} \mid D, \{\theta_i\}) = \frac{D!}{(D-k)!m_1!\cdots m_l!}\theta_0^{D-M}\prod_{i=1}^{k}\theta_i^{m_i}$$

1. Fix $\hat{D}$=k
2. Estimate the parameter of the Poisson distribution via maximum likelihood and calculate the respective maximized log-likelihood. (What is the MLE of λ?
3. Do $\hat{D}+1$ in one unit and repeat previous step
4. Keep incrementing if the maximised log-likelihood is increasing
5. The estimate of D is the value immediately before when the maximized log-likelihood starts decreasing

How to estimate species richness?

Exercise: Data_lecture_13_TCR_diversity.csv

Estimate the species richness D for the DP CD3low cells using the second solution.

	Thymus			Lymph nodes	
i	DP CD3low	SP CD4$^+$	SP CD8$^+$	LN CD4$^+$	LN CD8$^+$
1	79	33	16	34	17
2	17	6	3	8	8
3	6	2	3	2	1
4	5	2	5	1	2
5	1	0	3	0	1
6	1	0	1	0	0
7	1	0	1	0	0
8		0	1	1	0
10		1	0	1	0
11		0	1	0	0
16		1		0	0
20		0		1	0
21		0			1
28		1			0
52					1

How to estimate species richness?

Abundance	Number of Species
0	D-k
1	m_1
2	m_2
3	m_3
....	
l	m_l

Augmented Species-Abundance distribution

Calculation of a 95% confidence interval using the profile likelihood

Use the critical value of the Wilks's ratio test

$$H_0 : D = D_0 \text{ versus } H_1 : D \neq D_0$$

$$\Lambda = -2(\log L_{D_0} - \log L_{\hat{D}}) \,|\, H_0 \rightsquigarrow \chi^2_{(1)}$$

critical value $= q_{95\%,\chi^2_{(1)}}$

accept H_0 if $\Lambda < q_{95\%,\chi^2_{(1)}}$ reject H_0, otherwise

How to estimate species richness?

Abundance	Number of Species
0	D-k
1	m_1
2	m_2
3	m_3
....	
l	m_l

Augmented Species-Abundance distribution

Calculation of a 95% confidence interval using the profile likelihood

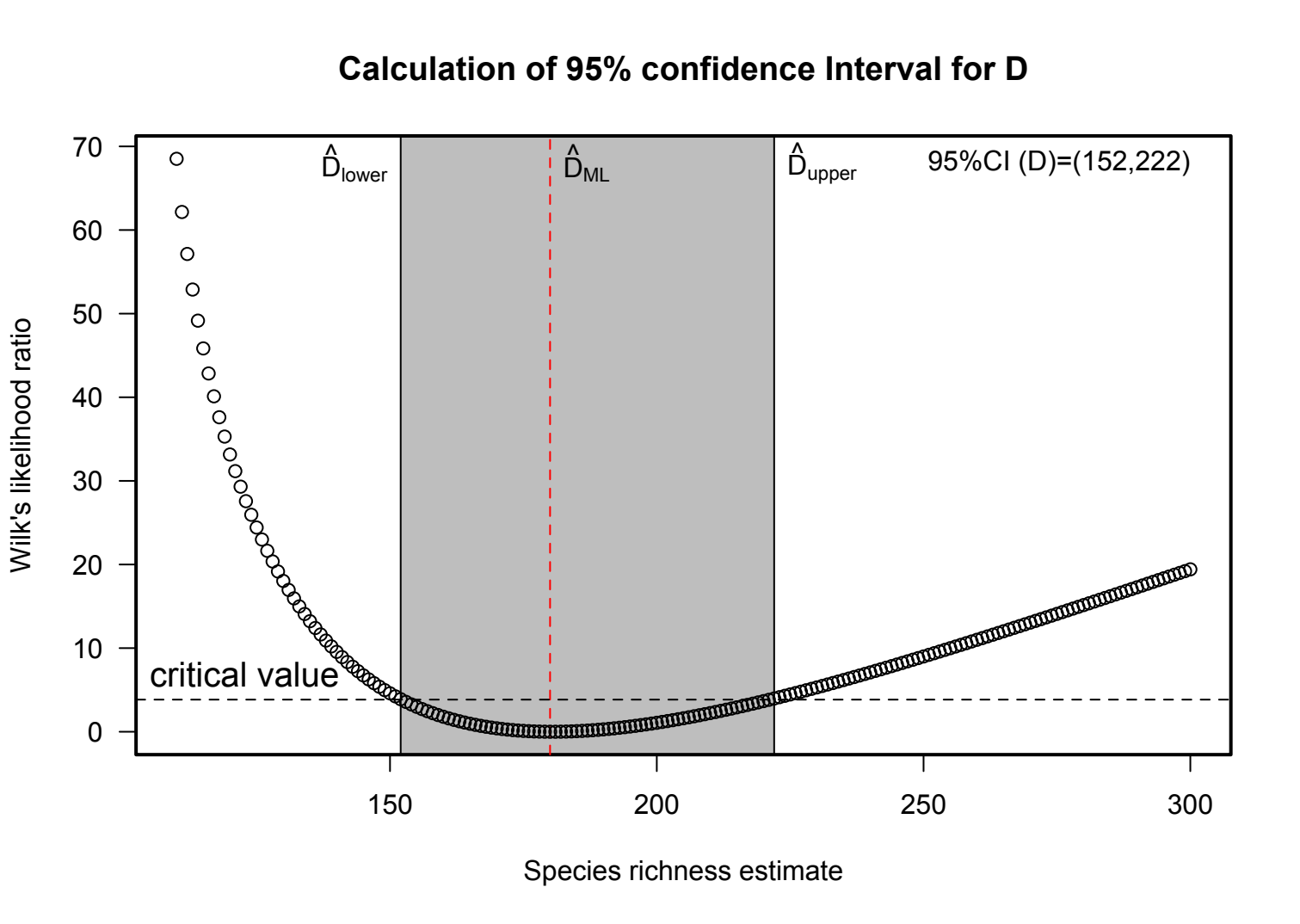
Use the critical value of the Wilks's ratio test

$$q_{95\%,\chi^2_{(1)}} = -2(\log L_{D_0} - \log L_{\hat{D}})$$

$$(\hat{D}_{lower}; \hat{D}_{upper})$$

$\hat{D}_{lower}$ and $\hat{D}_{upper}$ are the solutions of the above question

How to estimate species richness?



How to estimate species richness?

Abundance	Number of Species
1	m_1
2	m_2
3	m_3
....	
l	m_l
>l	0

Augmented Species-Abundance distribution

Pearson's goodness of fit test to check whether the model fits the data well

Use only the observed data

$$f(\{m_i\} \mid k, \{\theta_i\}) = \frac{k!}{m_1! \cdots m_l!} \prod_{i=1}^{k} \left(\frac{\theta_i}{1-\theta_0} \right)^{m_i}$$

$$\hat{\theta}_i = \frac{e^{-\hat{\lambda}} \hat{\lambda}^i}{i!}$$

Exercise: Data_lecture_13_TCR_diversity.csv

Calculate the confidence interval for the species richness D for the DP CD3low cells using the profile likelihood plot. Check whether the Poisson model fits the data well using the Pearson's goodness of fit test.

	Thymus			Lymph nodes	
i	DP CD3low	SP $CD4^+$	SP $CD8^+$	LN $CD4^+$	LN $CD8^+$
1	79	33	16	34	17
2	17	6	3	8	8
3	6	2	3	2	1
4	5	2	5	1	2
5	1	0	3	0	1
6	1	0	1	0	0
7	1	0	1	0	0
8		0	1	1	0
10		1	0	1	0
11		0	1	0	0
16		1		0	0
20		0		1	0
21		0			1
28		1			0
52					1

Poisson-Gamma mixture model for estimating diversity richness

Abundance	Number of Species
0	D-k
1	m_1
2	m_2
3	m_3
....	
l	m_l

Augmented Species-Abundance distribution

Modelling θ_i

$$\theta_i = P[X = i \mid \lambda]$$

$X \mid \lambda \rightsquigarrow \text{Poisson}(\lambda)$ $\qquad$ $\lambda \mid \alpha, \beta \rightsquigarrow \text{Gamma}(\alpha, \beta)$

$$P[X = x] = \int_0^\infty P[X = x \mid \lambda] P[\lambda] d\lambda = \int_0^\infty \frac{e^\lambda \lambda^x}{x!} \times \frac{\beta^\alpha \lambda^{\alpha-1} e^{\beta\lambda}}{\Gamma(\alpha)} d\lambda$$

$$= \frac{\Gamma(i+\alpha)}{\Gamma(i+1)\Gamma(\alpha)} \left(\frac{\beta}{\beta+1}\right)^\alpha \left(\frac{1}{\beta+1}\right)^i$$

Negative Binomial

How to estimate species richness?

Abundance	Number of Species
0	D-k
1	m_1
2	m_2
3	m_3
....	
l	m_l

Augmented Species-Abundance distribution

First solution (truncated Negative Binomial)

$$k \mid D, \theta_0 \rightsquigarrow \text{Binomial}(D, 1-\theta_0)$$

$$f(\{m_i\} \mid k, \{\theta_i\}) = \frac{k!}{m_1! \cdots m_l!} \prod_{i=1}^{k} \left(\frac{\theta_i}{1-\theta_0} \right)^{m_i}$$

1. Estimate a Poisson truncated at zero using raw data only
2. Estimate D from the binomial using $\hat{D} = \dfrac{k}{1-\hat{\theta}_0}$

$$\hat{\theta}_0 = e^{-\hat{\lambda}}$$

How to estimate species richness?

Abundance	Number of Species
0	D-k
1	m_1
2	m_2
3	m_3
....	
l	m_l

Augmented Species-Abundance distribution

Second solution (profile likelihood)

$$f(\{m_i\} \mid D, \{\theta_i\}) = \frac{D!}{(D-k)!m_1!\cdots m_l!}\theta_0^{D-M}\prod_{i=1}^{k}\theta_i^{m_i}$$

1. Fix $\hat{D}$=k
2. Estimate the parameters of the Negative distribution via maximum likelihood and calculate the respective maximized log-likelihood. (What is the MLE of λ?
3. Do $\hat{D}+1$ in one unit and repeat previous step
4. Keep incrementing if the maximised log-likelihood is increasing
5. The estimate of D is the value immediately before when the maximized log-likelihood starts decreasing

Exercise: Data_lecture_13_TCR_diversity.csv

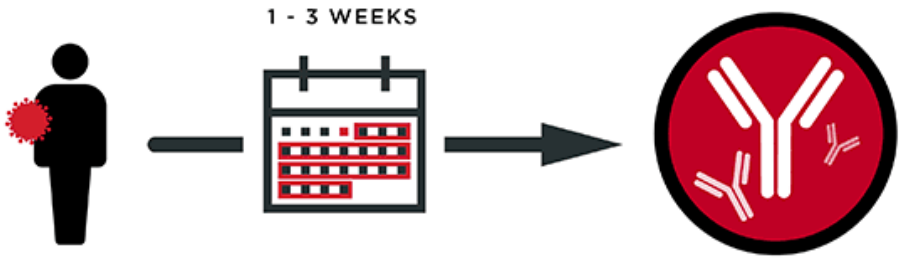
Estimate the species richness D for the DP CD3low cells using the Negative Binomial distribution. Estimate via the second solution.

	Thymus			Lymph nodes	
i	DP CD3low	SP $CD4^+$	SP $CD8^+$	LN $CD4^+$	LN $CD8^+$
1	79	33	16	34	17
2	17	6	3	8	8
3	6	2	3	2	1
4	5	2	5	1	2
5	1	0	3	0	1
6	1	0	1	0	0
7	1	0	1	0	0
8		0	1	1	0
10		1	0	1	0
11		0	1	0	0
16		1		0	0
20		0		1	0
21		0			1
28		1			0
52					1

Serology

ANTIBODY AGAINST SARS-CoV-2

Serology, or **antibody**, testing checks a sample of a person's blood to look for antibodies against SARS-CoV-2, the virus that causes COVID-19. Antibodies usually become detectable in the blood **1-3 weeks** after someone is infected.

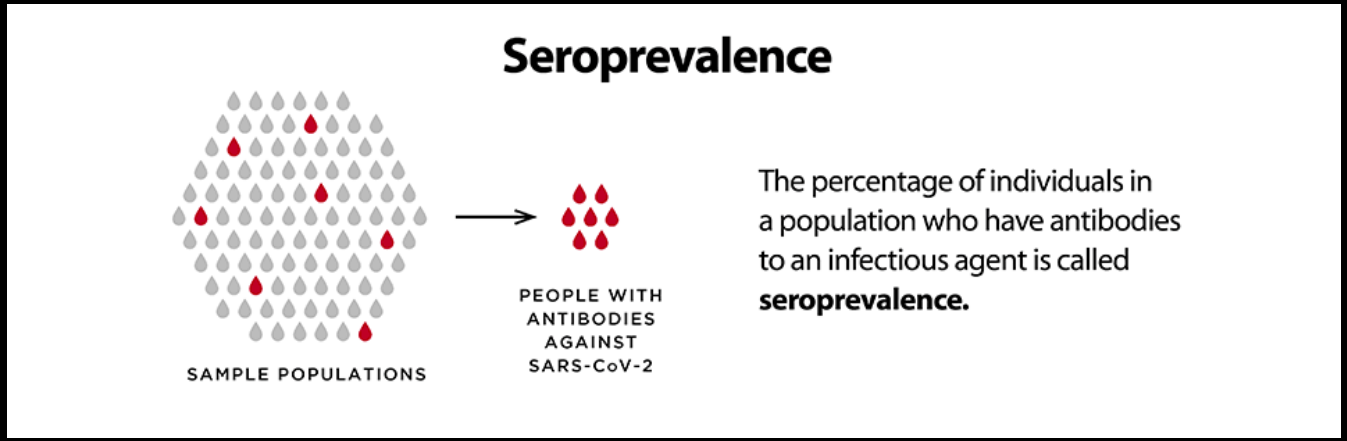


Person has detectable level of antibodies.*

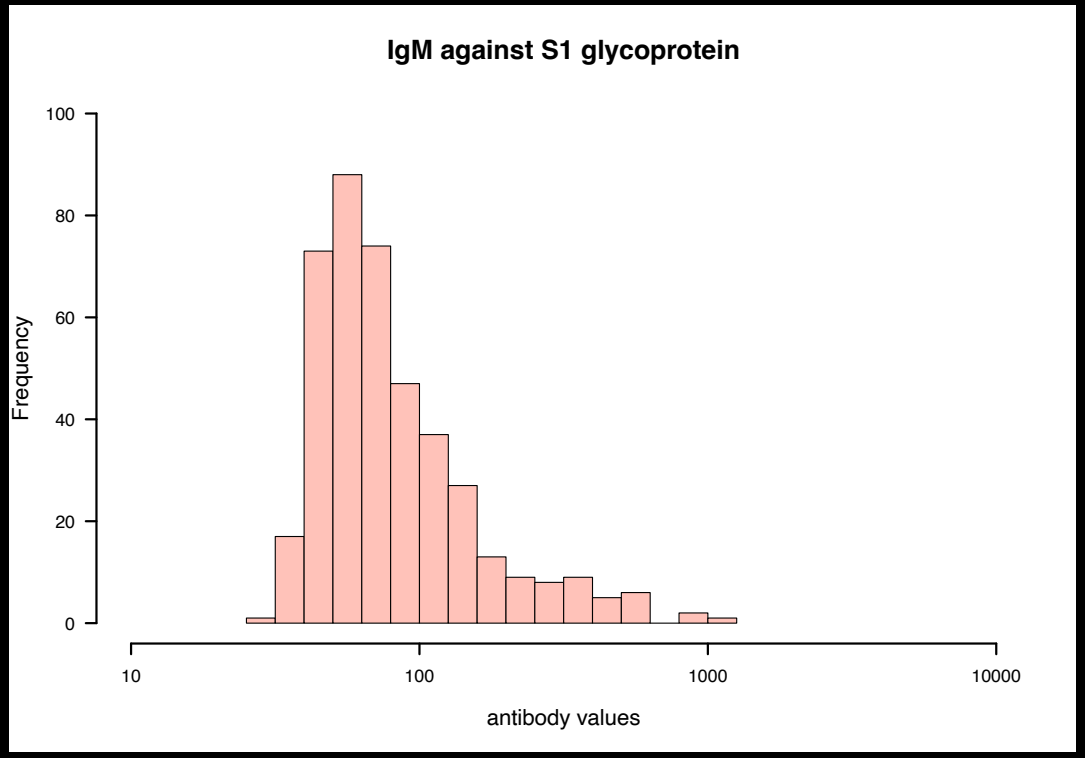
PERSON INFECTED

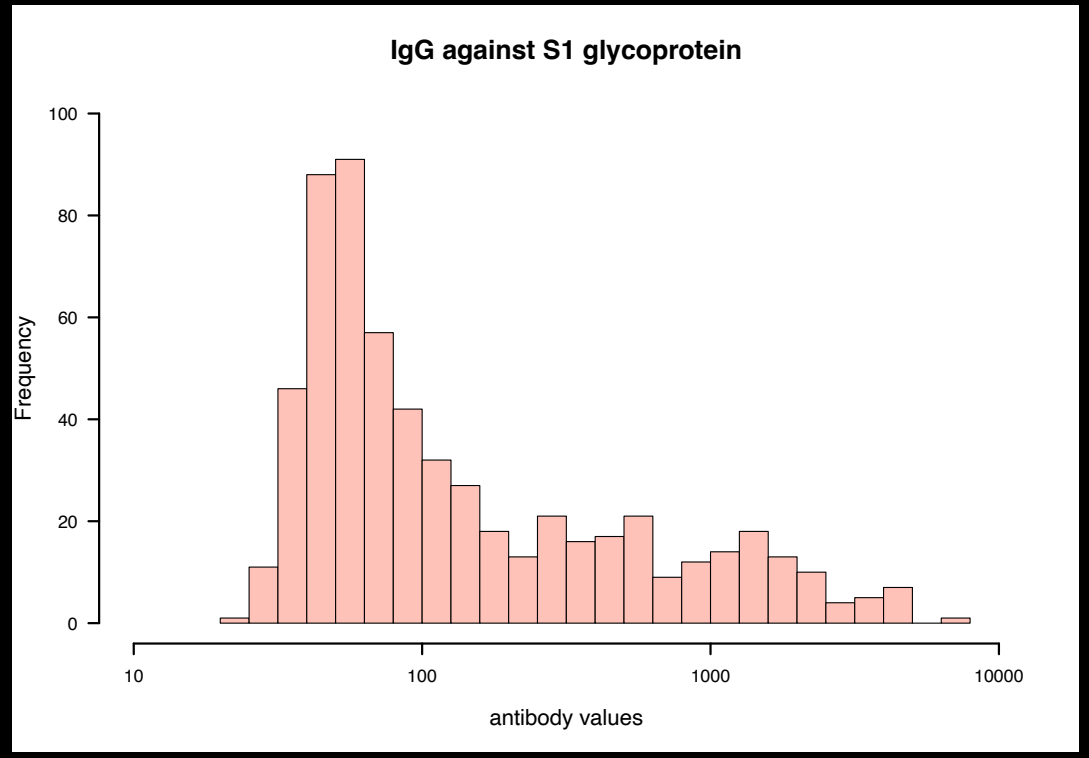
*Some people may take longer than 3 weeks to develop antibodies, and some people may not develop antibodies. It is currently unknown how long antibodies are detectable after infection.

Sero-epidemiological surveys



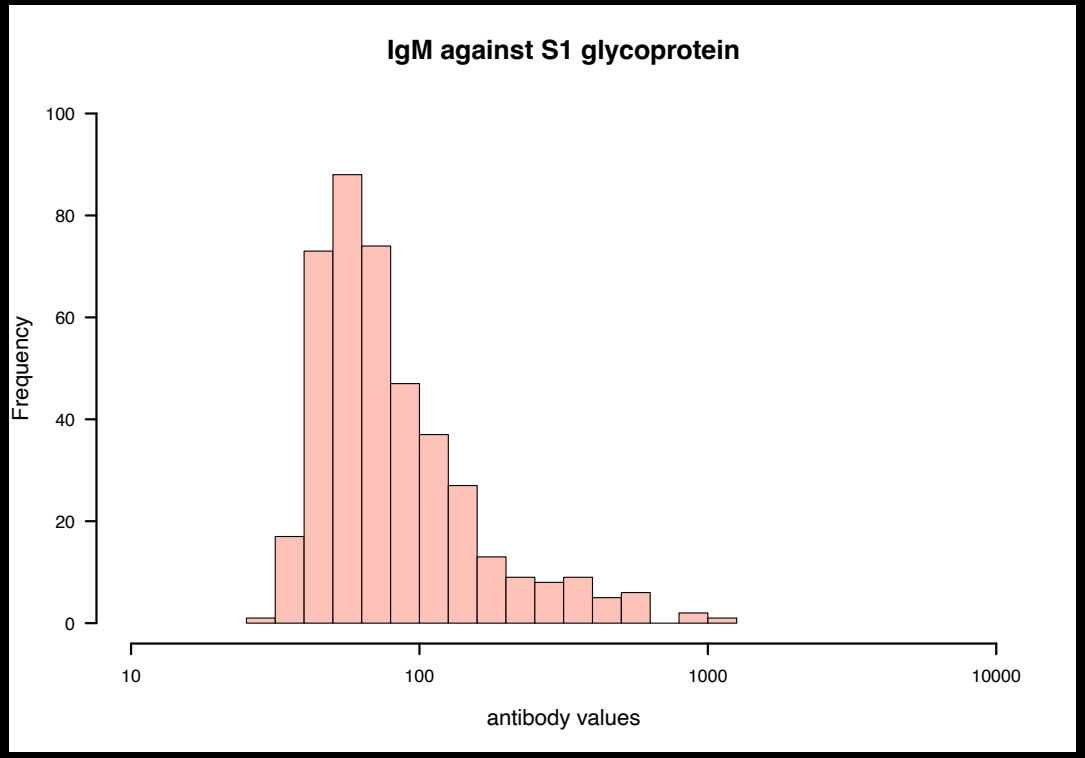
Antibody data are intrinsically quantitative

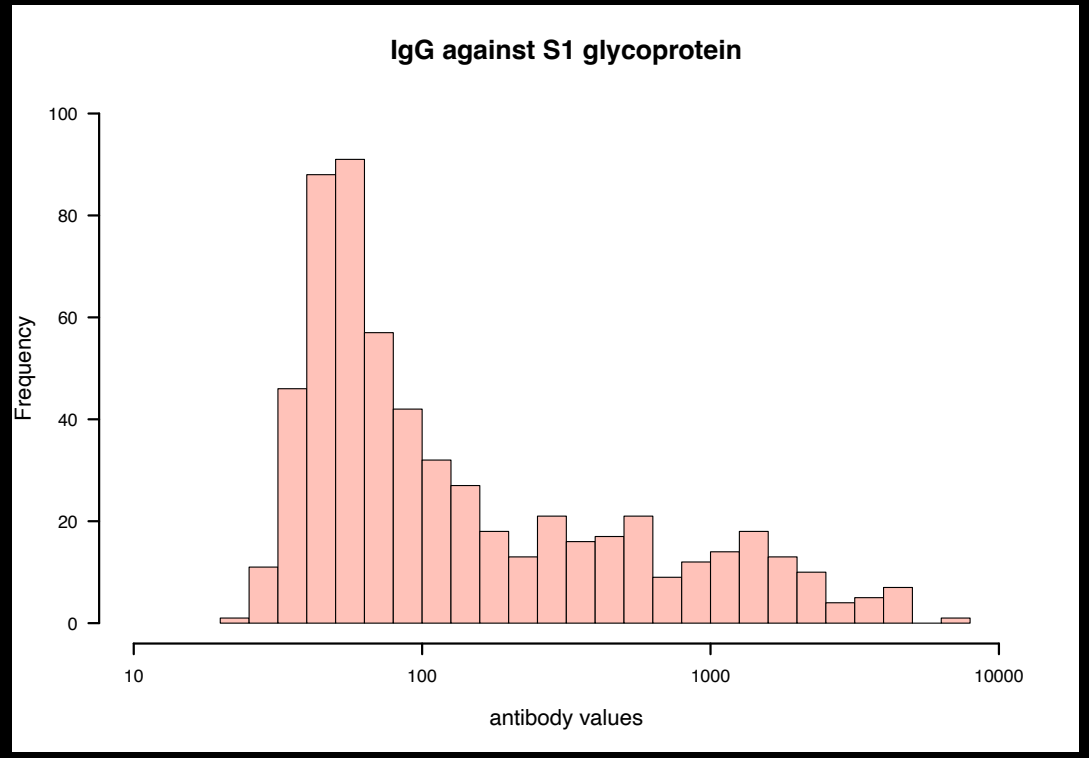




Rosado et al (2020). Serological signatures of SARS-CoV-2 infection: Implications for antibody-based diagnostics. medRxiv 2020.05.07.20093963.

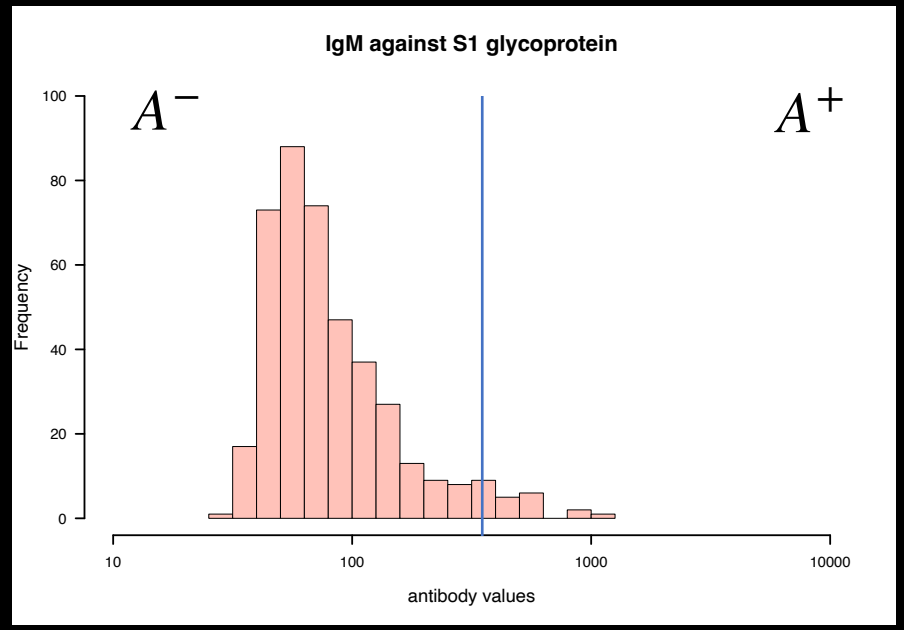
Who are the seropositive individuals?

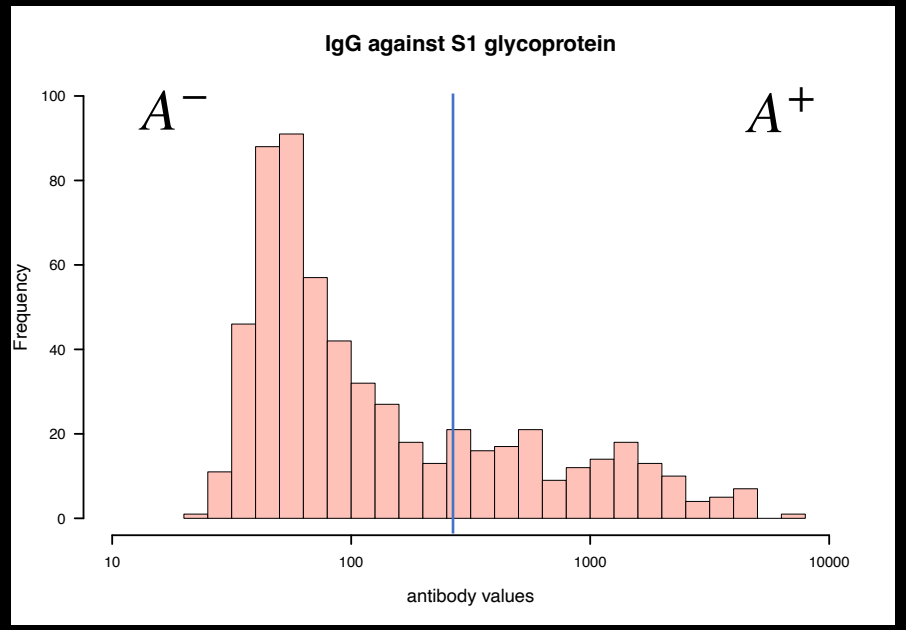


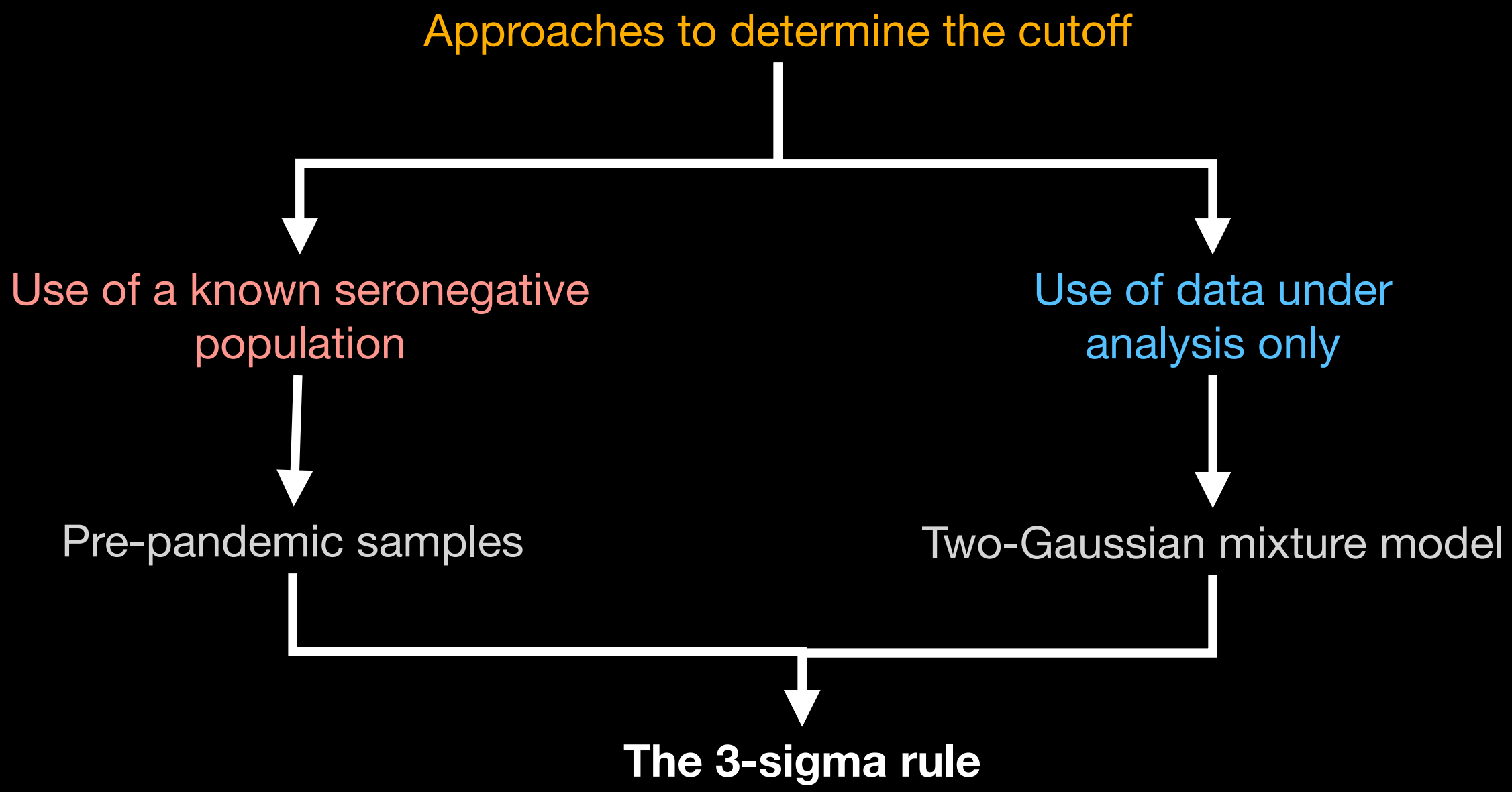


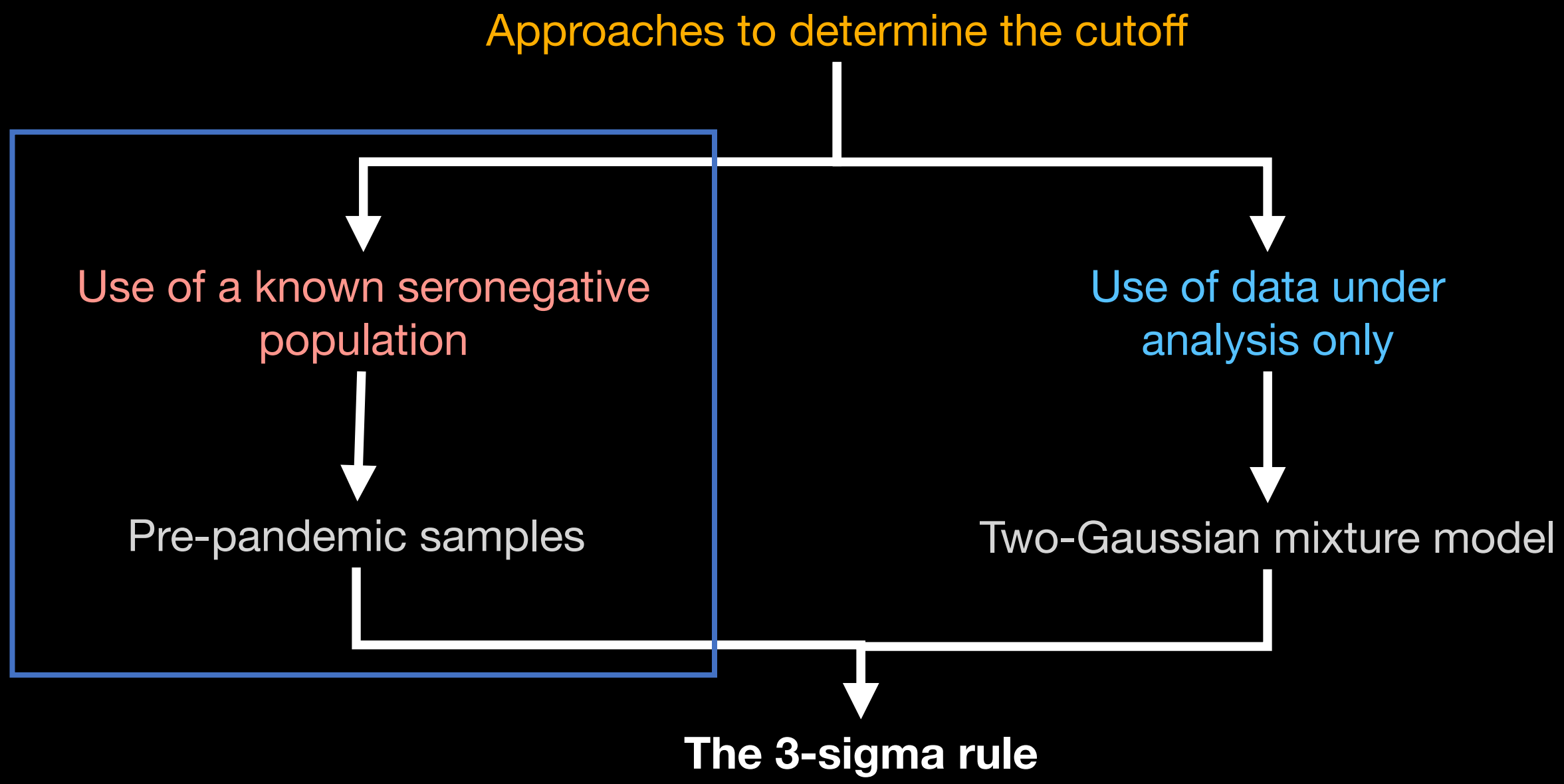
Rosado et al (2020). Serological signatures of SARS-CoV-2 infection: Implications for antibody-based diagnostics. medRxiv 2020.05.07.20093963.

How to determine the cut-off?






Approaches to determine the cutoff
Use of a known seronegative population
Use of data under analysis only
Pre-pandemic samples
Two-Gaussian mixture model
The 3-sigma rule


Approaches to determine the cutoff
Use of a known seronegative population
Pre-pandemic samples
Use of data under analysis only
Two-Gaussian mixture model
The 3-sigma rule

The 3-sigma rule

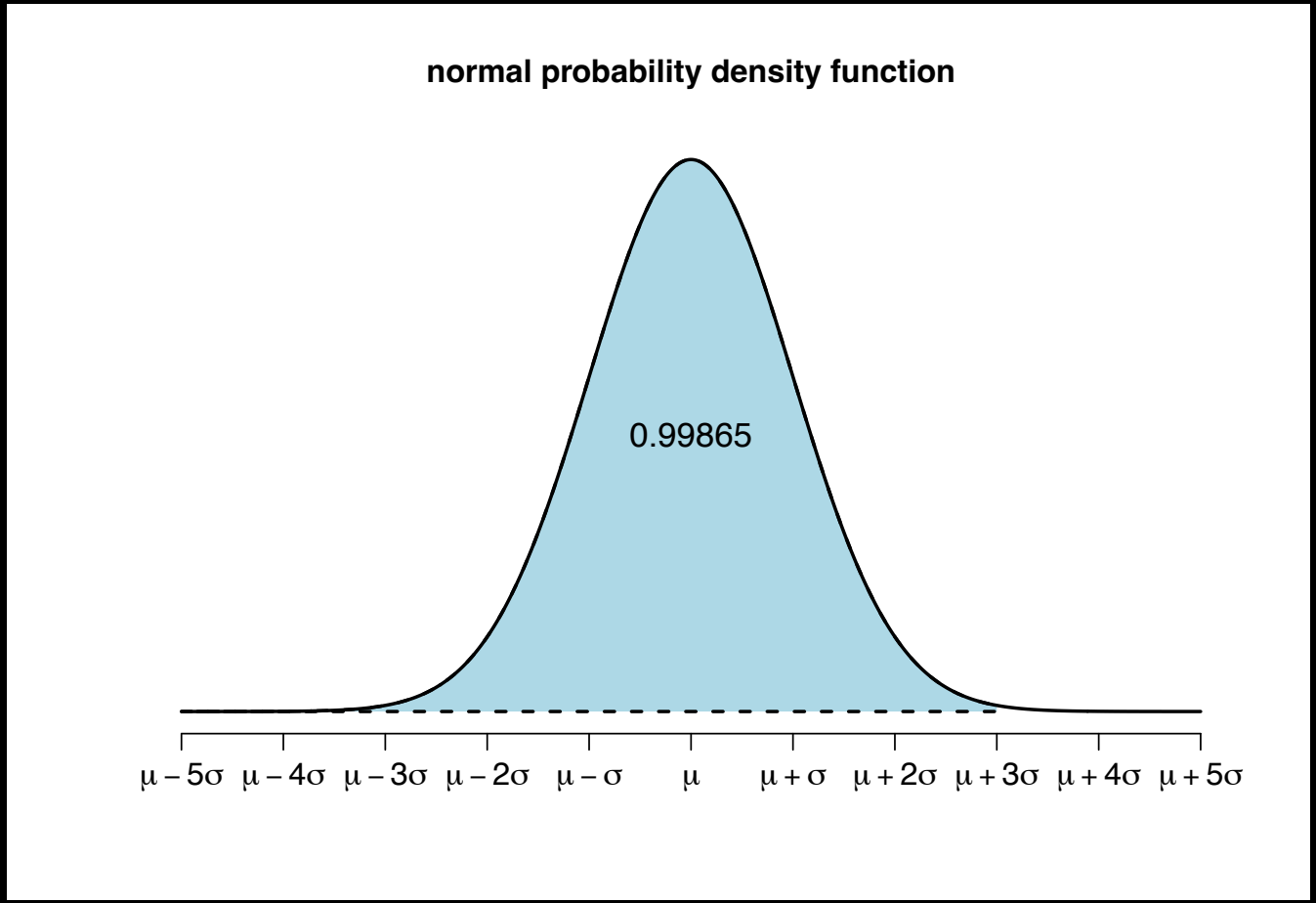
$$\mu_{A^-} = E\left[X|A^-\right]$$

$$\sigma_{A^-} = \sqrt{Var\left[X|A^-\right]}$$

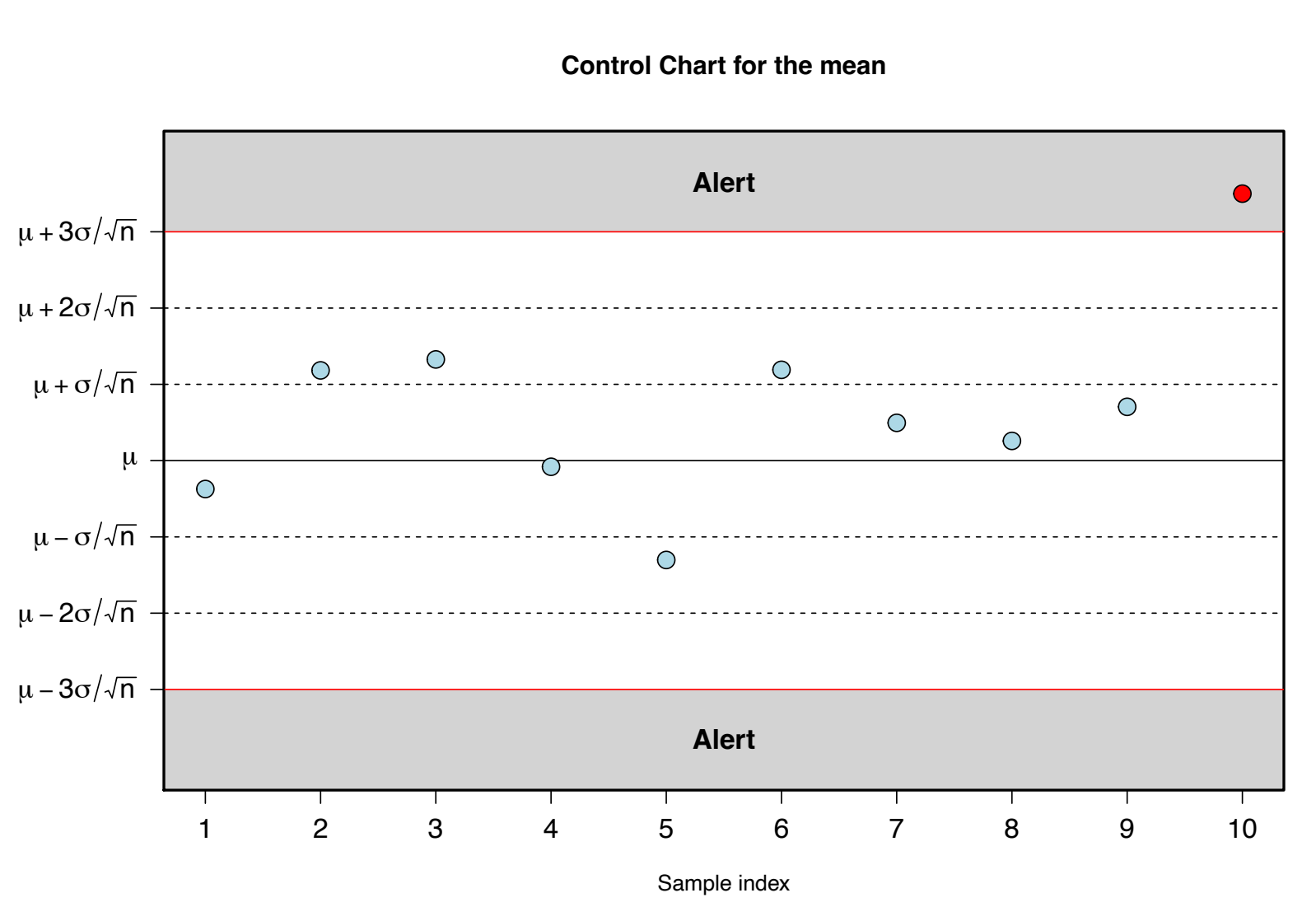
Seronegative, if $X_i \leq \mu_{A^-} + 3\sigma_{A^-}$

Seropositive, otherwise

The link to the Normal distribution



Quality control (Shewhart)



In practice (known seronegative population)

$\mu_{A^-} \rightarrow \bar{X}_{A^-}$

$\sigma_{A^-} \rightarrow S_{A^-}$

Seronegative, if $x_i \leq \bar{X}_{A^-} + 3s_{A^-}$

Seropositive, otherwise

Theoretical property of the 3-sigma

Cantelli-Chebyshev inequality

$$P\left[X \geq \mu + \lambda\right] \leq \frac{\sigma^2}{\sigma^2 + \lambda^2}, \text{ if } \lambda > 0$$

$$\mu = E\left[X\right] \qquad \sigma^2 = Var\left[X\right] < \infty$$

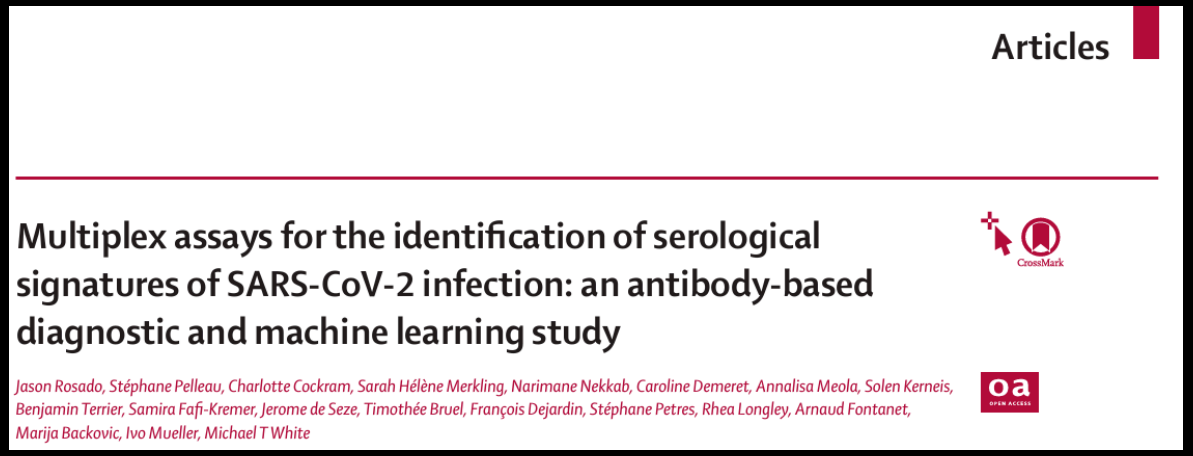
Application to $\lambda = 3\sigma$

$$P\left[X \geq \mu_{A^-} + 3\sigma_{A^-}\right] \leq \frac{1}{10} \equiv 0.1$$

$$\downarrow$$

$$P\left[X < \mu_{A^-} + 3\sigma_{A^-}\right] > 0.9$$

Exercise: data_lecture_14_SARS_COV2_serology.csv


Articles

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

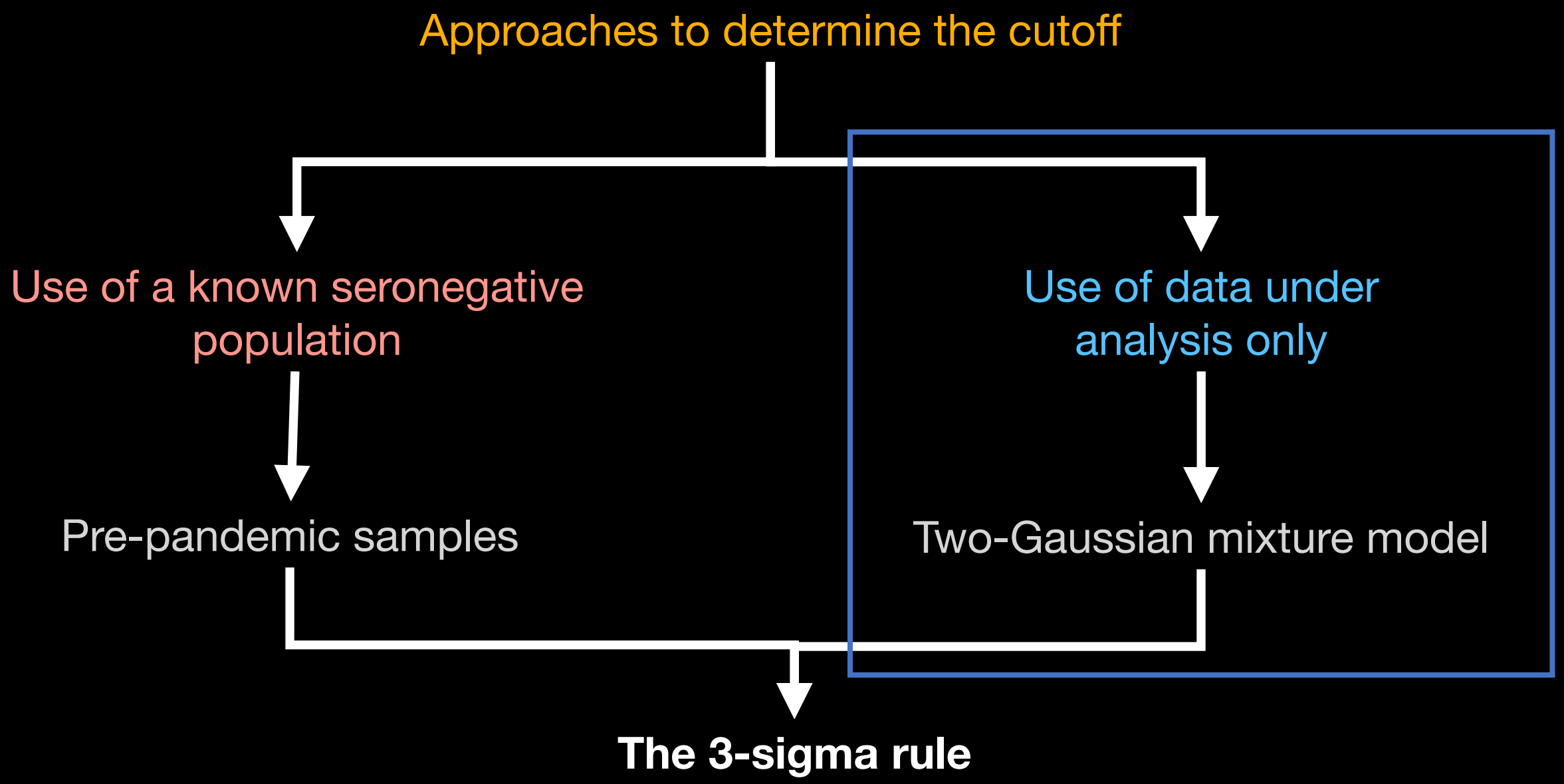
Jason Rosado, Stéphane Pelleau, Charlotte Cockram, Sarah Hélène Merkling, Narimane Nekkab, Caroline Demeret, Annalisa Meola, Solen Kerneis, Benjamin Terrier, Samira Fafi-Kremer, Jerome de Seze, Timothée Bruel, François Dejardin, Stéphane Petres, Rhea Longley, Arnaud Fontanet, Marija Backovic, Ivo Mueller,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

Apply the 3s-rule to pre-pandemic samples (status=negative) to calculate the cut-off for seropositivity of the anti-Spike-protein antibodies (Spike_IPP_IgG_MFI).

Calculate the proportion of these samples are above the threshold and check if this proportion agrees with the Cantelli-Chebyshev inequality.

Is the Normal distribution a reasonable distribution for the samples of SARS-CoV2-infected individual?

Apply this cutoff to calculate seroprevalence in SARS-CoV2-infected individual (status=positive).


Approaches to determine the cutoff
Use of a known seronegative population
Use of data under analysis only
Pre-pandemic samples
Two-Gaussian mixture model
The 3-sigma rule

Gaussian mixture models

$$f_X(x) = \sum_{i=1}^{k} \pi_i f_{N(\mu_i,\sigma_i)}(x)$$

where $\sum_{i=1}^{k} \pi_i = 1$

The most common model $\rightarrow k = 2$

$$f_X(x) = (1-\pi) f_{N(\mu_{S^-},\sigma_{S^-})}(x) + \pi f_{N(\mu_{S^+},\sigma_{S^+})}(x)$$

Definition of $S^- \Rightarrow \mu_{S^-} < \mu_{S^-}$

Estimation of the model

EM (Expectation-Maximization) Algorithm

1. Start with initial estimates for the parameters
2. E-Step - calculate the probability of each individual belonging to a given subpopulation according to estimates at 1.
3. M-Step - re-estimate the parameters using these probabilities and repeat the E-step with these new estimates
4. Stop with the increment in the log-likelihood is below a given tolerance error.

Package mixtools

Estimation of the model

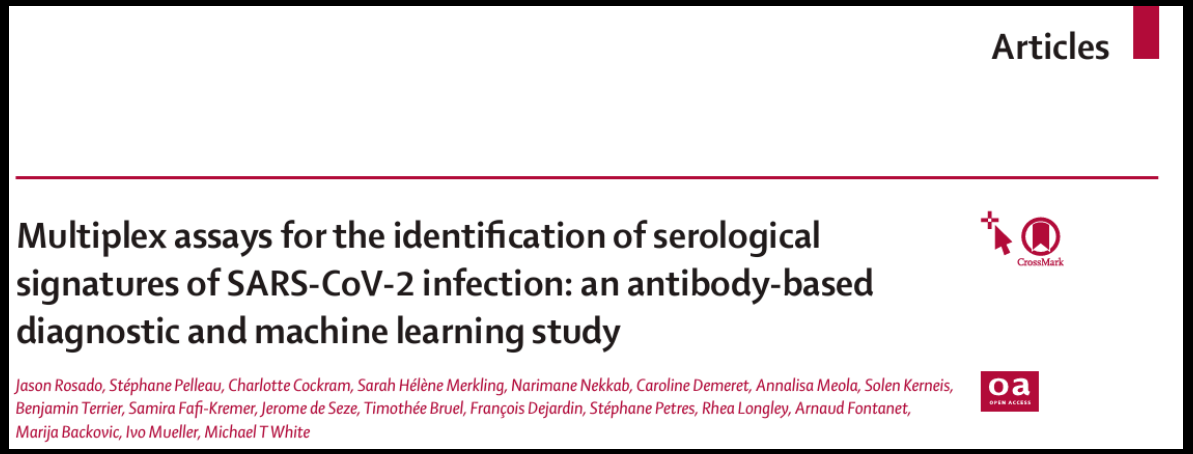
EM (Expectation-Maximization) Algorithm

1. Start with initial estimates for the parameters
2. E-Step - calculate the probability of each individual belonging to a given subpopulation according to estimates at 1.
3. M-Step - re-estimate the parameters using these probabilities and repeat the E-step with these new estimates
4. Stop with the increment in the log-likelihood is below a given tolerance error.

Calculate the cutoff for seropositivity according to $\hat{\mu}_{S^-}$ and $\hat{\sigma}_{S^-}$

Package mixtools

Exercise: data_lecture_14_SARS_COV2_serology.csv


Articles

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

Use the normalmixEM from the mixtools package to estimate a two-Gaussian mixture model to the data of anti-Spike-protein antibodies (Spike_IPP_IgG_MFI) from the SARS-CoV2-infected individual (status=positive).

Apply the 3s-rule to calculate the respective cut-off for seropositivity of the anti-Spike-protein antibodies (Spike_IPP_IgG_MFI).

Apply this cutoff to estimate the seroprevalence in these individuals.