Biostatistics

Other topics

Nuno Sepúlveda, 22.01.2024

Syllabus

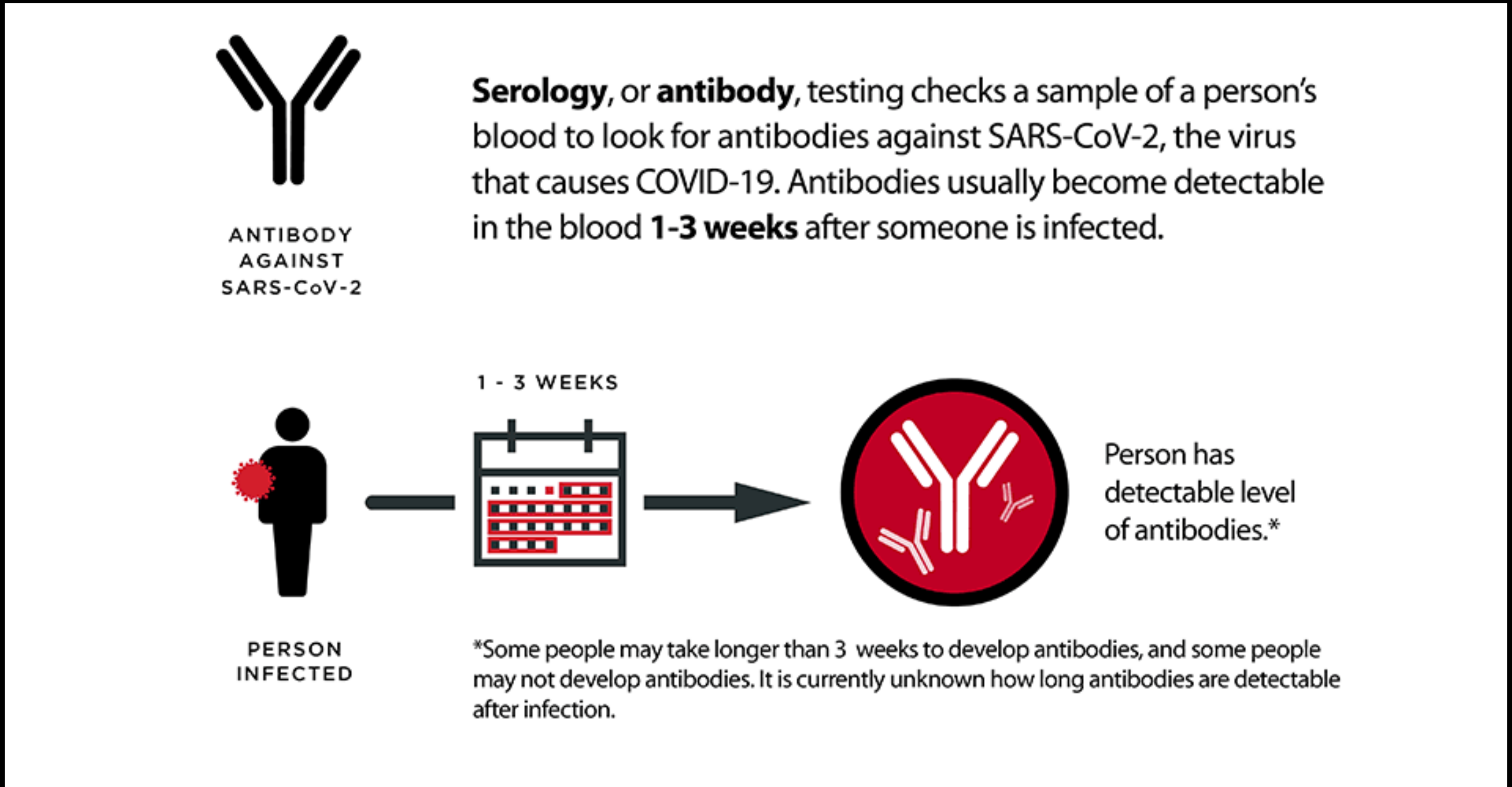
1. **General review**
   a. **What is Biostatistics?**
   b. **Population/Sample/Sample size**
   c. **Type of Data – quantitative and qualitative variables**
   d. **Common probability distributions**
   e. **Work example – Malaria in Tanzania**

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**
   c. **Advanced sample size and power calculations**

Serology

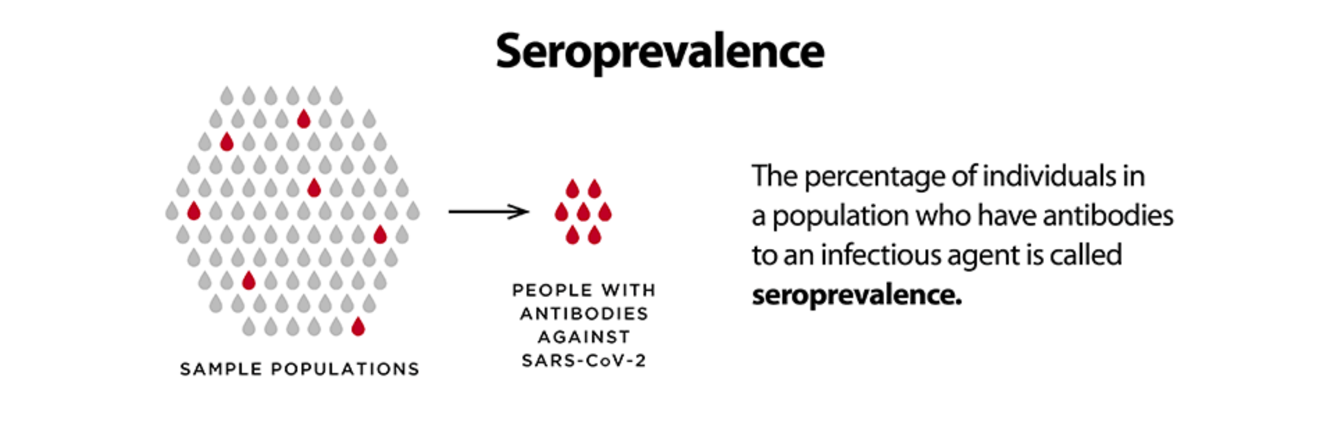


Serology, or **antibody**, testing checks a sample of a person's blood to look for antibodies against SARS-CoV-2, the virus that causes COVID-19. Antibodies usually become detectable in the blood **1-3 weeks** after someone is infected.

Person has detectable level of antibodies.*

*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.

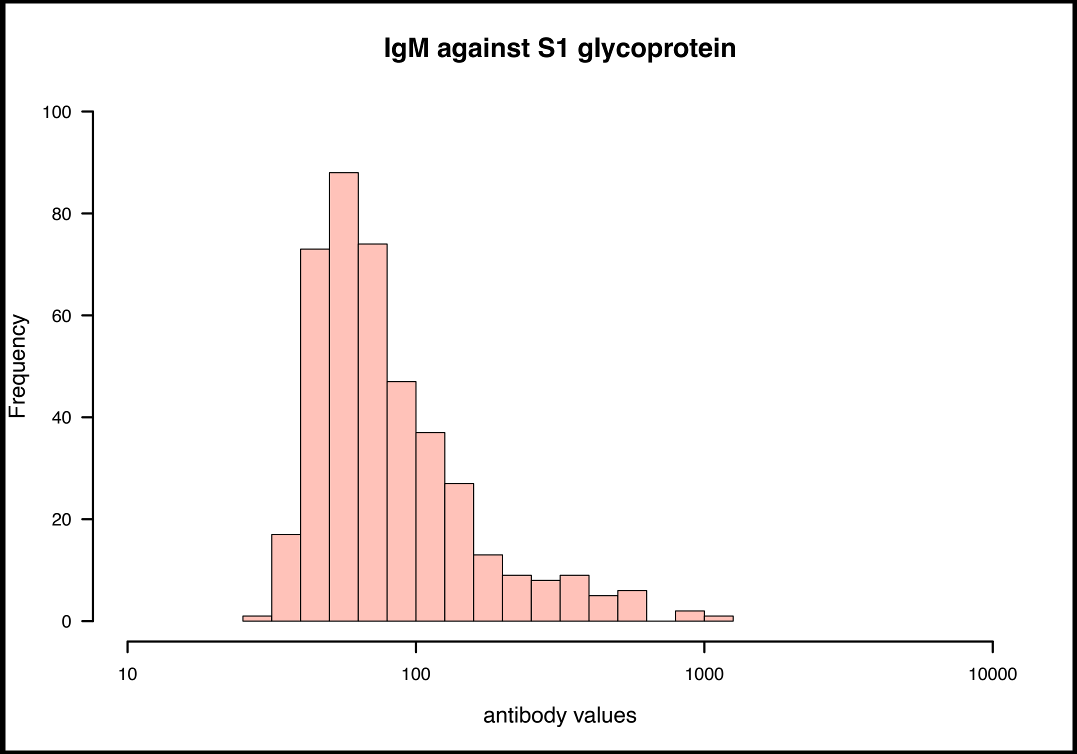
Sero-epidemiological surveys

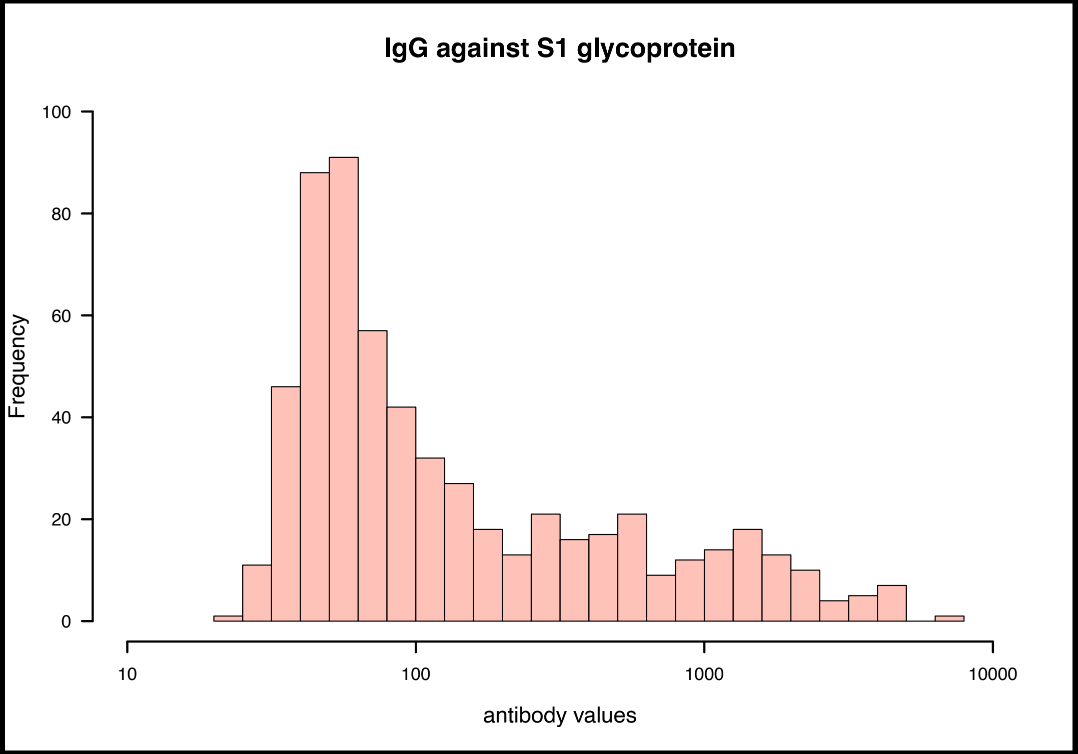


Seroprevalence

The percentage of individuals in a population who have antibodies to an infectious agent is called **seroprevalence.**

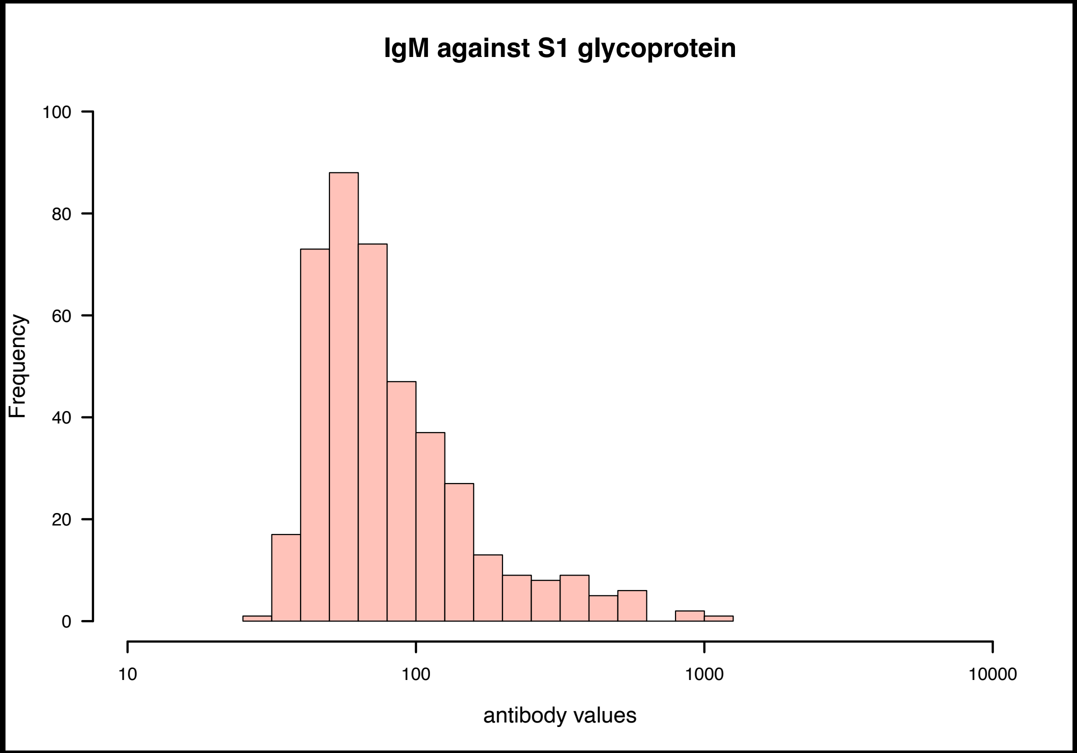
Antibody data are intrinsically quantitative

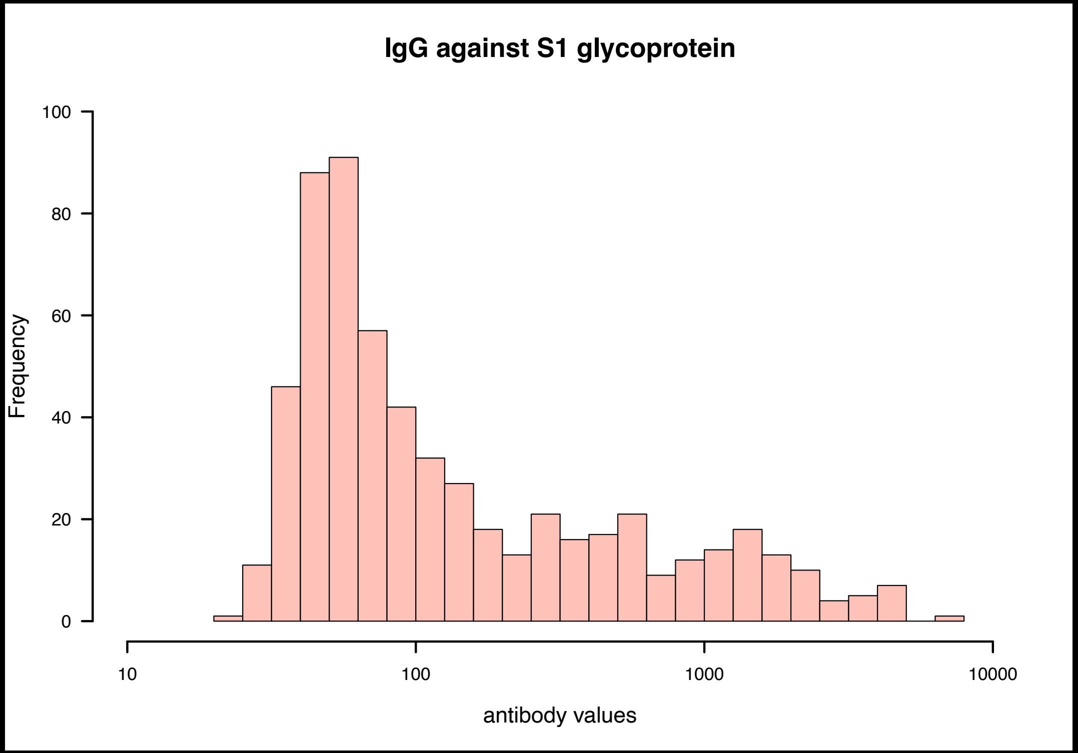




Rosado et al (2020). Serological signatures of SARS-CoV-2 infection: Implications for antibody-based diagnostics. medRxiv 2020.05.07.20093963.

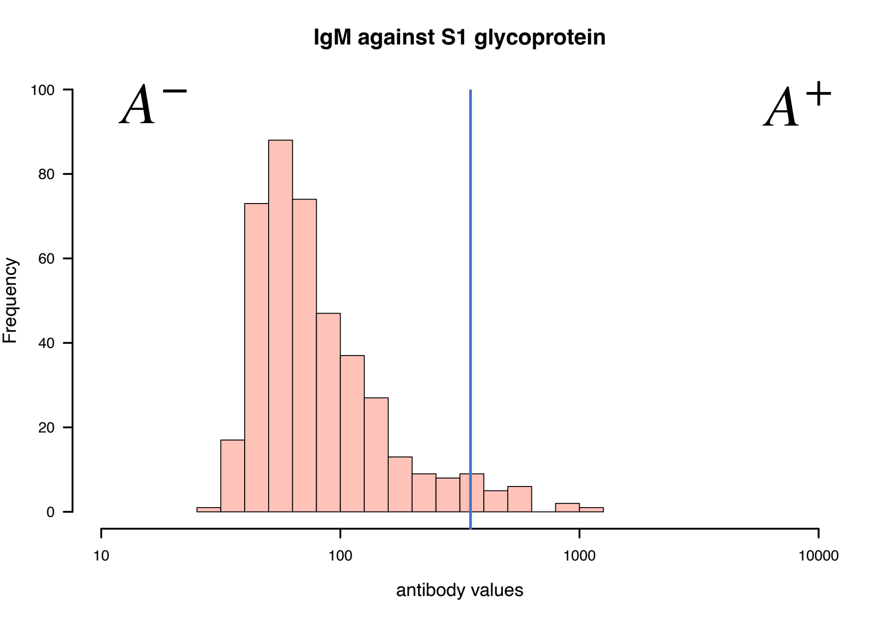
Who are the seropositive individuals?

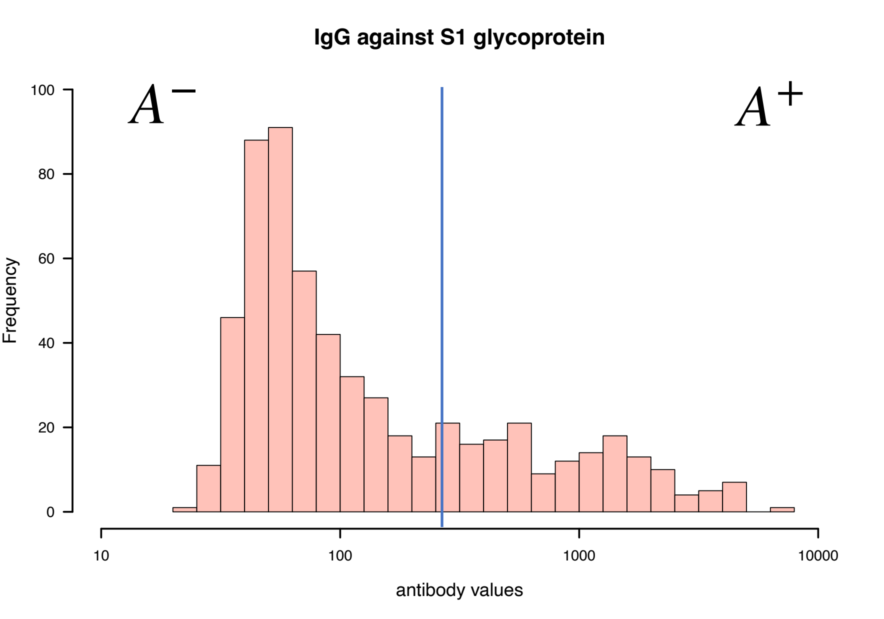


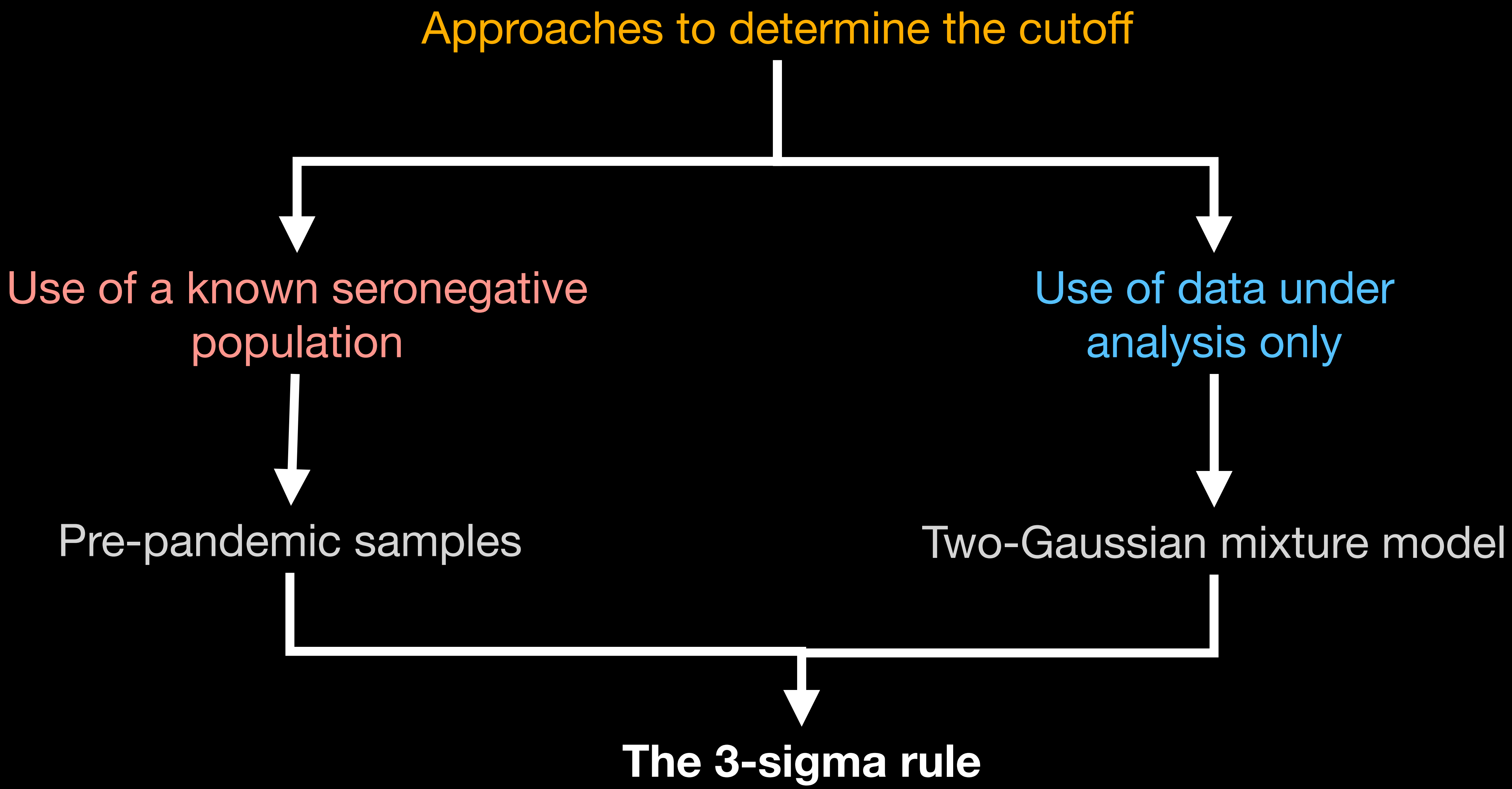


Rosado et al (2020). Serological signatures of SARS-CoV-2 infection: Implications for antibody-based diagnostics. medRxiv 2020.05.07.20093963.

How to determine the cut-off?







Approaches to determine the cutoff

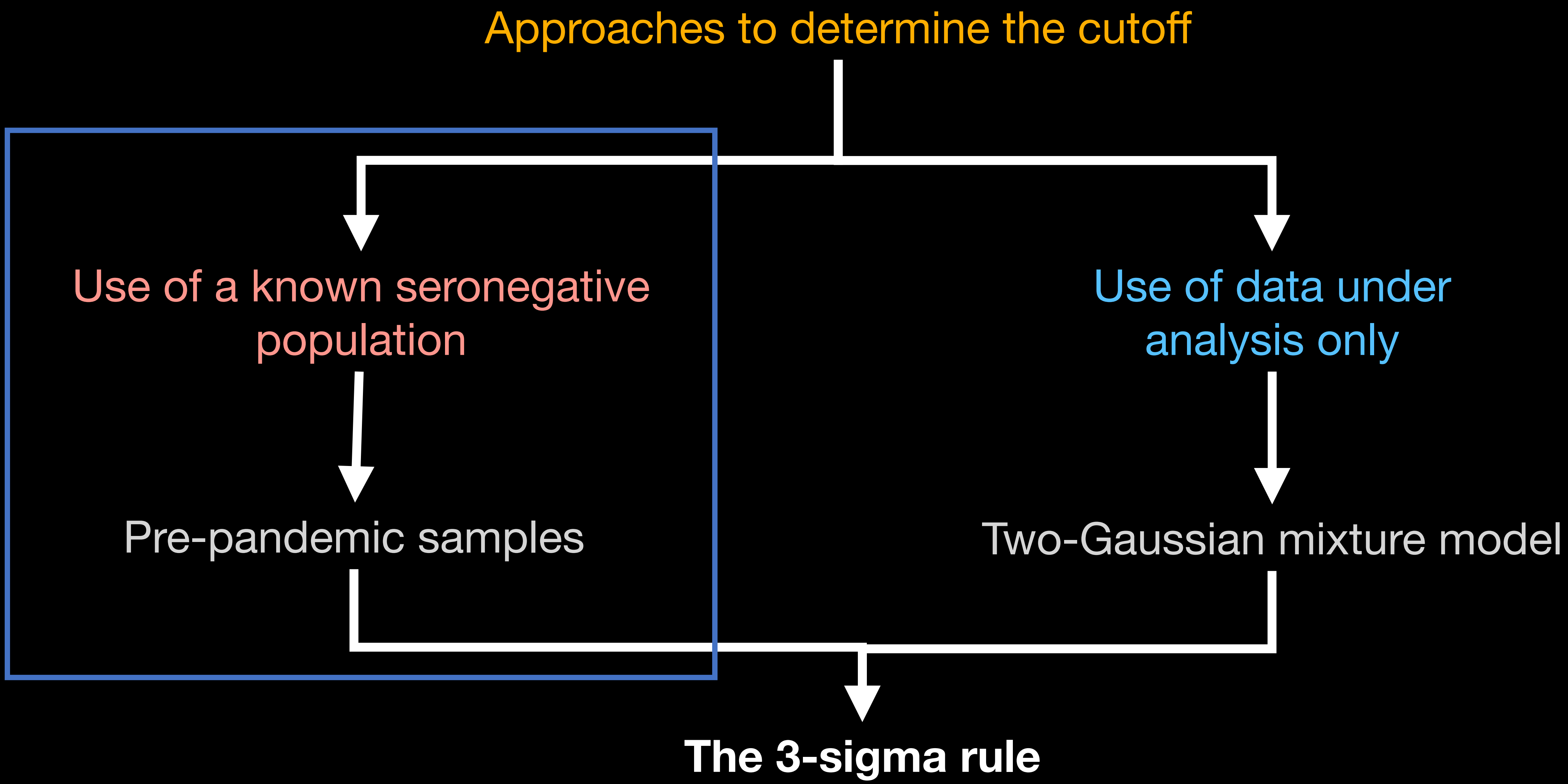
Use of a known seronegative population

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

Use of data under analysis only

Pre-pandemic samples

Two-Gaussian mixture model

The 3-sigma rule

The 3-sigma rule

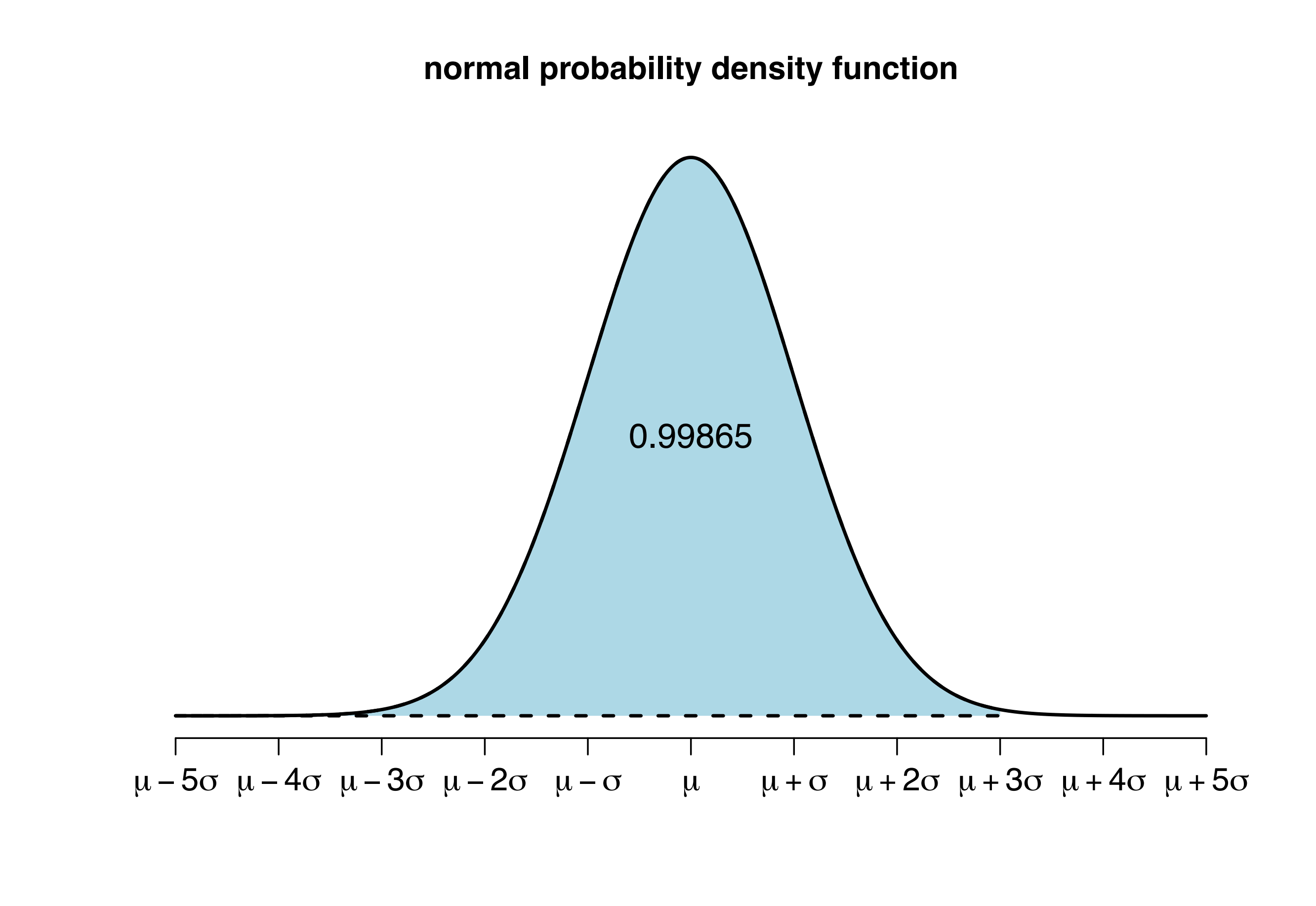
$$\mu_{A^-} = E\left[X \mid A^-\right]$$

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

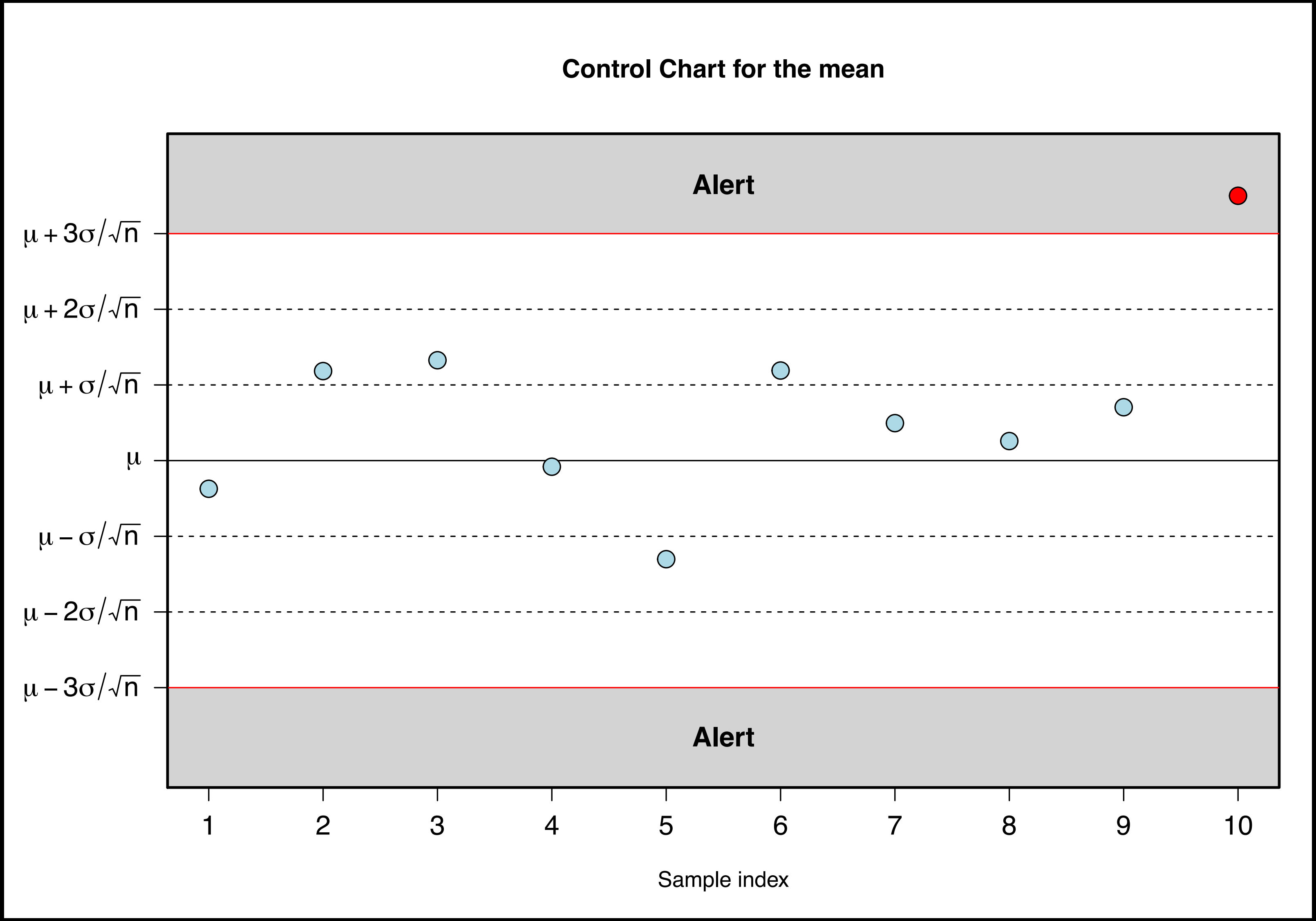
Seronegative, if $X_i \leq \mu_{A^-} + 3\sigma_{A^-}$

Seropositive, otherwise

The link to the Normal distribution



Quality control (Shewhart)



In practice (known seronegative population)

$\mu_{A-} \to \bar{X}_{A-}$

$\sigma_{A-} \to S_{A-}$

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

Seropositive, otherwise

Theoretical property of the 3-sigma

Cantelli-Chebyshev inequality

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

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

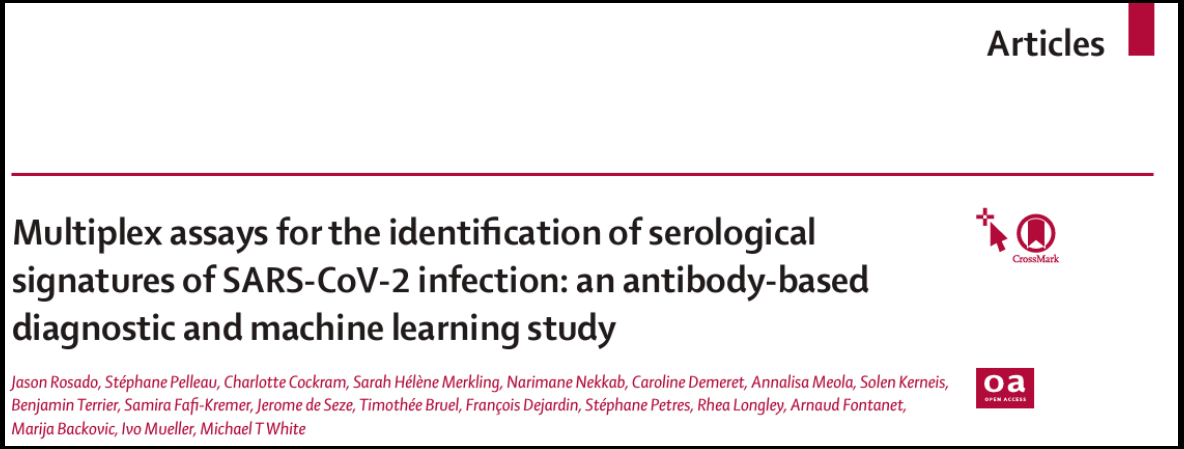
Application to $\lambda = 3\sigma$

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

↓

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

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

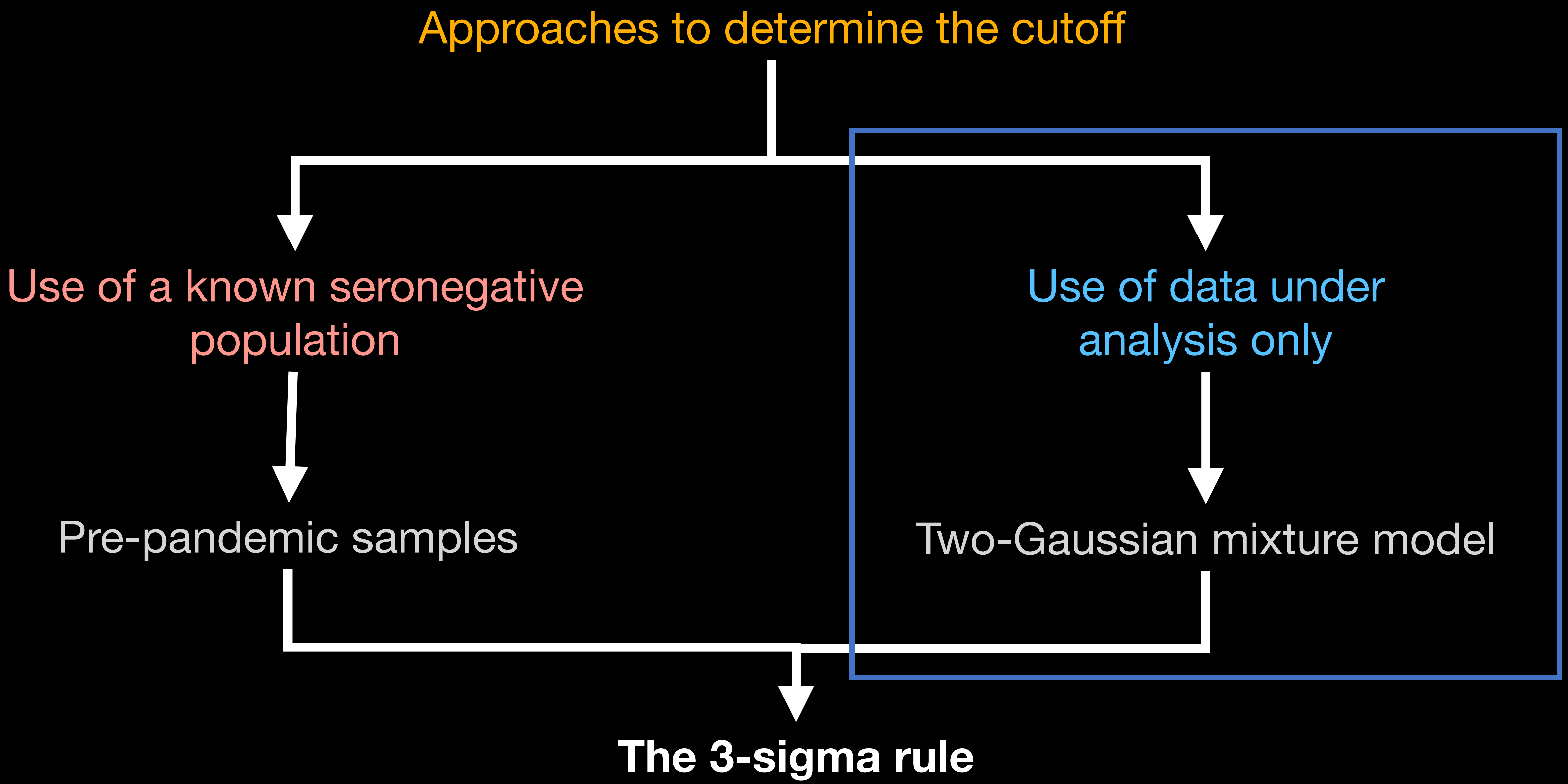
Jason Rosado, Stéphane Pelleau, Charlotte Cockram, Sarah Hélène Merkling, Narimane Nekkab, Caroline Demeret, Annalisa Meola, Solen Kerneis, Benjamin Terrier, Samira Fafi-Kremer, Jerome de Seze, Timothée Bruel, François Dejardin, Stéphane Petres, Rhea Longley, Arnaud Fontanet, Marija Backovic, Ivo Mueller,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