Lecture 4
Visualizing & standardizing serological data

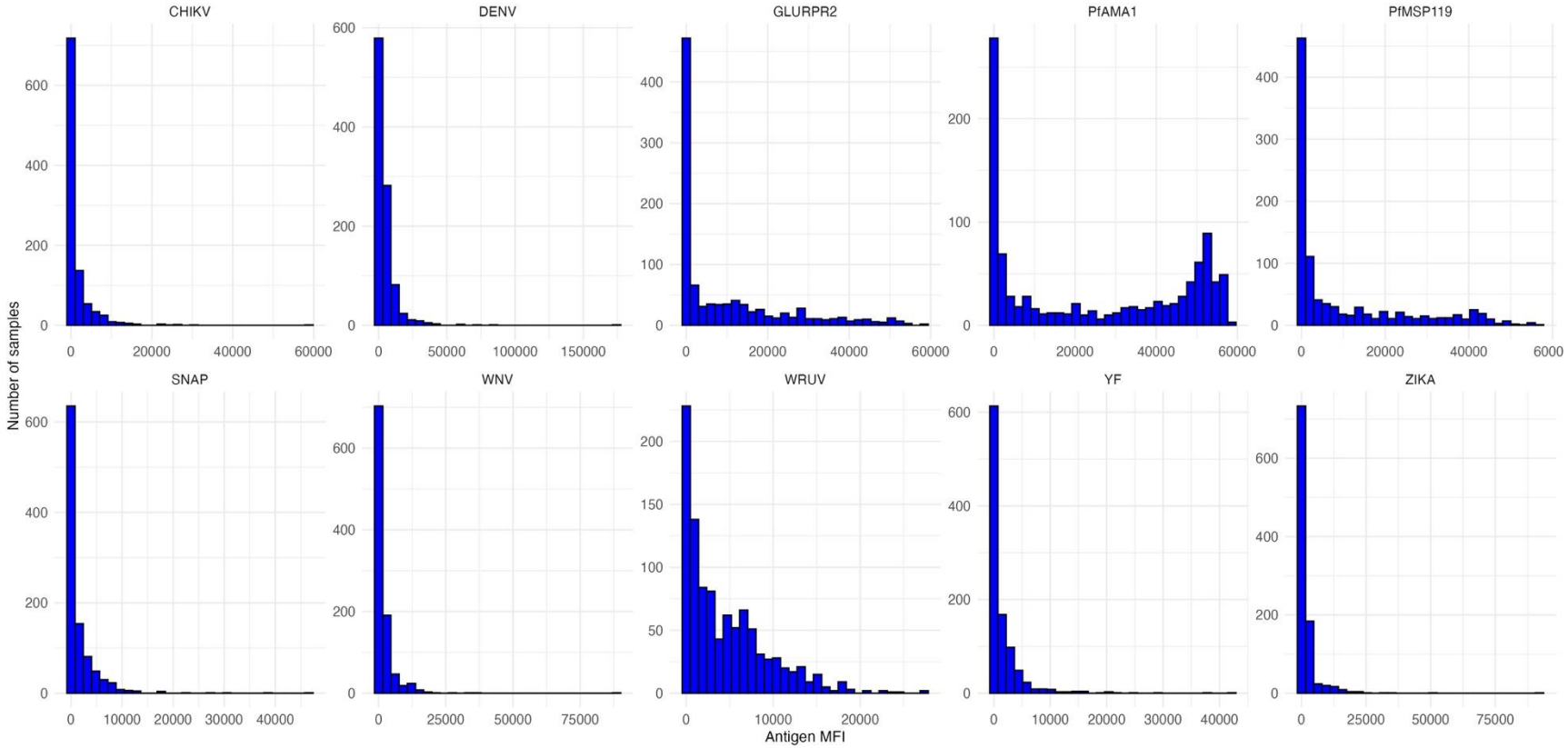
May 22, 2025

Seroanalytics Training
Blantyre, Malawi

Lecture outline

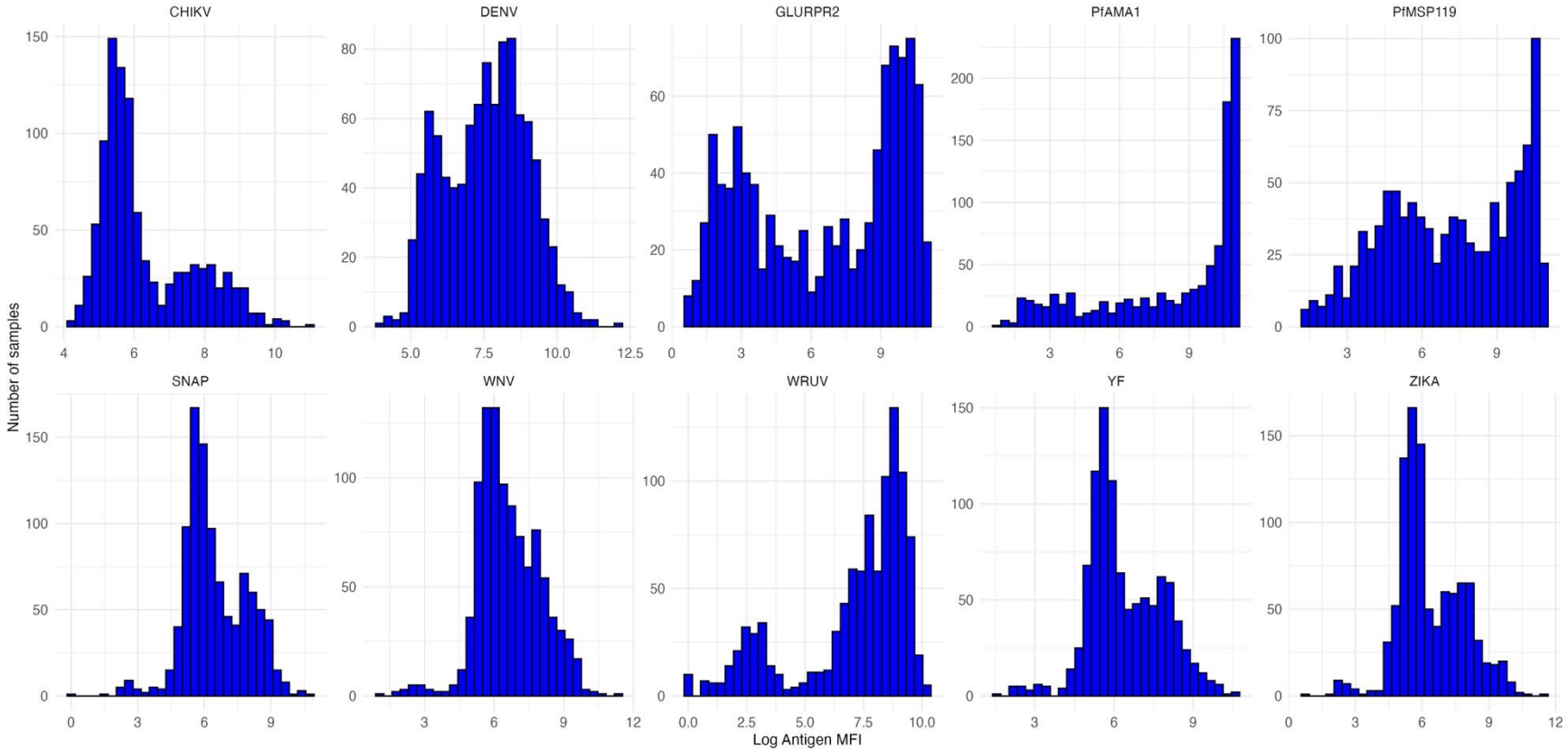
- Data visualization
- Why binarize serological data?
- Calculating seroprevalence from a cutoff
- Selecting controls

Visualizing serological data with histograms



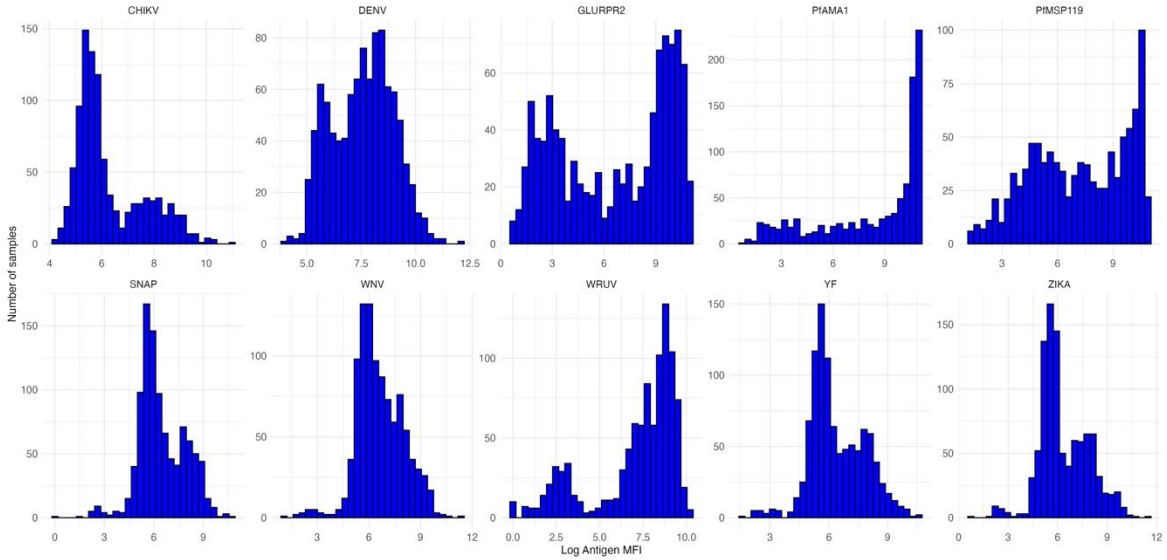
How would you compare the different distributions of data (untransformed)?

Visualizing serological data with histograms



How would you compare the different distributions of data (log transformed)?

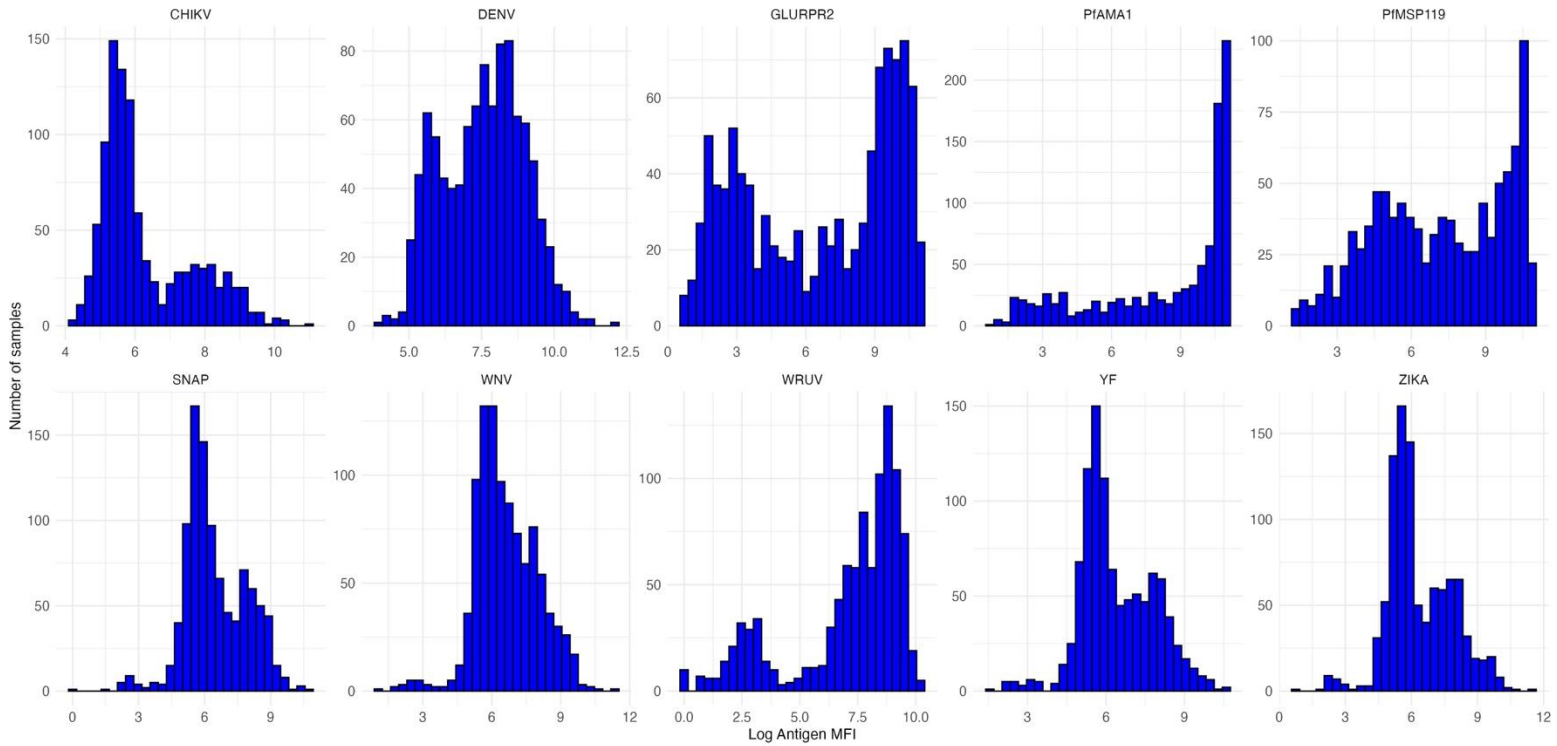
Visualizing serological data with histograms



How would you compare the different distributions of data?

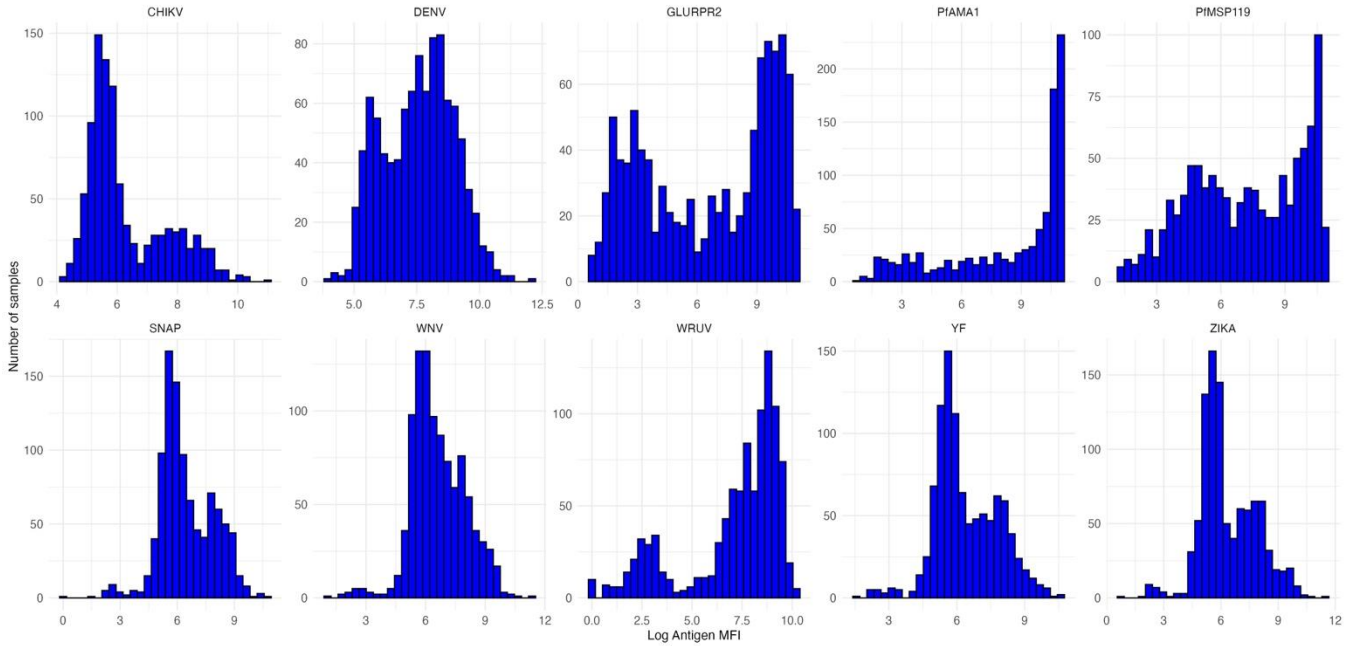
- Skewed vs. non-skewed distributions
- Unimodal vs. bimodal vs. multimodal
- Narrow or wide distribution
- Are there outliers?

Visualizing serological data with histograms



What underlying differences might cause the different distributions of data?

Visualizing serological data with histograms

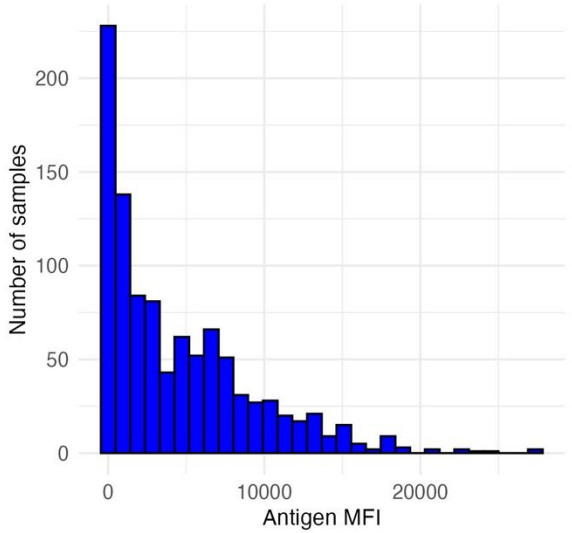


What underlying differences might cause the different distributions of data?

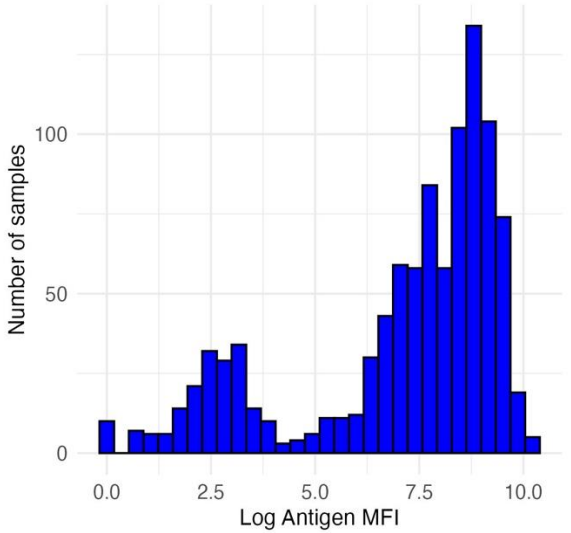
- Mix of exposed and unexposed in population
- Mix of vaccinated / unvaccinated
- Waning antibody responses

Histograms of rubella antibody responses

Untransformed MFI Values



Log transformed MFI values

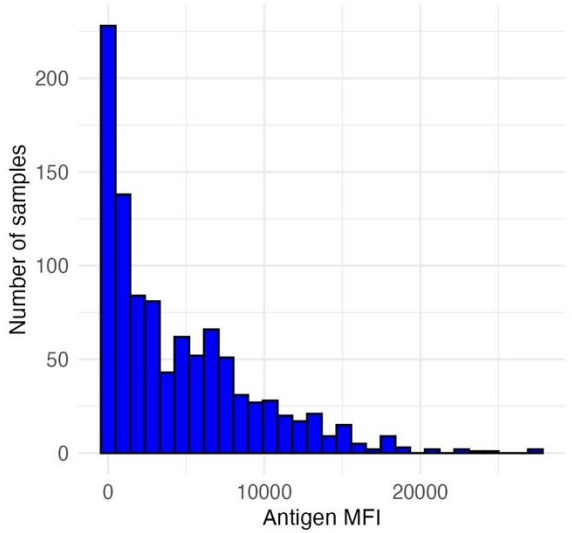


How do we get information from histograms like these?

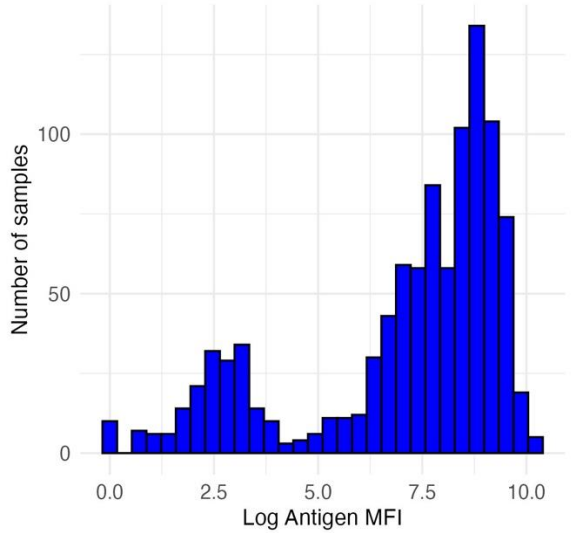
- How do we compare the results of different histograms?
- What inferences can we make about pathogen exposure based on a histogram?
- A method of making inferences from distributions is **binarizing** data
  - Setting a cutoff, and everything above that cutoff is positive

Histograms of rubella antibody responses

Untransformed MFI Values



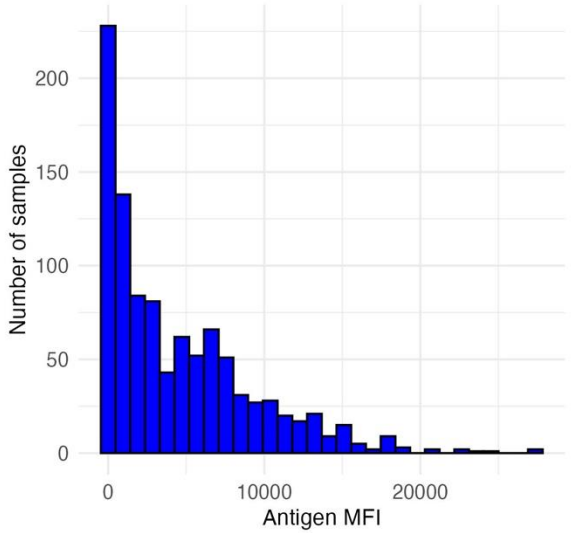
Log transformed MFI values



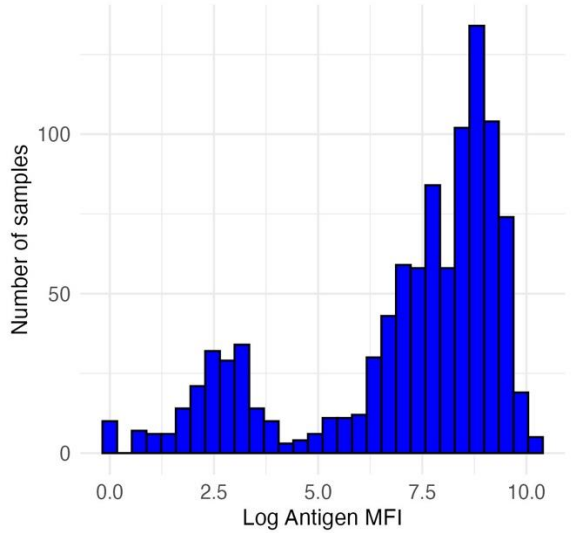
- Rubella has a **correlate of protection**
  - Individuals with antibody responses above a certain cutoff are expected to be protected from rubella infection
  - By applying a cutoff equivalent to the correlate of protection, we can calculate seroprevalence (here, also equals the immune proportion)

Histograms of rubella antibody responses

Untransformed MFI Values



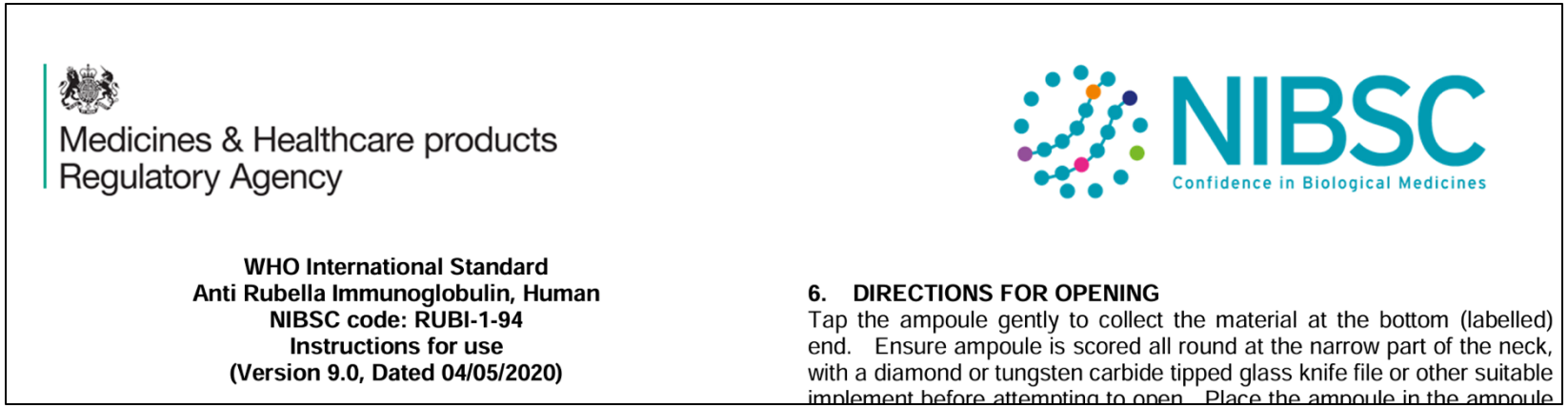
Log transformed MFI values



- Rubella correlate of protection is 9.36 IU/ml (international units per milliliter)
- However, the Luminex assay gives us antibody units in MFI (median fluorescence intensity)
- **We can convert between MFI and IU/ml using a standard curve:** we convert all MFI values to IU/ml, and apply the 9.36 IU/ml cutoff

Getting MFI values from standard curves

- The standard curve is created using **known concentrations** of antibody standards (e.g., the WHO International Standard for rubella, below), ideally measured on the same plate(s) as the unknown samples.



Medicines & Healthcare products Regulatory Agency

NIBSC
Confidence in Biological Medicines

WHO International Standard
Anti Rubella Immunoglobulin, Human
NIBSC code: RUBI-1-94
Instructions for use
(Version 9.0, Dated 04/05/2020)

6. DIRECTIONS FOR OPENING

Tap the ampoule gently to collect the material at the bottom (labelled) end. Ensure ampoule is scored all round at the narrow part of the neck, with a diamond or tungsten carbide tipped glass knife file or other suitable implement before attempting to open. Place the ampoule in the ampoule

https://nibsc.org/documents/ifu/RUBI-1-94.pdf

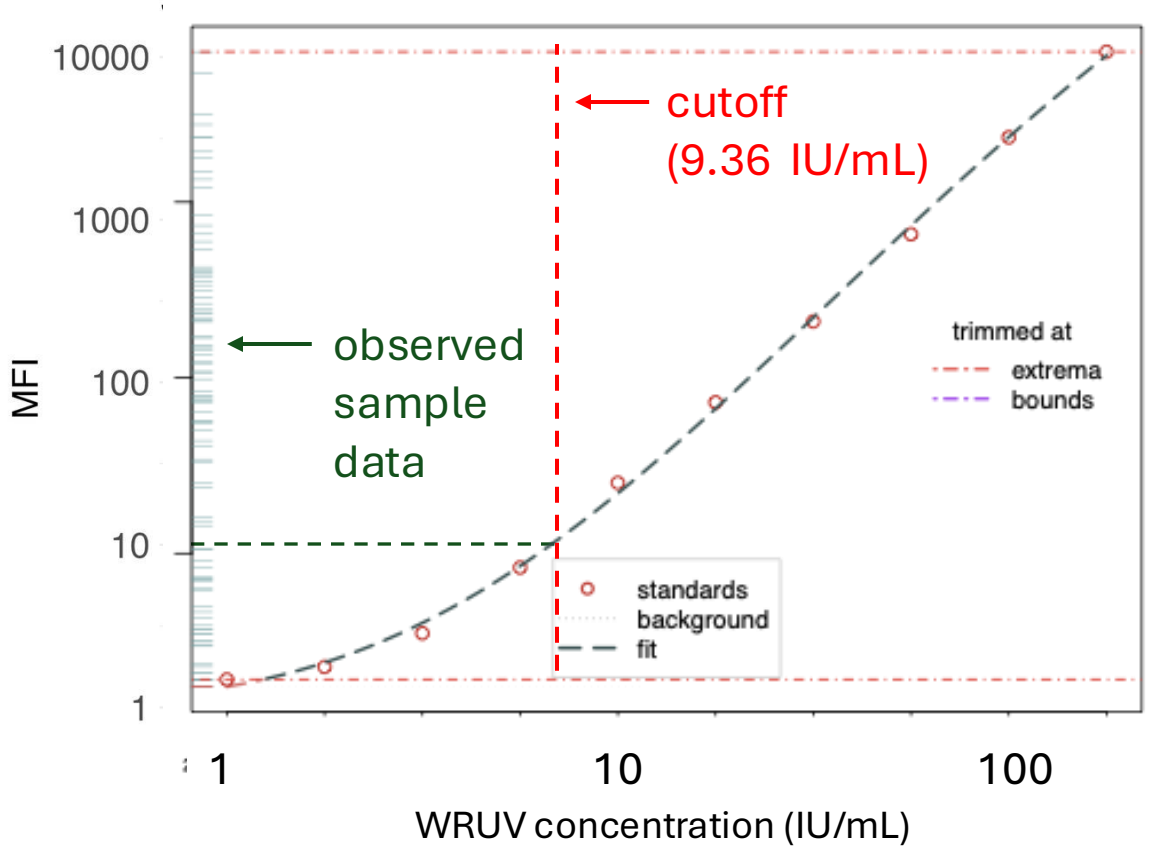
Converting MFI values to standardized units (IU/ml) from standard curves

(Recall we used standard curves to estimate relative antibody units.)

1. First, we fit a model to our standard curve data to establish the relationship between **MFI values** and known **IU/ml concentrations**.
2. Then, we can use this model to convert MFI values for the samples of interest to IU/ml.
3. Lastly, we apply the 9.36 IU/ml correlate of protection cutoff to the samples of interest to determine a binary serostatus for each sample.

Converting MFI values to standardized units (IU/ml) from standard curves

For rubella (the “WRUV” antigen), we can apply the cutoff (based on correlate of protection) of 9.36 IU/ml to determine individual serostatus.



Determining serostatus from a cutoff

For each sample in the dataset, we convert the MFI value to a concentration value (IU/mL), and then classify each sample as seropositive (1) or seronegative (0)

Sample ID	WRUV MFI	WRUV IU/mL	Serostatus
1	MFI_1	conc_1	1
2	MFI_2	conc_2	1
3	MFI_3	conc_3	0
4	MFI_4	conc_4	1
5	MFI_5	conc_5	1
6	MFI_6	conc_6	0
7	MFI_7	conc_7	0

Calculating seropositivity *or seroprevalence*

For WRUV, with a cutoff of 9.36 IU/ml, we have:

- 767 seropositive samples
- 233 seronegative samples

$$Seropositivity\ (\%\ Seropositive) = \frac{Number\ seropositive}{Total\ number\ tested} = \frac{767}{1000} = 76.7\%$$

Calculating seropositivity – confidence intervals

- Estimating uncertainty is important for capturing true population seroprevalence; it typically accounts for variance around a mean estimate.
- Estimating a 95% confidence interval means that in 95 out of 100 calculations of seroprevalence from the same source population using 1000 individuals, the range will include the true population-level seroprevalence.
- This method uses a binomial distribution and accounts for the number of people sampled – luckily, we can use R to easily compute it:

$$\Pr(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$$

With 767 seropositive individuals and 1000 individuals total, the 95% CI is: **74.0% - 79.3%.**

Calculating seropositivity – confidence intervals

With 767 seropositives and 1000 total samples:

Seroprevalence = 76.7% (95% CI: 74.0%, 79.3%)

Question: How do we interpret seroprevalence?

Interpretation strongly depends on how seroprevalence is determined, including controls used and selected cutoff.

For the seroprevalence of rubella (WRUV):

- Percentage of people who have been exposed to vaccine or natural infection.
- If we know there's no/little vaccination, we may assume seroprevalence is wholly due to natural infection, or vice versa.
- Since cutoff is a correlate as protection, seroprevalence could also indicate which individuals are susceptible to future infection, and whether there could be outbreaks in a region.

Question: How do we interpret seroprevalence?

How would we interpret seroprevalence if there is NOT a correlate of protection?

Question: How do we interpret seroprevalence?

How would we interpret seroprevalence if there is NOT a correlate of protection?

- For antigens in general, seroprevalence could indicate:
  - Population ever exposed to pathogen or vaccine
  - Population with recent infection
  - Population with symptomatic infection
  - Cross-reactive antibody responses

Controls

- If we don’t have a correlate of protection, we can use well-characterized control samples to determine cutoffs for seropositivity.
  - Ideally, we would have positive and negative controls.
  - Often, we only have negative controls, and sometimes we don’t have positive or negative controls.
- The controls we use to establish cutoffs can affect our interpretation of seroprevalence.

Non-target antigen controls

- We also have controls with non-target antigens (SNAP in our dataset; other common ones are GST and Vero cell).
- These controls can ensure that the plate ran correctly, whether values are similar between plates, and high values can indicate non-specific binding.
- The non-target antigen values should ideally not differ between samples with high and low antigen values.

ID	SNAP MFI
1	80
2	74
3	214
4	1495
5	226
6	189
7	2211
8	228
9	67
10	915
11	50

Positive controls

- Often, positive controls are based on antibody responses from people who are known to have had symptomatic disease.
- **How might positive controls differ from positives in samples?**

Positive controls

- Often, positive controls are based on antibody responses from people who are known to have had symptomatic disease.
- **How might positive controls differ from positives in samples?**
    - Vaccination vs. natural infection
    - Intensity of infection – severe, symptomatic, asymptomatic – (controls may be more likely to have clinical or severe infection)
    - Timing since infection (controls likely were taken from acute infection phase)

Negative controls

- **What populations might be best to get negative controls from?**

Negative controls

- **What populations might be best to get negative controls from?**
  - Non-endemic area (likely to be adults)
  - Very young presumed unexposed (likely to be from target population)

Negative controls

- **What populations might be best to get negative controls from?**
  - Non-endemic area (likely to be adults)
  - Very young presumed unexposed (likely to be from target population)
- **Why would you choose to get - or not get - controls from each of these populations?**

Negative controls

- **What populations might be best to get negative controls from?**
  - Non-endemic area (likely to be adults)
  - Very young presumed unexposed (likely to be from target population)
- **Why would you choose to get - or not get - controls from each of these populations?**
  - Controls from non-endemic area may differ from target population in many ways (geographic, socioeconomic, disease history)
  - Very young immune systems vs. adult immune systems

Controls

You may not have ideal controls.

You should understand the assumptions/biases of your controls!

Conclusions

- You should visualize your data before conducting analysis.
- It can be useful to classify samples as seropositive and seronegative, and to calculate seropositivity.
- It’s important to understand the controls you use, and think about how these may affect your inference.