Inferring Transmission Dynamics from serological data

2025-04-28

1. Set up our environment

```
# clear everything in the environment
rm(list=ls())

# load packages
library(ggplot2)
library(epitools)
```

2. Read in our data

You will need to change the "path" variable to the location of the files on your computer. In this lab we will focus on the DENV and CHIKV data which have already been classified as seropositive (1) or seronegative (0).

```
# set my_path to be the working directory location of where the *seroanalytics_workshop* folder is stor
my_path <- "/Users/soniahegde/Library/CloudStorage/OneDrive-JohnsHopkins/seroanalytics_workshop"
df <- read.csv(paste(my_path, "Data/Lab_6_data.csv", sep="/"))

# lets take a look at our dataset
# in this lab we will focus on the dengue (DENV) and chikungunya (CHIKV) data
head(df)
```

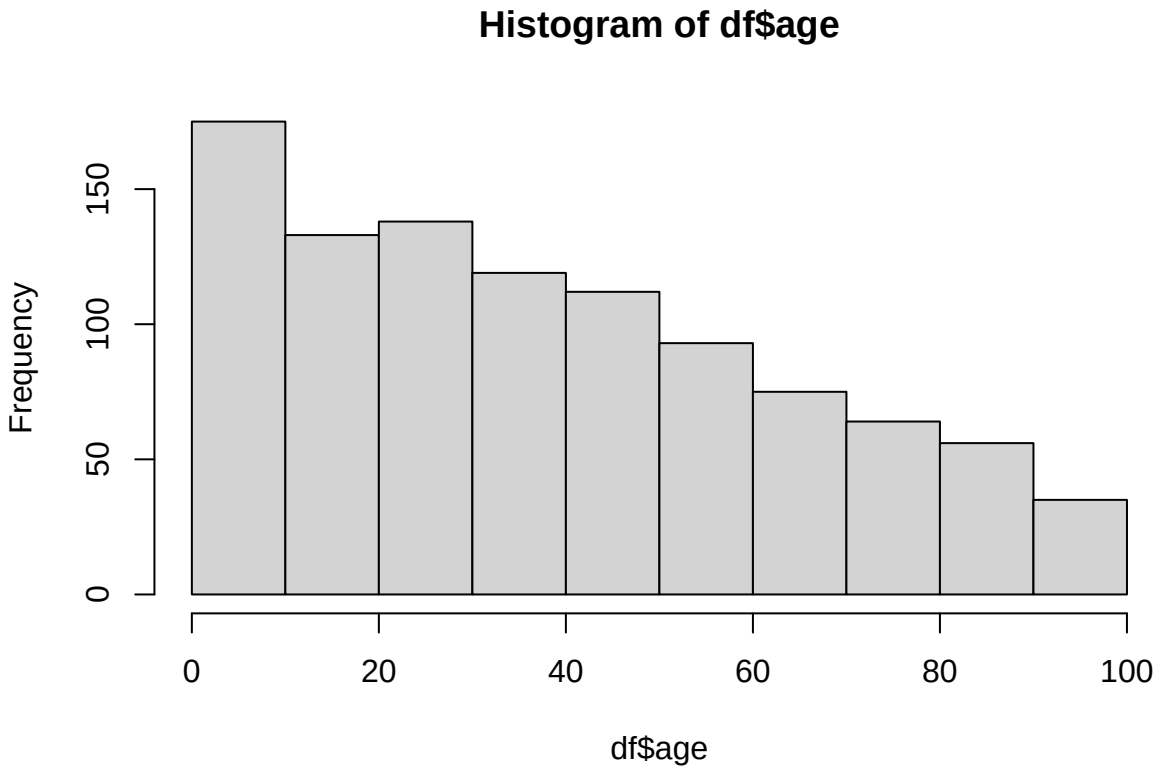
```
##   age DENV CHIKV   loc
## 1  58    1     1 rural
## 2  33    1     1 rural
## 3  49    1     0 urban
## 4  10    0     0 rural
## 5  98    1     1 rural
## 6  50    1     1 urban
```

3. Calculate age-specific seroprevalence

To investigate the past transmission dynamics of these pathogens we first want to look at age-specific trends in seroprevalence. Here, we will calculate age-specific seroprevalence, based on the individual-level serostatus that has already been defined.

Its important to check the age-distribution of the study population before deciding what age groups to use for analysis. This is to ensure that we choose age groups with a sufficient number of people. In general, the more age-groups we can use, the more information we can get about past transmission. However, if the number of study participants in a particular age group is too low, we will not be confident in the seroprevalence

estimate (i.e. confidence intervals will be wide). If the sample sizes within an age-group are too small, wider age groupings should be used.



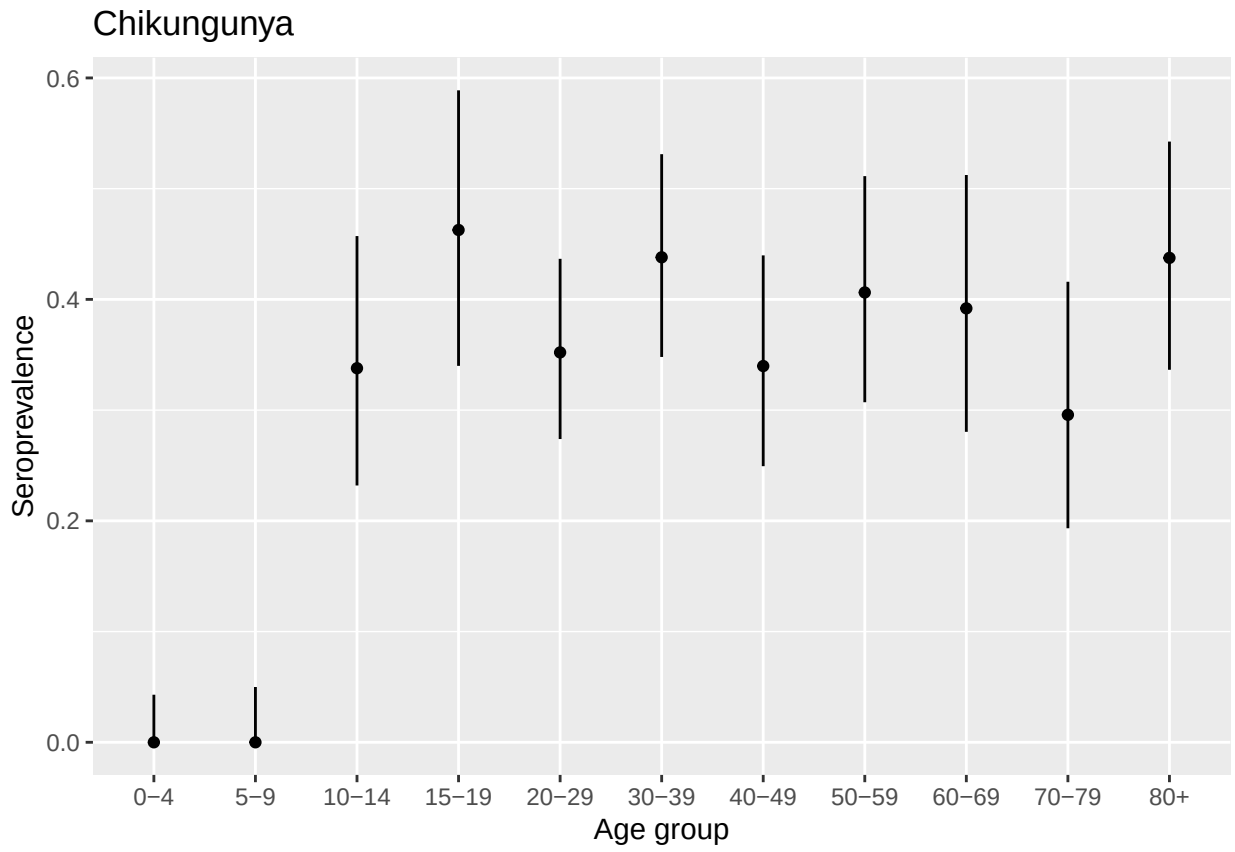

```
##
##   0-4   5-9 10-14 15-19 20-29 30-39 40-49 50-59 60-69 70-79   80+
##    84    72    74    67   142   121   103    96    74    71    96
```

4. Investigate transmission dynamics

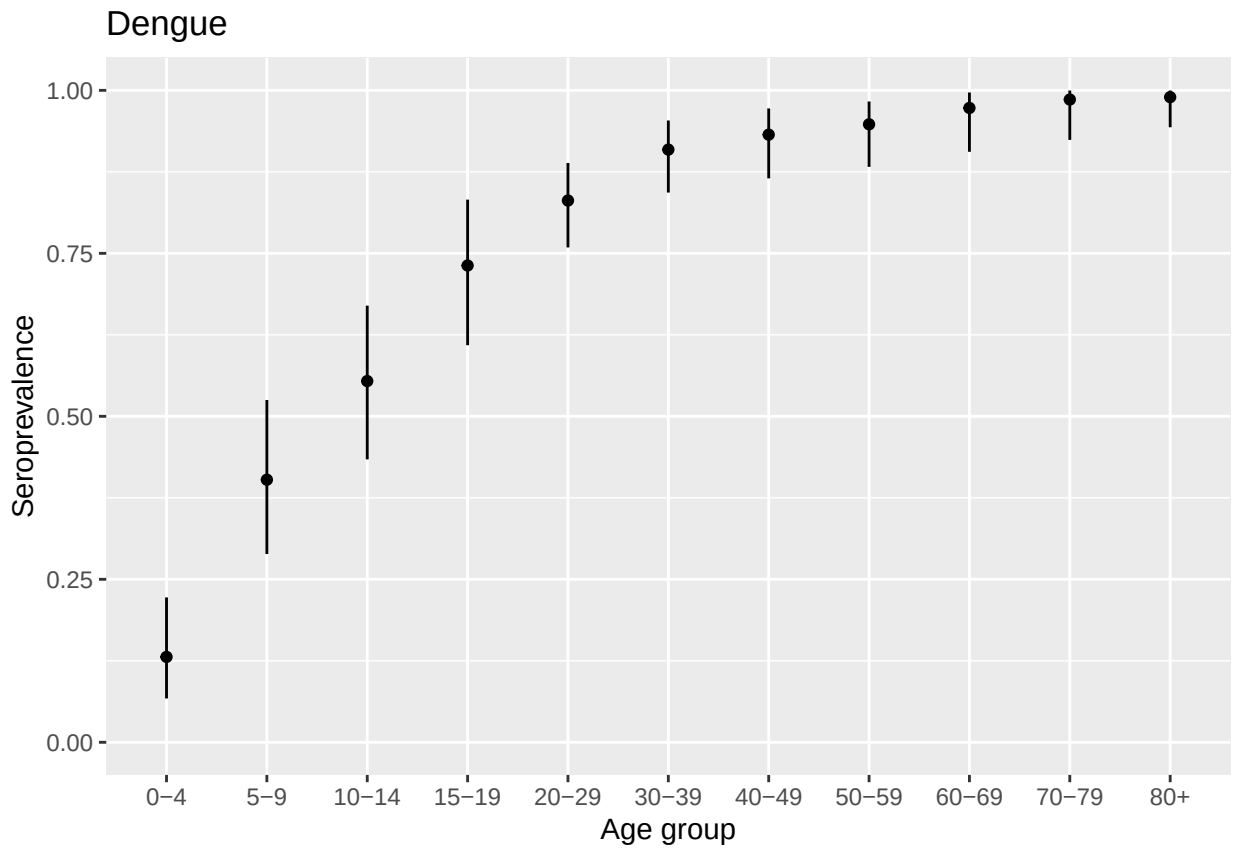
Now that we have our age-specific seroprevalence estimates, we can visualize these trends for each pathogen and answer the questions below:

- Question 1: What do we think has happened with CHIKV transmission here?
- Question 2: What kind of transmission pattern do we see for DENV?

```
# plot age-specific CHIKV seroprevalence
ggplot(seroprev$CHIKV, aes(age_group, prev))+ geom_point()+ ggtitle("Chikungunya")+
  geom_linerange(aes(ymin=ciL, ymax=ciU))+ ylab("Seroprevalence")+ xlab("Age group")
```




```
# plot age-specific DENV seroprevalence
ggplot(seroprev$DENV, aes(age_group, prev))+ geom_point()+ ggtitle("Dengue")+ ylim(0,1)+
  geom_linerange(aes(ymin=ciL, ymax=ciU))+ ylab("Seroprevalence")+ xlab("Age group")
```



5. Estimating FOI (force of infection)

We will now fit a catalytic model to the age-specific DENV seroprevalence data to estimate force of infection (FOI). We will do this using a generalized linear model (GLM) with a complementary log-log link. In this model we assume:

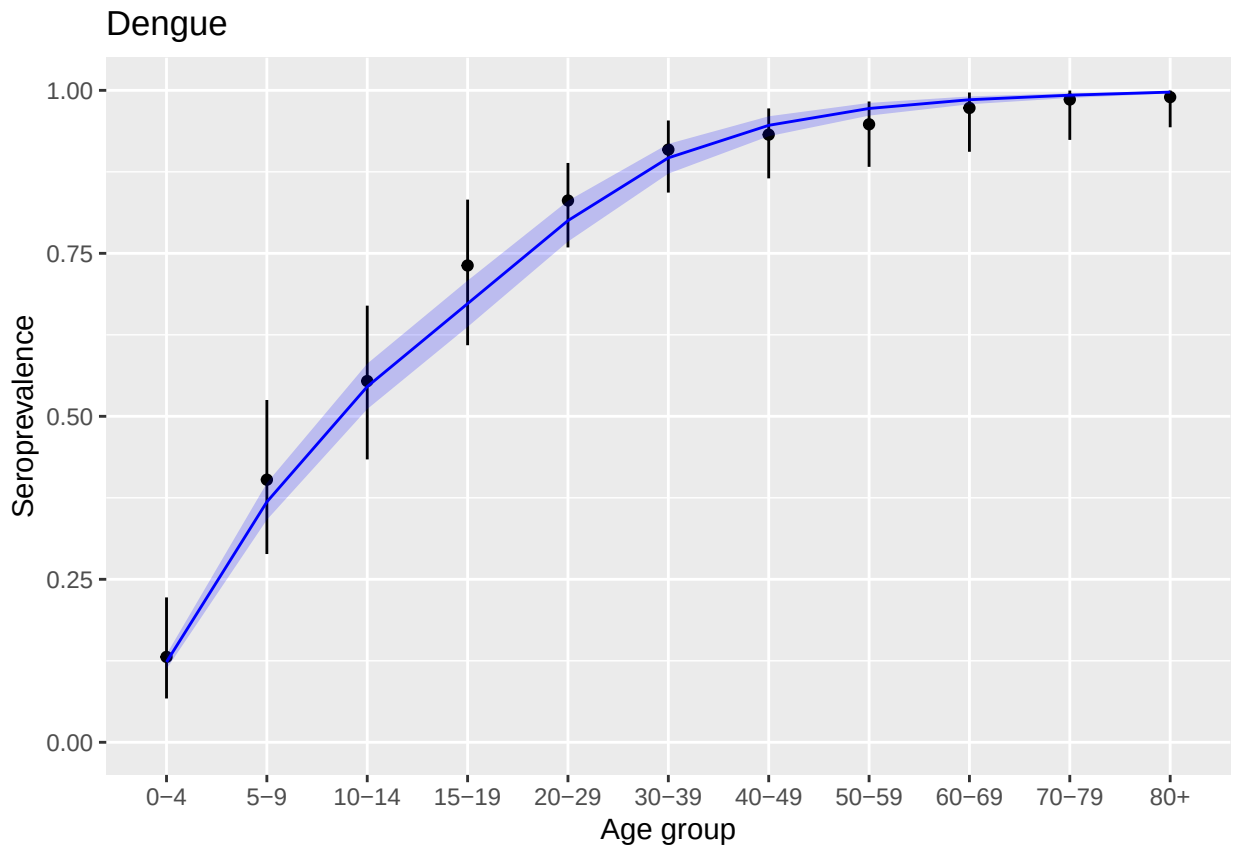
- that FOI is constant over time (i.e. the result can be interpreted as the long-term average annual FOI)
- that FOI does not vary with age (i.e. age does not impact the risk of infection)

```
# for this model we will use the mid-points of each age group
age_mid <- c(2, 7, 12, 17, 24.5, 34.5, 44.5, 54.5, 64.5, 74.5, 90)
seroprev$DENV$age_mid <- age_mid

# fit catalytic model
mod <- glm(cbind(npos, n-npos) ~ 1, offset=log(age_mid),
           data=seroprev$DENV, family = binomial(link = "cloglog"))

# model results
summary(mod)
```

```
##
## Call:
## glm(formula = cbind(npos, n - npos) ~ 1, family = binomial(link = "cloglog"),
##     data = seroprev$DENV, offset = log(age_mid))
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2926  -0.7154   0.1468   0.5303   1.0383
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.72228    0.04964  -54.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6.8663  on 10  degrees of freedom
## Residual deviance: 6.8663  on 10  degrees of freedom
## AIC: 49.57
##
## Number of Fisher Scoring iterations: 4
```

```r
# the model provides an estimate of the log of the FOI, requiring us to exponentiate the result
log_FOI <- mod$coefficients[1]
FOI <- exp(log_FOI)

# the model estimates a 6.6% force of infection

# lets calculate the confidence intervals around this estimate
se_log_FOI <- summary(mod)$coefficients[1,2]
ci_log_FOI <- log_FOI + c(-1.96, 1.96) * se_log_FOI
ci_FOI <- exp(ci_log_FOI)


# lets now see how well this estimate fits our data

# calculate predicted seroprevalence by age group
pred_seroprev <- data.frame(age_group=seroprev$DENV$age_group,
                            pred=1-exp(-FOI*age_mid),
                            ciL=1-exp(-ci_FOI[1]*age_mid),
                            ciU=1-exp(-ci_FOI[2]*age_mid))


# plot observed vs model estimated seroprevalence
ggplot(seroprev$DENV, aes(age_group, prev))+ geom_point()+ ggtitle("Dengue")+ ylim(0,1)+
  geom_linerange(aes(ymin=ciL, ymax=ciU))+ ylab("Seroprevalence")+ xlab("Age group")+
  geom_line(data=pred_seroprev, aes(age_group, pred, group=1), col="blue")+
  geom_ribbon(data=pred_seroprev, aes(x=age_group, y=pred, ymin=ciL, ymax=ciU, group=1), fill="blue", al
```




```
# we can see that the model estimate of FOI matches our observed data pretty well!
```

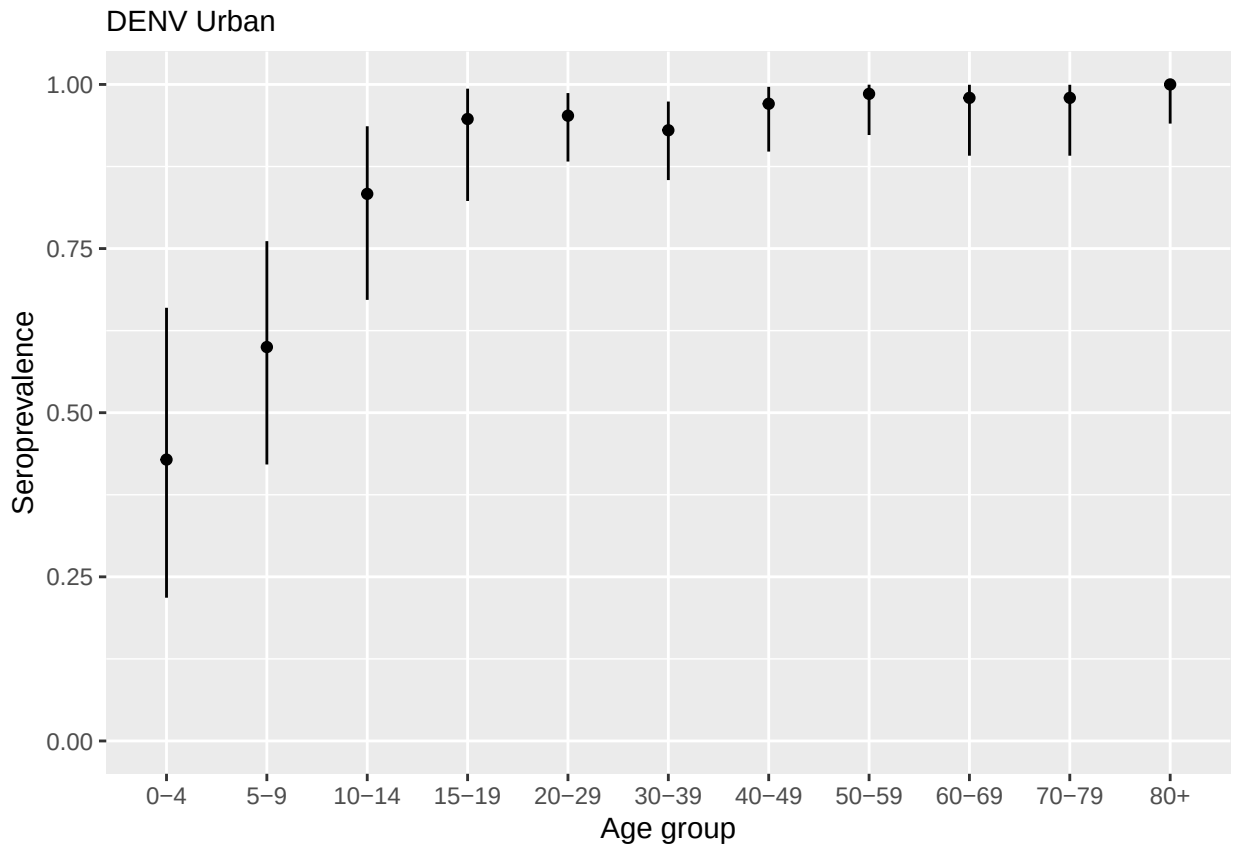
6. FOI in urban vs rural locations

In our dataset we know whether each study participant lives in a rural or urban location ("loc" variable in the data). Lets see if there are any differences in DENV FOI by location. We will apply the same code as before to calculate DENV age-specific seroprevalence (this time by urban/rural) and estimate FOI.

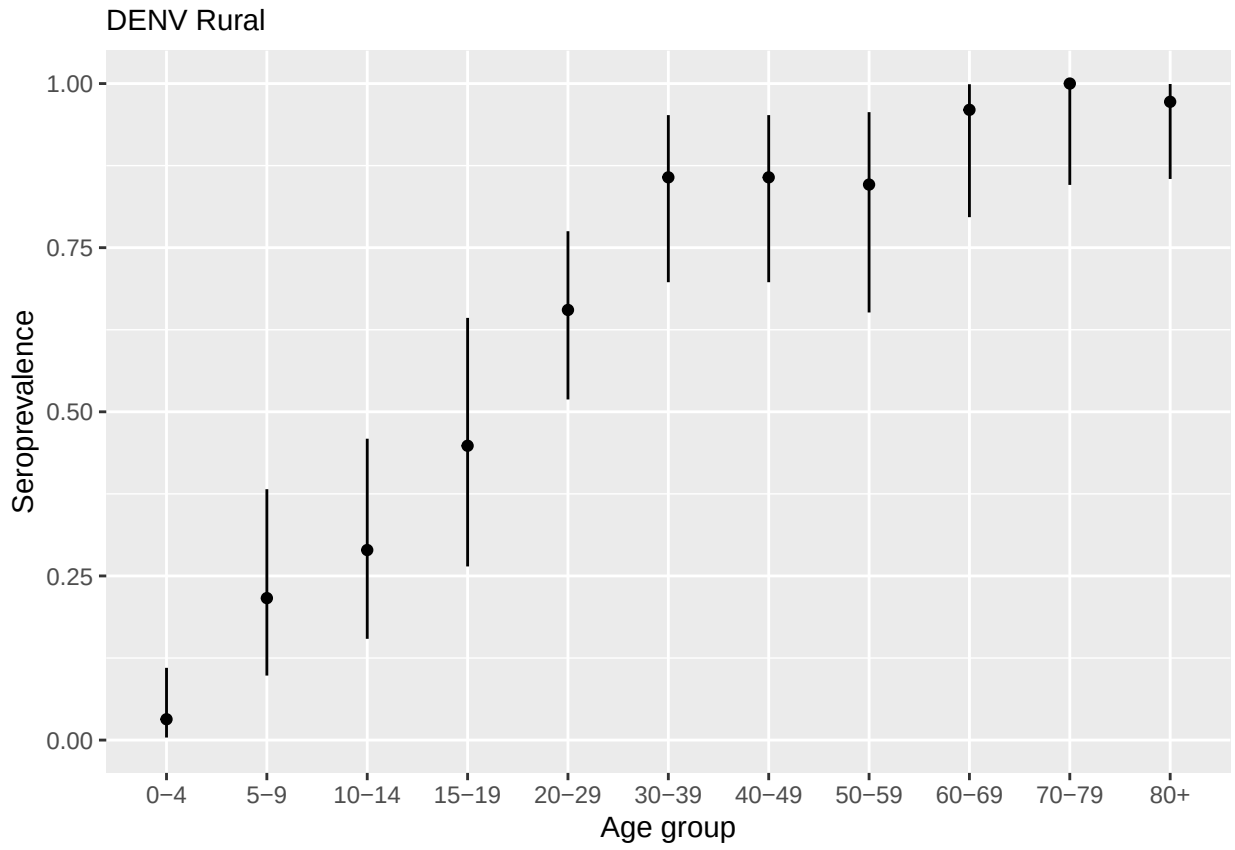
- Question: does DENV FOI vary significantly between urban and rural locations?

```
# create a list to store the seroprevalence data
seroprevUR <- list()

# we now loop through each location
for(loc in c("urban","rural")){

  # create a dataframe and store it in the list
  seroprevUR[[loc]] <- data.frame(age_group=factor(age_groups, levels=age_groups), n=NA, npos=NA)

  # loop through each age group
  for(a in 1:length(age_groups)){
    seroprevUR[[loc]]$n[a] <- nrow(df[df$age_group==age_groups[a] & df$loc==loc, ]) # how many people i
    seroprevUR[[loc]]$npos[a] <- nrow(df[df$age_group==age_groups[a] & df$DENV==1 & df$loc==loc, ]) # h
  }
```

```
  # we can now calculate seroprevalence (npos/n)
  seroprevUR[[loc]]$prev <- seroprevUR[[loc]]$npos/seroprevUR[[loc]]$n

  # and get binomial confidence intervals
  seroprevUR[[loc]][,c("ciL","ciU")] <- binom.exact(seroprevUR[[loc]]$npos, seroprevUR[[loc]]$n)[,c("lo

}


# we can now plot the age-specific seroprevalence
ggplot(seroprevUR$urban, aes(age_group, prev))+ geom_point()+ geom_linerange(aes(ymin=ciL, ymax=ciU))+
  ylab("Seroprevalence")+ xlab("Age group")+ ylim(0,1)+ labs(subtitle="DENV Urban")
```

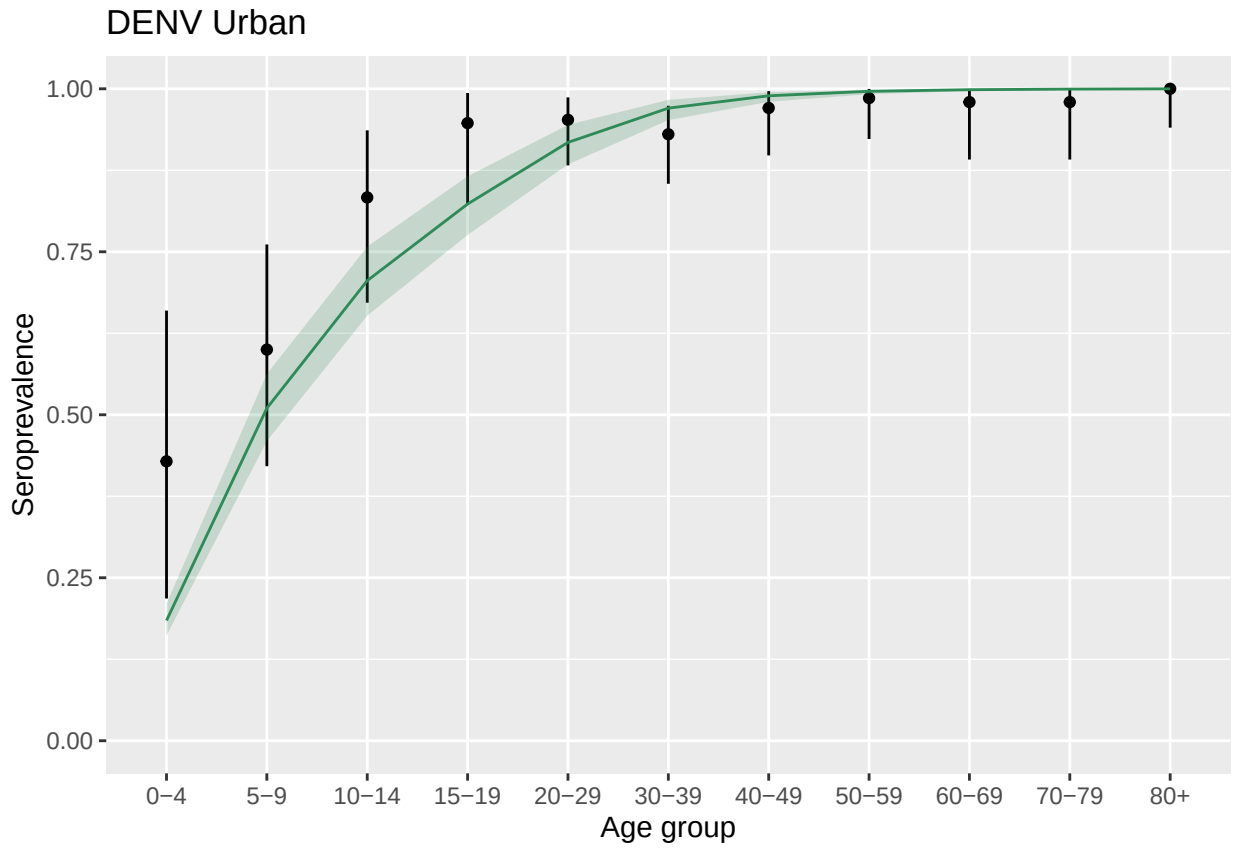



```
ggplot(seroprevUR$rural, aes(age_group, prev))+ geom_point()+ geom_linerange(aes(ymin=ciL, ymax=ciU))+
  ylab("Seroprevalence")+ xlab("Age group")+ ylim(0,1)+ labs(subtitle="DENV Rural")
```

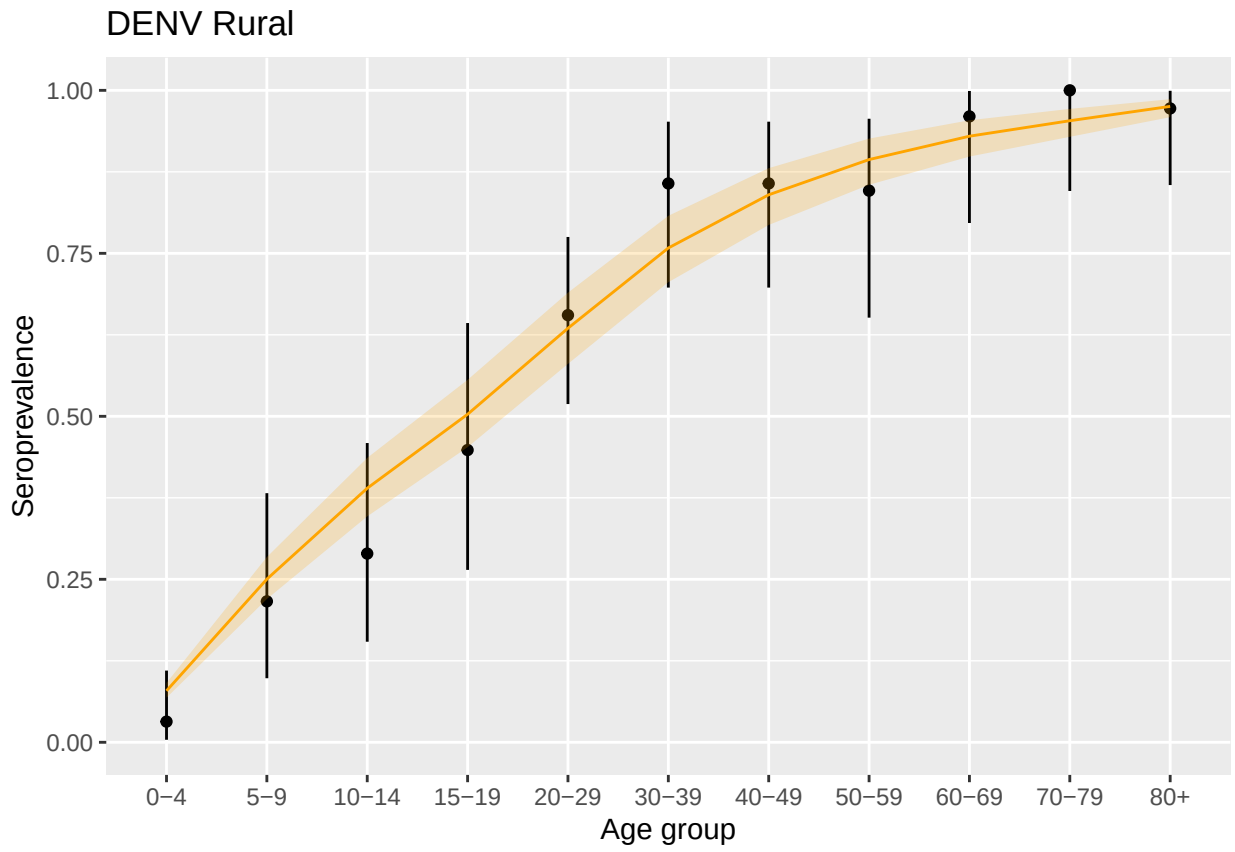



```
# assign the age-group mid-points for fitting the catalytic model
seroprevUR$urban$age_mid <- age_mid
seroprevUR$rural$age_mid <- age_mid


# fit catalytic model for each location
modU <- glm(cbind(npos, n-npos) ~ 1, offset=log(age_mid),
            data=seroprevUR$urban, family = binomial(link = "cloglog")) # urban model

modR <- glm(cbind(npos, n-npos) ~ 1, offset=log(age_mid),
            data=seroprevUR$rural, family = binomial(link = "cloglog")) # rural model


# extract FOI estimates
FOI_urban <- exp(modU$coefficients[1])
FOI_rural <- exp(modR$coefficients[1])

# calculate confidence intervals
ci_FOI_urban <- exp(log(FOI_urban) + c(-1.96, 1.96) * summary(modU)$coefficients[1,2])
ci_FOI_rural <- exp(log(FOI_rural) + c(-1.96, 1.96) * summary(modR)$coefficients[1,2])


# calculate predicted seroprevalence by age group
pred_seroprevU <- data.frame(age_group=age_groups,
                             pred=1-exp(-FOI_urban*age_mid),
                             ciL=1-exp(-ci_FOI_urban[1]*age_mid),
```

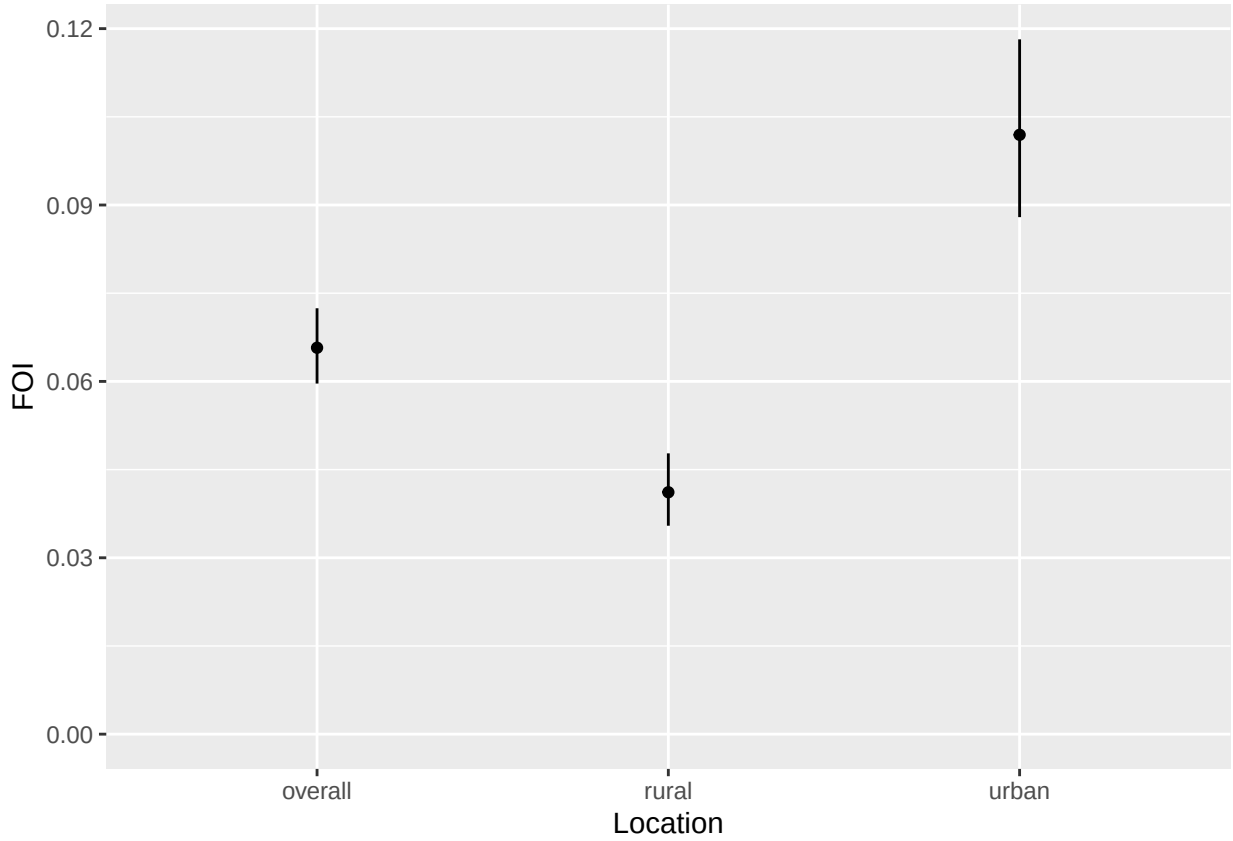
```
                            ciU=1-exp(-ci_FOI_urban[2]*age_mid)) # predictions for urban

pred_seroprevR <- data.frame(age_group=age_groups,
                            pred=1-exp(-FOI_rural*age_mid),
                            ciL=1-exp(-ci_FOI_rural[1]*age_mid),
                            ciU=1-exp(-ci_FOI_rural[2]*age_mid)) # predictions for rural

# plot observed vs model estimated seroprevalence
ggplot(seroprevUR$urban, aes(age_group, prev))+ geom_point()+ ggtitle("DENV Urban")+ ylim(0,1)+
  geom_linerange(aes(ymin=ciL, ymax=ciU))+ ylab("Seroprevalence")+ xlab("Age group")+
  geom_line(data=pred_seroprevU, aes(age_group, pred, group=1), col="seagreen")+
  geom_ribbon(data=pred_seroprevU, aes(x=age_group, y=pred, ymin=ciL, ymax=ciU, group=1), fill="seagreen
```




```
ggplot(seroprevUR$rural, aes(age_group, prev))+ geom_point()+ ggtitle("DENV Rural")+ ylim(0,1)+
  geom_linerange(aes(ymin=ciL, ymax=ciU))+ ylab("Seroprevalence")+ xlab("Age group")+
  geom_line(data=pred_seroprevR, aes(age_group, pred, group=1), col="orange")+
  geom_ribbon(data=pred_seroprevR, aes(x=age_group, y=pred, ymin=ciL, ymax=ciU, group=1), fill="orange"
```




```
# we can now also plot a comparison of our FOI estimates
foiests <- data.frame(loc=c("overall","urban","rural"),
                      FOI=NA, ciL=NA, ciU=NA)
foiests[1,2:4] <- c(FOI, ci_FOI)
foiests[2,2:4] <- c(FOI_urban, ci_FOI_urban)
foiests[3,2:4] <- c(FOI_rural, ci_FOI_rural)

# plot
ggplot(foiests, aes(loc, FOI))+ geom_point()+ geom_linerange(aes(ymin=ciL, ymax=ciU))+
  ylim(0,NA)+ xlab("Location")
```



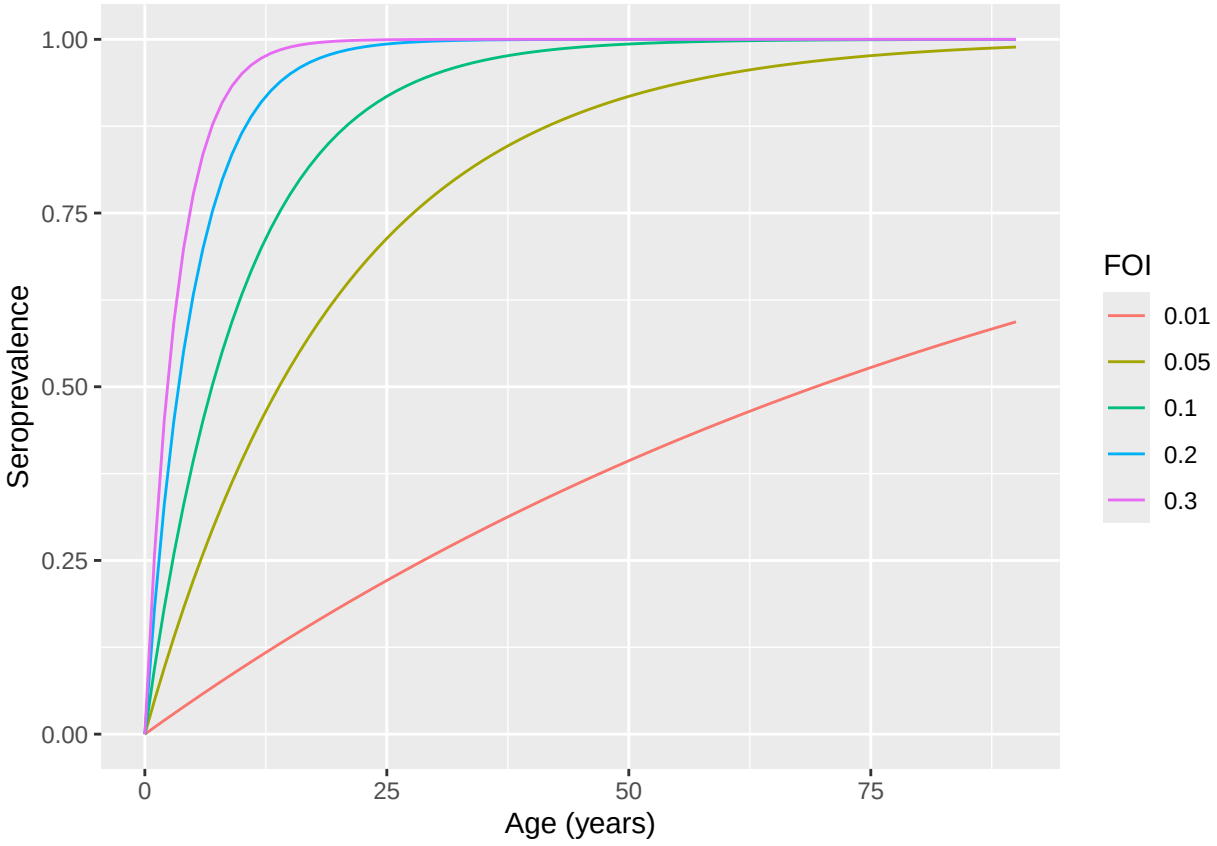
7. Further exploration of FOI (extra exercises if time allows)

Here we will explore what different values of FOI mean for age-specific seroprevalence and susceptibility.

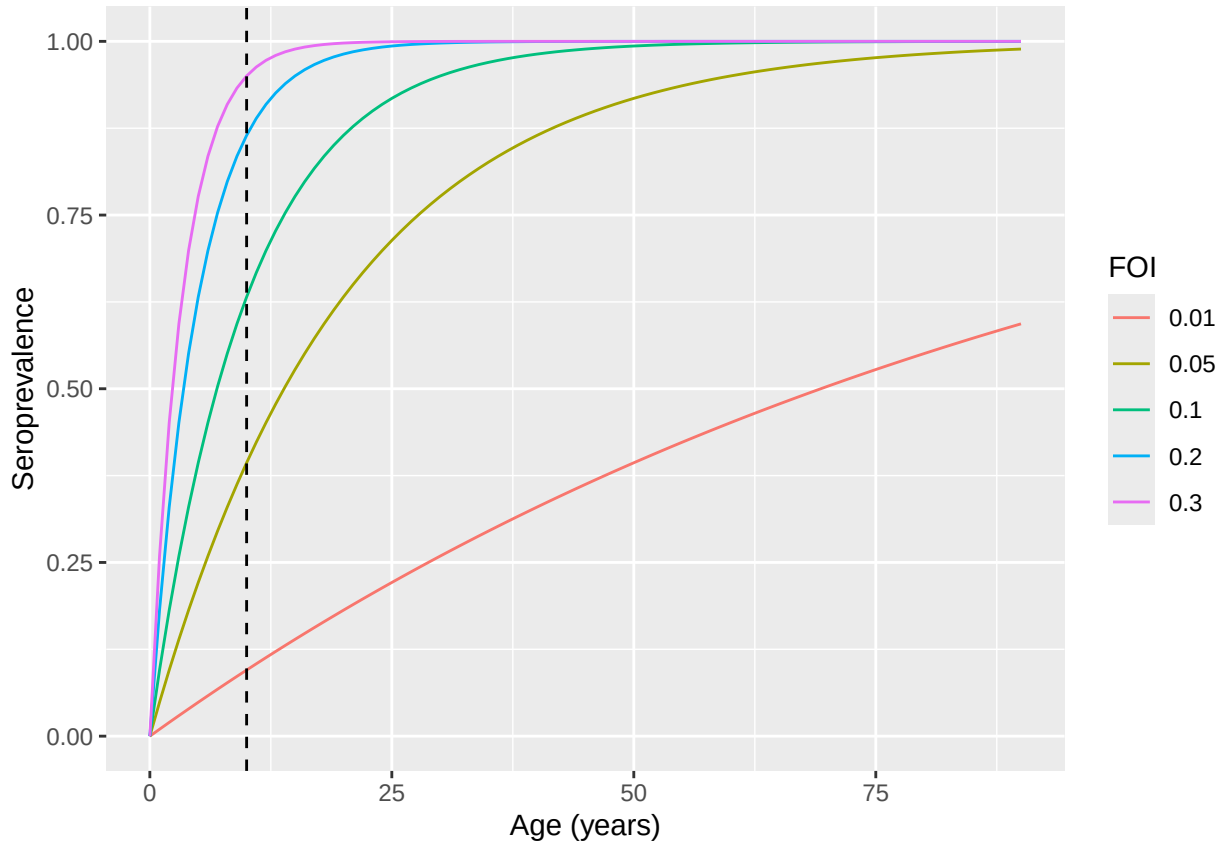
- Question: What proportion of 10 year olds would we expect to have been infected vs susceptible for different values of FOI?

```
# lets take some example FOI values (feel free to add your own values here)
explore_FOIs <- c(0.01, 0.05, 0.1, 0.2, 0.3)

# we will loop through each of these values and simulate expected age-specific seroprevalence values
explore_data <- list() # create a list to store the simulated data
for(i in 1:length(explore_FOIs)){

  edf <- data.frame(age=seq(0,90),
                    FOI=explore_FOIs[i],
                    prev=1-exp(-explore_FOIs[i]*seq(0,90)),
                    susc=exp(-explore_FOIs[i]*seq(0,90)))
  explore_data[[i]] <- edf

}

# here we unlist and bind the data together for plotting
explore_data <- do.call("rbind", explore_data)
```

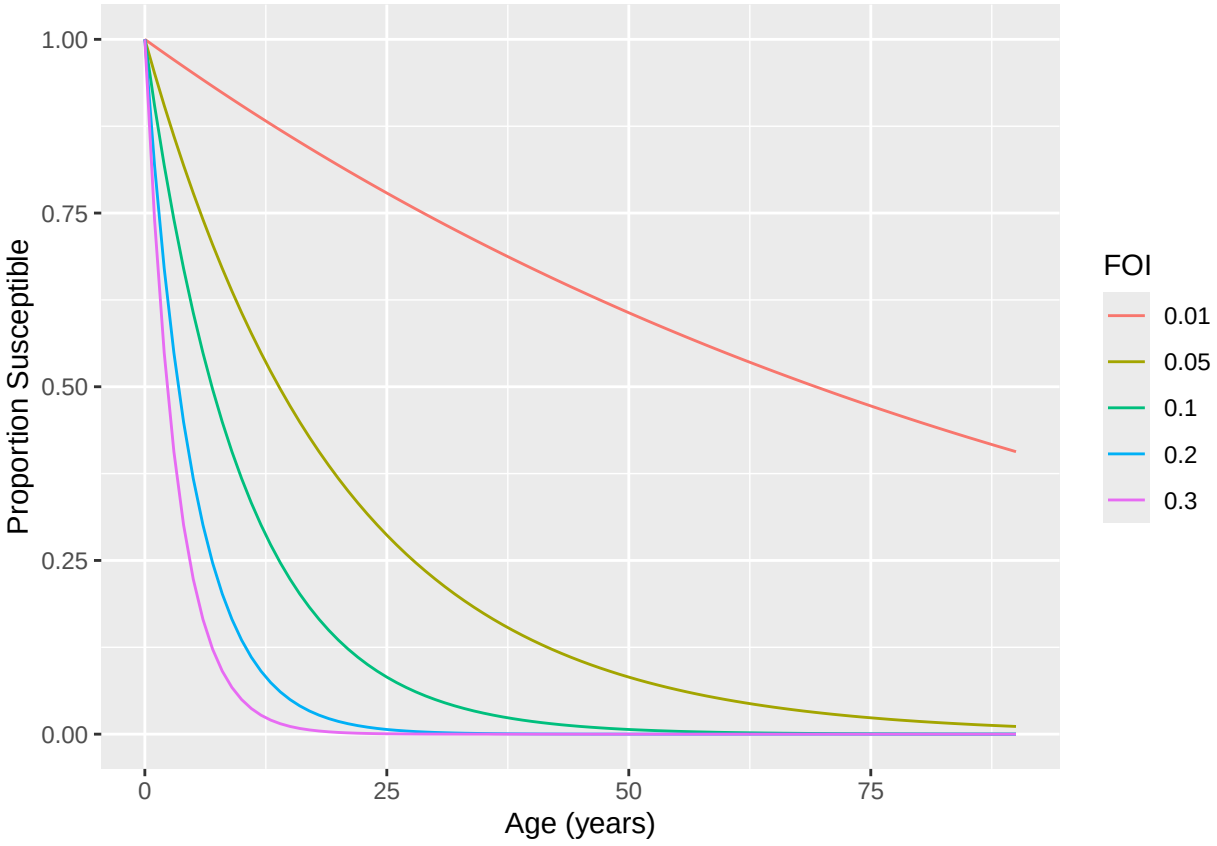
```
# plotting expected age-specific prevalence for different FOIs
ggplot(explore_data, aes(age, prev, col=factor(FOI)))+ geom_line()+
  ylab("Seroprevalence")+ xlab("Age (years)")+ labs(col="FOI")
```




```
# add a line at age 10 years
ggplot(explore_data, aes(age, prev, col=factor(FOI)))+ geom_line()+
  ylab("Seroprevalence")+ xlab("Age (years)")+ labs(col="FOI")+
  geom_vline(aes(xintercept=10), linetype="dashed")
```




```
# now lets also see what the proportion susceptible would look like in each scenario
ggplot(explore_data, aes(age, susc, col=factor(FOI)))+ geom_line()+
  ylab("Proportion Susceptible")+ xlab("Age (years)")+ labs(col="FOI")
```


1.00
0.75
0.50
0.25
0.00
Proportion Susceptible
0
25
50
75
Age (years)
FOI
0.01
0.05
0.1
0.2
0.3