Transmission Risk Comparison

Remo Schmutz

2022-12-21

Assumptions and parameter estimations

We use Rudnick's revised Wells-Riley equation to estimate the transmission risk for tuberculosis. There are different parameter which have to be estimated from data or from the literature. The equation used is described in the paper from Rudnick on page 238 and 239.

The risk of infection can be computed as follows:

$P = exp(-\frac{\bar{f}*I*q*t}{n})$ (1)

The following parameters/helpfunctions are used:

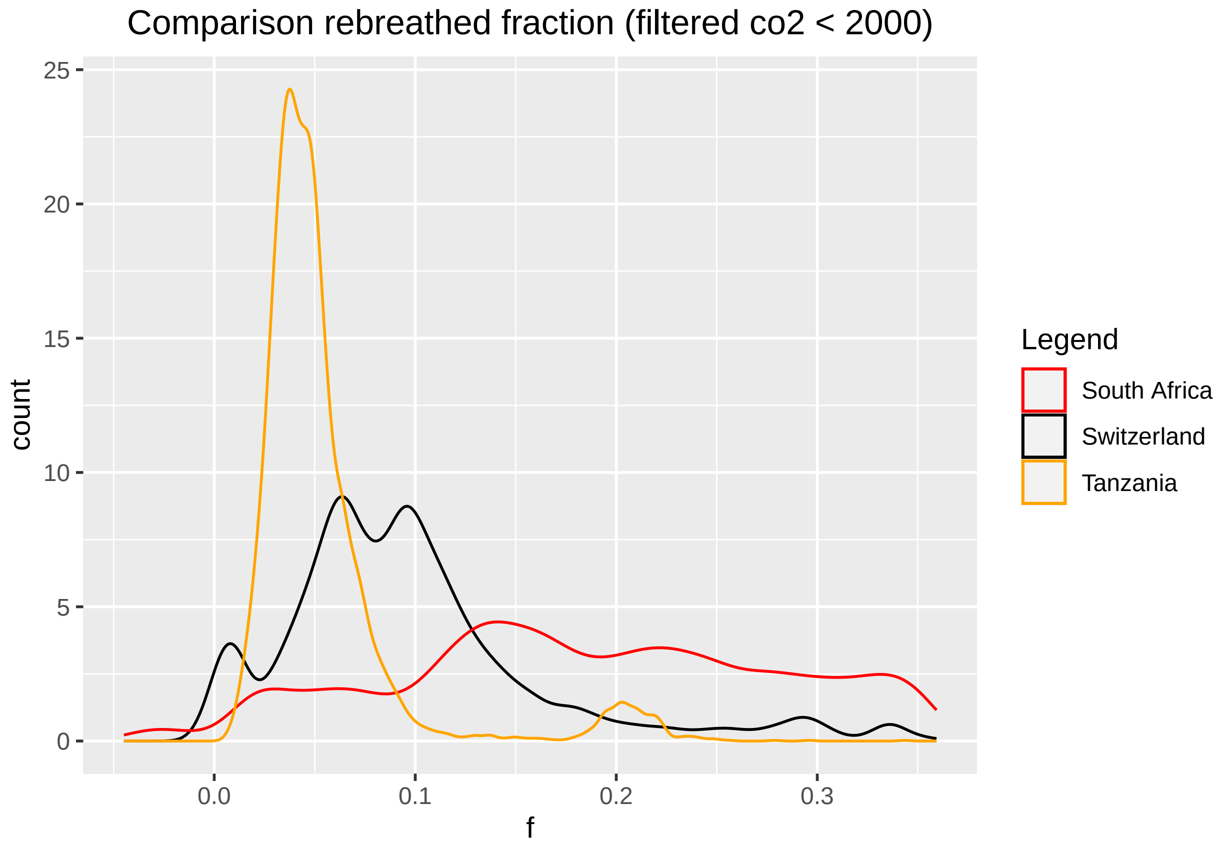
$f = \frac{C-C_o}{C_a}$ (2)

- $C :=$ Indoor Co2 concentration
  - mean or distribution from data
- $C_o :=$ Outdoor Co2 concentration
  - from literature https://www.fsis.usda.gov/sites/default/files/media_file/2020-08/Carbon-Dioxide.pdf
- $C_a :=$ Volume fraction of CO2 added to exhaled breath during breathing
  - Persily and de Jonge [Table 3 and 4] doi: 10.1111/ina.12383
- $\bar{f} := \int_{t=0}^{t=max} f dt$
  - integrating over f values from different times (2) or using a distribution based on the data
- $I :=$ Number of infectors in the class
  - estimated using prevalence of the age group in the country
- $q :=$ Quantum per hour
  - assuming a distribution from literature
- $t :=$ time
  - changing this parameter to compare
- $n :=$ number of people in the class
  - data (Switzerland) or assumption (South Africa, Tanzania)

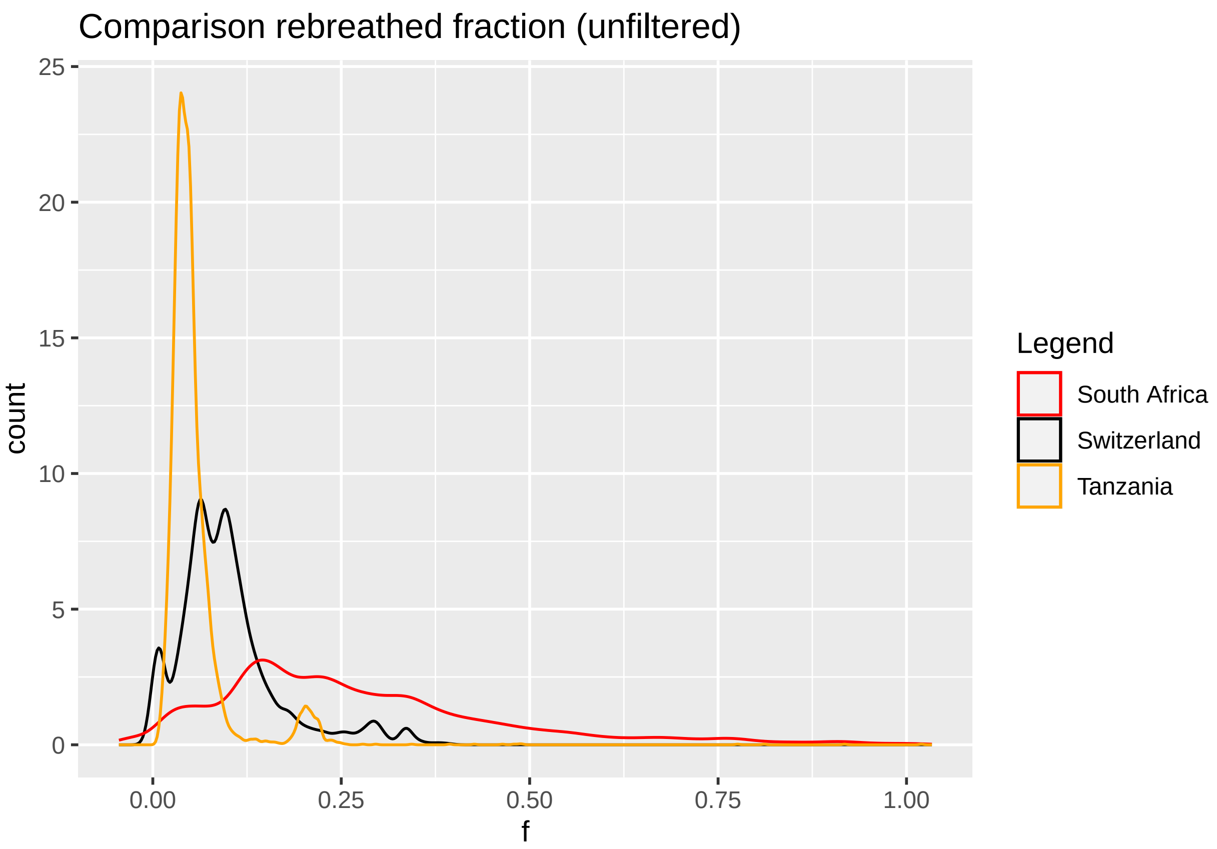
Calculate the f-values per time:

```
C_a <- (0.0048+0.0041)/2 #using M=1.4 in table 3  (mean of "Sitting reading, writing, typing" and "Sitt
C_o <- 400 #p.p.m (taking a higher estimate because higher values ar possible when a lot of traffic ect

ch1 <- ch %>%
  filter(co2 <2000) %>% #filtering extreme values
  mutate(
    f = ((co2-C_o)/C_a)/1000000) # f in decimal

ch1_unfiltered <- ch %>%
  mutate(
    f = ((co2-C_o)/C_a)/1000000) # f in decimal

satz1 <- satz %>%
  mutate(
    f = ((co2-C_o)/C_a)/1000000)# f in decimal

tz1 <- satz1 %>%
  filter(co2 < 2000) %>%
  filter(country == "Tanzania")

tz1_unfiltered <- satz1 %>%
  filter(country == "Tanzania")

sa1 <- satz1 %>%
  filter(co2 < 2000) %>%
  filter(country == "South Africa")

sa1_unfiltered <- satz1 %>%
  filter(country == "South Africa")

colors <- c("Switzerland" = "black", "South Africa" = "red", "Tanzania" = "orange")

plot_distribution_f <- ggplot(ch1, aes(x=f)) +
  geom_density(aes(color = "Switzerland")) +
  geom_density(data = sa1, aes(color = "South Africa")) +
  geom_density(data = tz1, aes(color = "Tanzania")) +
  ggtitle("Comparison rebreathed fraction (filtered co2 < 2000)") +
  labs(x = "f",
       y = "count",
       color = "Legend") +
  scale_color_manual(values = colors) +
   theme(plot.title = element_text(hjust = 0.5))

plot_distribution_f_unfiltered <- ggplot(ch1_unfiltered, aes(x=f)) +
  geom_density(aes(color = "Switzerland")) +
  geom_density(data = sa1_unfiltered, aes(color = "South Africa")) +
  geom_density(data = tz1_unfiltered, aes(color = "Tanzania")) +
  ggtitle("Comparison rebreathed fraction (unfiltered)") +
  labs(x = "f",
       y = "count",
       color = "Legend") +
  scale_color_manual(values = colors)
```

```
plot_distribution_f
```



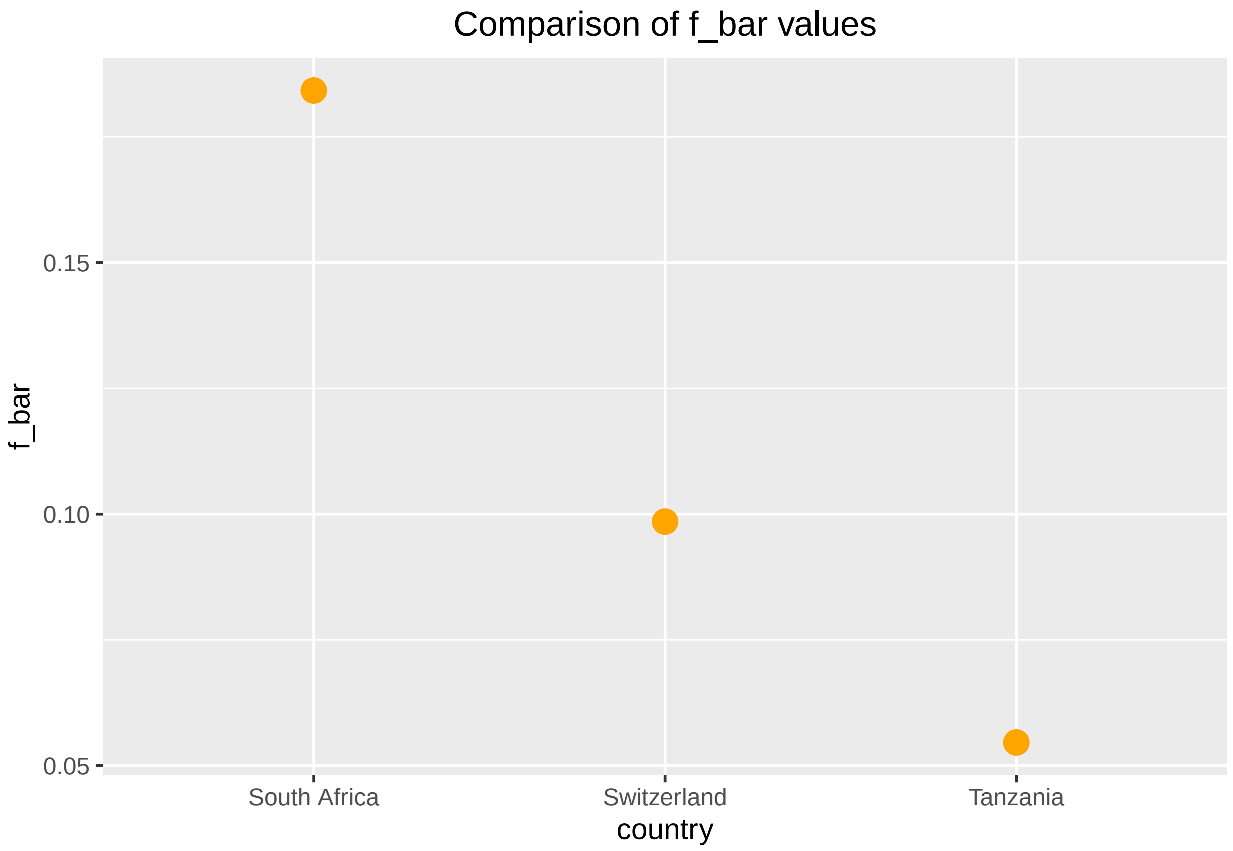
```
plot_distribution_f_unfiltered
```



If we look at the plot, we can see that the value does differ somewhat between countries, creating an distribution with a right tail. I filter values over 2000 p.p.m and compare it to the unfiltered values. For the following calculations i will use the filtered dataset. Remember f is equivalent to the fraction of indoor air that is exhaled breath, which is also the rebreathed fraction. Larger values for f mean that the air quality is worse. As i don't assume different values for outside air quality and the volume of co2 added to air during breathing between countries, the variance of the f_bar values and the co2 values between countries follow the same pattern. For this reason i will sometimes switch between the two. It is also evident that the distribution takes a different form for each country. If I want to work with probability distributions, I would have to use a different one for each country. This contains further sources of error, s.d the comparison between the countries thereby risks to become imprecise.

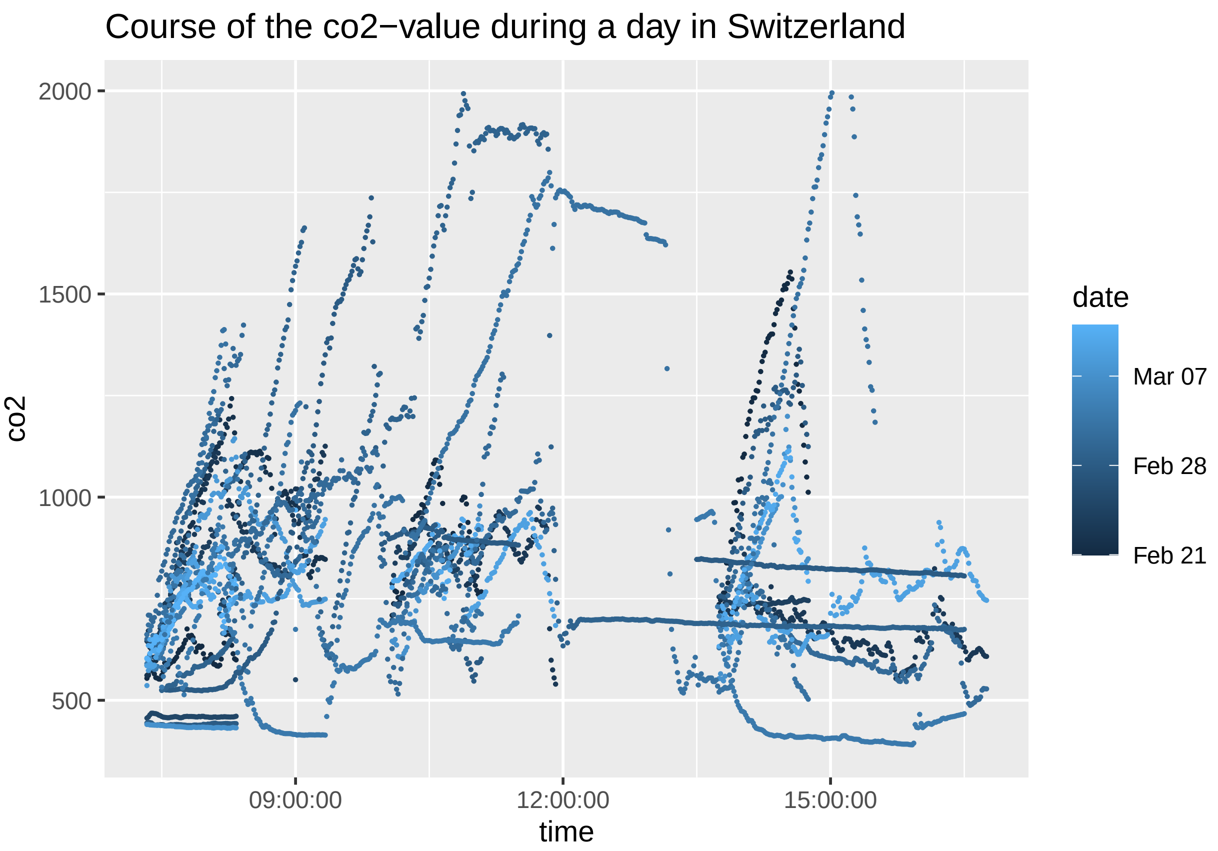
Calculate $\bar{f}$-values using the f-values $(a1)$:

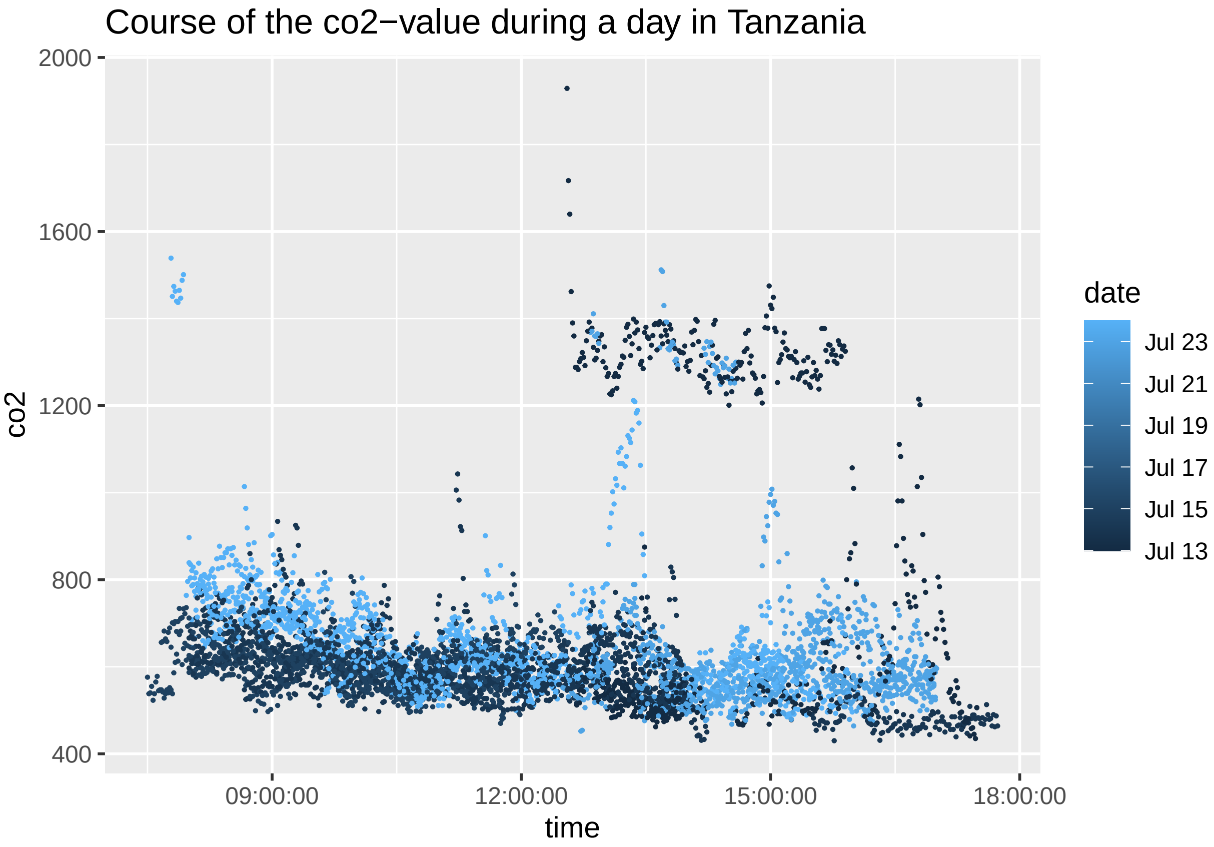
```
f_bar_ch <- ch1 %>%
  summarize(f_bar = mean(f, na.rm = TRUE))

f_bar_sa <- sa1 %>%
  summarize(f_bar = mean(f, na.rm = TRUE))

f_bar_tz <- tz1 %>%
  summarize(f_bar = mean(f, na.rm = TRUE))

f_bar_compare <- tibble(country=c("Switzerland", "South Africa", "Tanzania"), f_bar =  c(f_bar_ch[1,1,d

plot_f_bar <- ggplot(f_bar_compare, aes(x=country, y=f_bar)) +
  geom_point( color="orange", size=4) +
  ylab("f_bar") +
  ggtitle("Comparison of f_bar values") +
  theme(plot.title = element_text(hjust = 0.5))

plot_f_bar
```

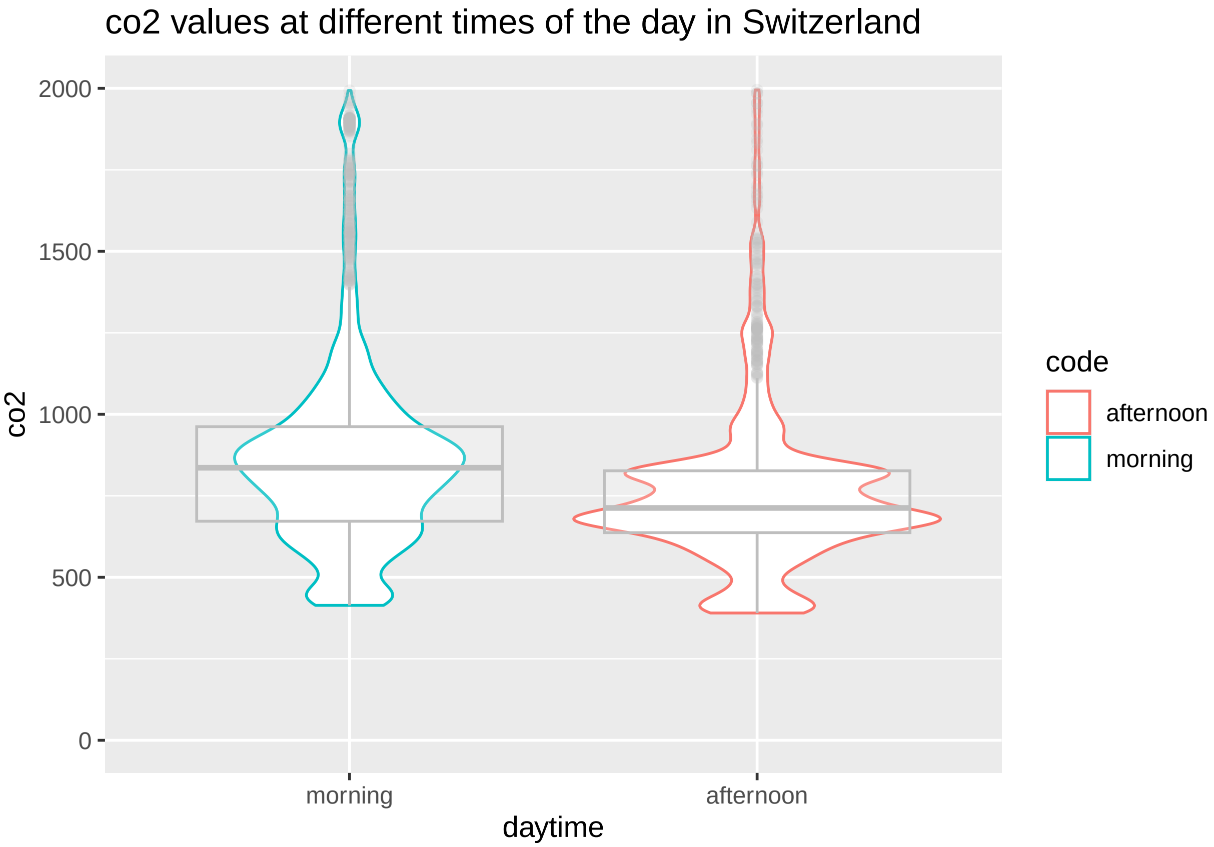


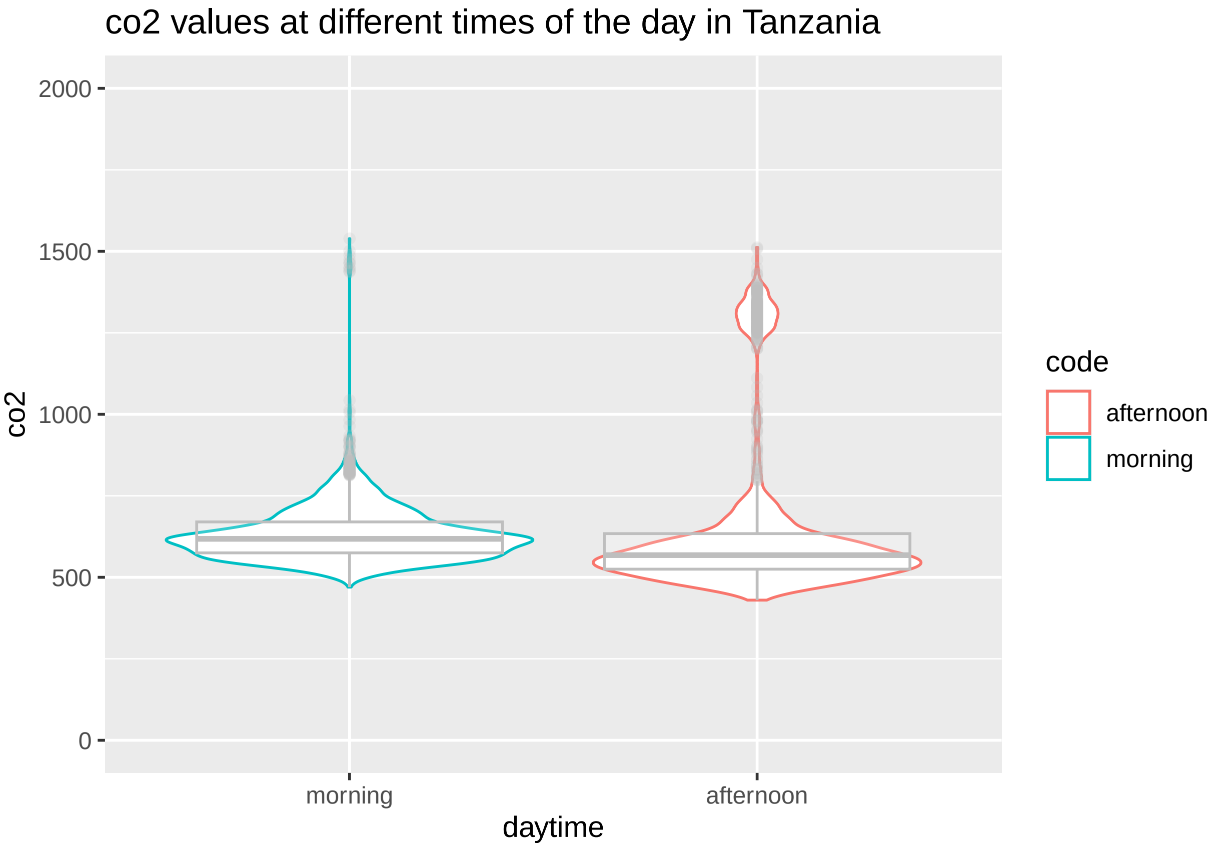
[For Tanzania and Switzerland only]

It would be interesting to see, where the variance comes from in the co2 data comes from inside a country. A possible reason could be daytime. In a first, simple analysis i will compare morning (07-1200) and afternoon (1330-17). In a second analysis i would like to incorparate the times where there is a break which is a possibility to open the windows and get fresh air. (possible for Switzerland and Tanzania)


Course of the co2−value during a day in Switzerland
co2
2000
1500
1000
500
time
09:00:00
12:00:00
15:00:00
date
Mar 07
Feb 28
Feb 21

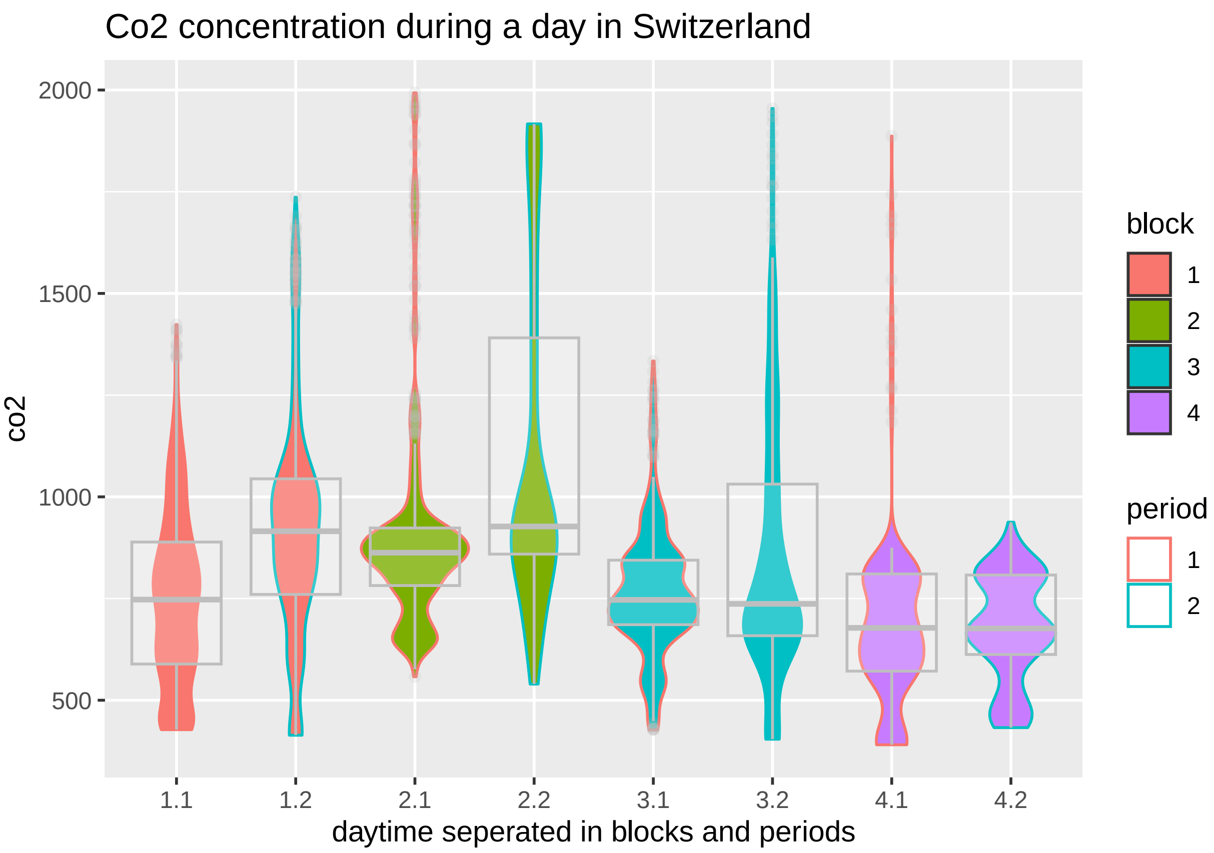

Course of the co2−value during a day in Tanzania
2000
1600
1200
800
400
co2
09:00:00
12:00:00
15:00:00
18:00:00
time
date
Jul 23
Jul 21
Jul 19
Jul 17
Jul 15
Jul 13

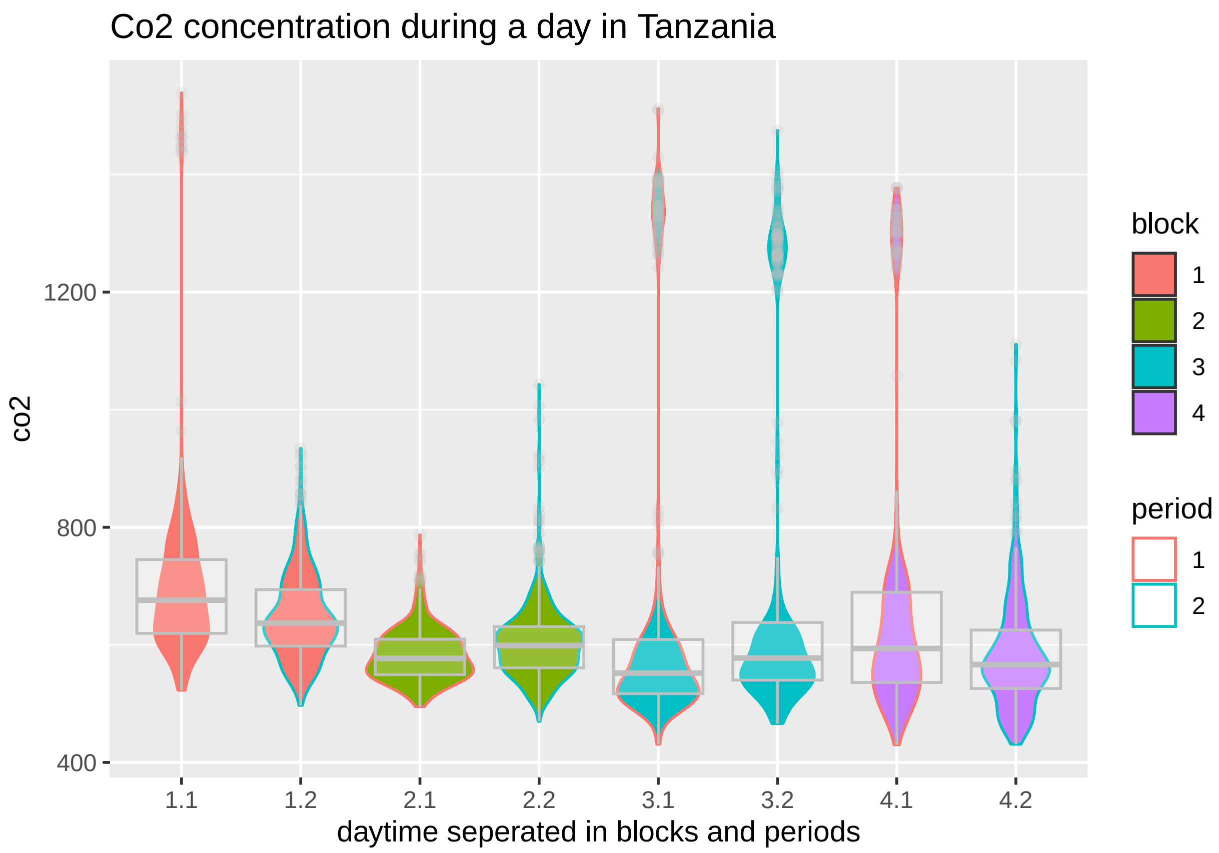

co2 values at different times of the day in Switzerland
2000
1500
1000
500
0
co2
morning
afternoon
daytime
code
afternoon
morning

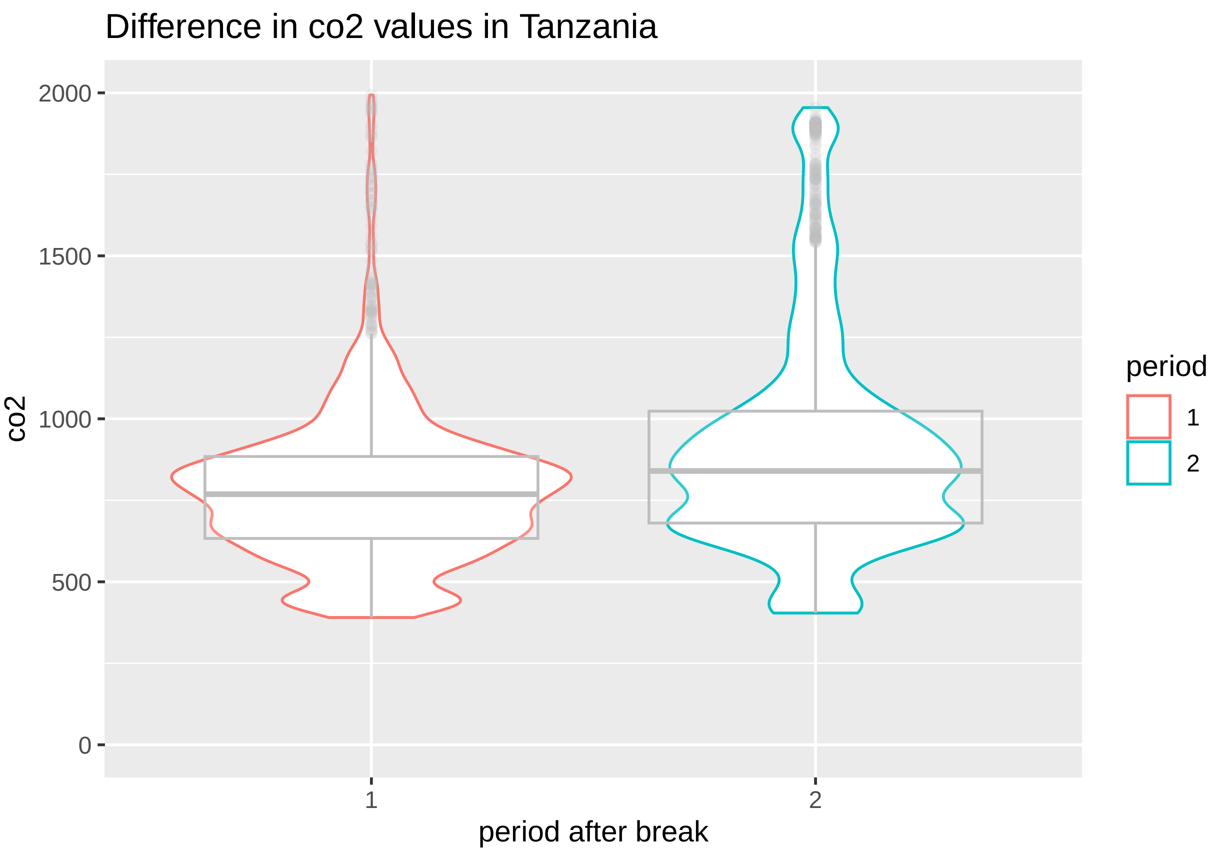


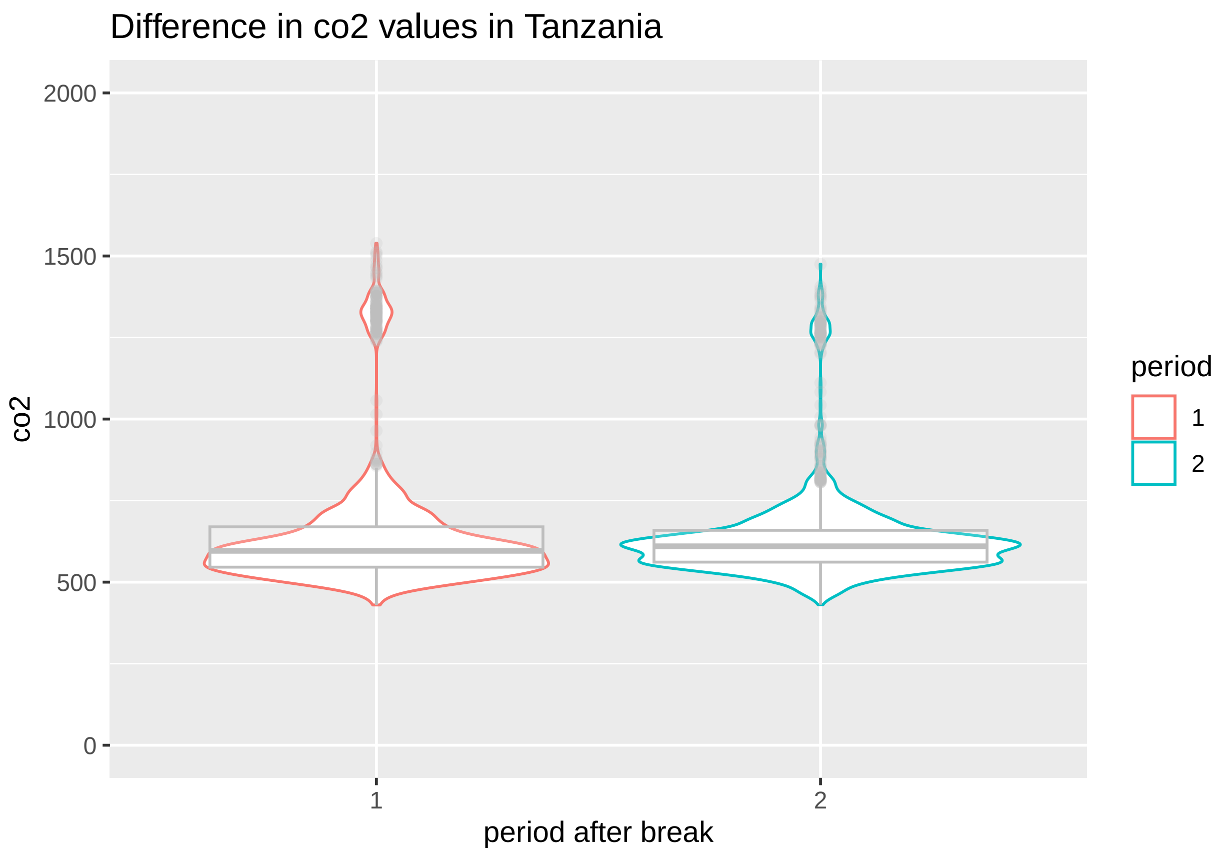
We see that after the lunch break, the levels decrease to the levels in beginning of the morning, also the averages do not really differ and are overall lower in the afternonn. Therefore we drop the assumption, that the concentration increases steadily during a day. Instead we follow the theory, that it decreases with duration after a break. I now assume that breaks are taken from 1000-1015, from 1200-1330 and from 1500-1515. That means we have four blocks:

block1: 0730-1000 block2: 1015-1200 block3: 1330-1500 block4: 1515-1645 period indicates if it is the first or the second half of the block.


Co2 concentration during a day in Switzerland
co2
2000
1500
1000
500
1.1
1.2
2.1
2.2
3.1
3.2
4.1
4.2
daytime seperated in blocks and periods
block
1
2
3
4
period
1
2


Co2 concentration during a day in Tanzania
co2
1200
800
400
1.1
1.2
2.1
2.2
3.1
3.2
4.1
4.2
daytime seperated in blocks and periods
block
1
2
3
4
period
1
2


Difference in co2 values in Tanzania
co2
2000
1500
1000
500
0
1
2
period after break
period
1
2



Indeed, we see variation in the data. The co2-values seem to increase in the second period after the break in Switzerland. This observation serves as motivation to pursue this approach further. [Here I should check if my assumptions for the break times are correct] In the data from tanzania there isn't the same pattern, especially after the lunch break we see very high values. The observation concerning the course of the co2 values could be interesting for modelling the change of transmission risk during a day.

Now i will proceed with the q-value and the parameter I. Here we will consider a truncted student distribution. It is important to take a distribution that has the most values at very low levels, but then has a long tail probability. Very high values should be unlikely but possible. The expression of the quanta in different persons is characterized by this characteristic.

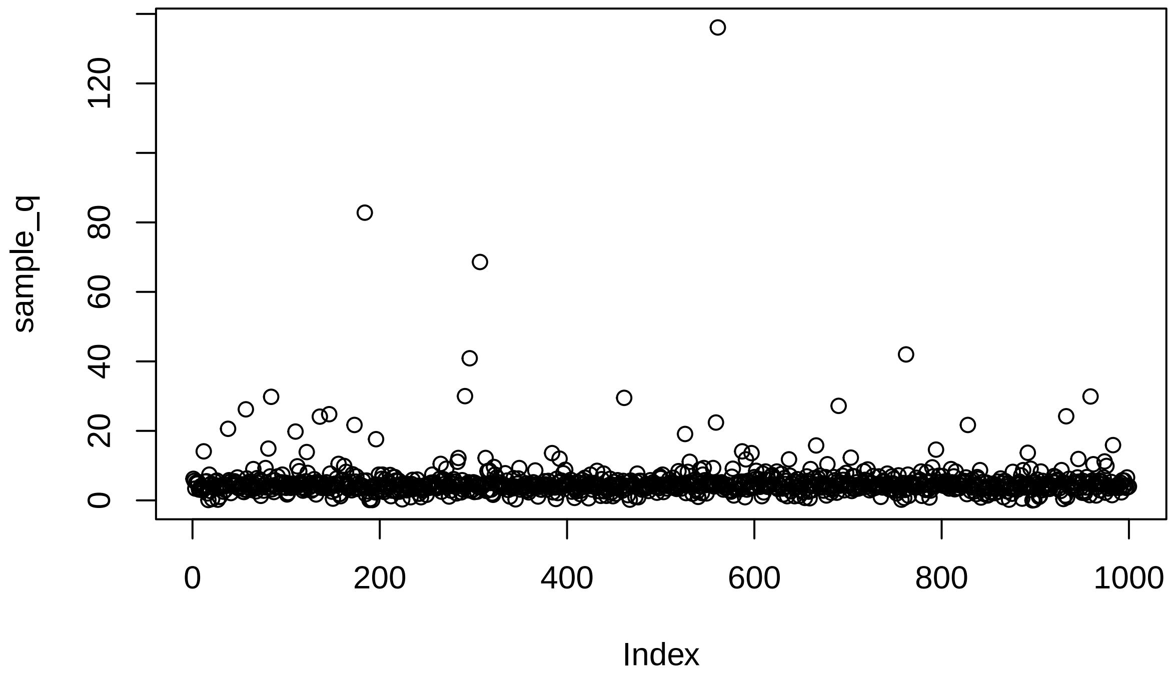
Since we have limited data to estimate a distribution on it, I had to be a little imprecise to determine the parameters. I determined the mean value via the weighted average of the different studies. I then determined the standard deviation and the degrees of freedom by hand. I did this in such a way that the simulations each had about the same number of high quanta (>10) as reported by Escombe. I would like to point out that this approach is not very exact and there may be better alternatives to estimate the parameters.

I'll use the following studies for calculating the meanparameter:

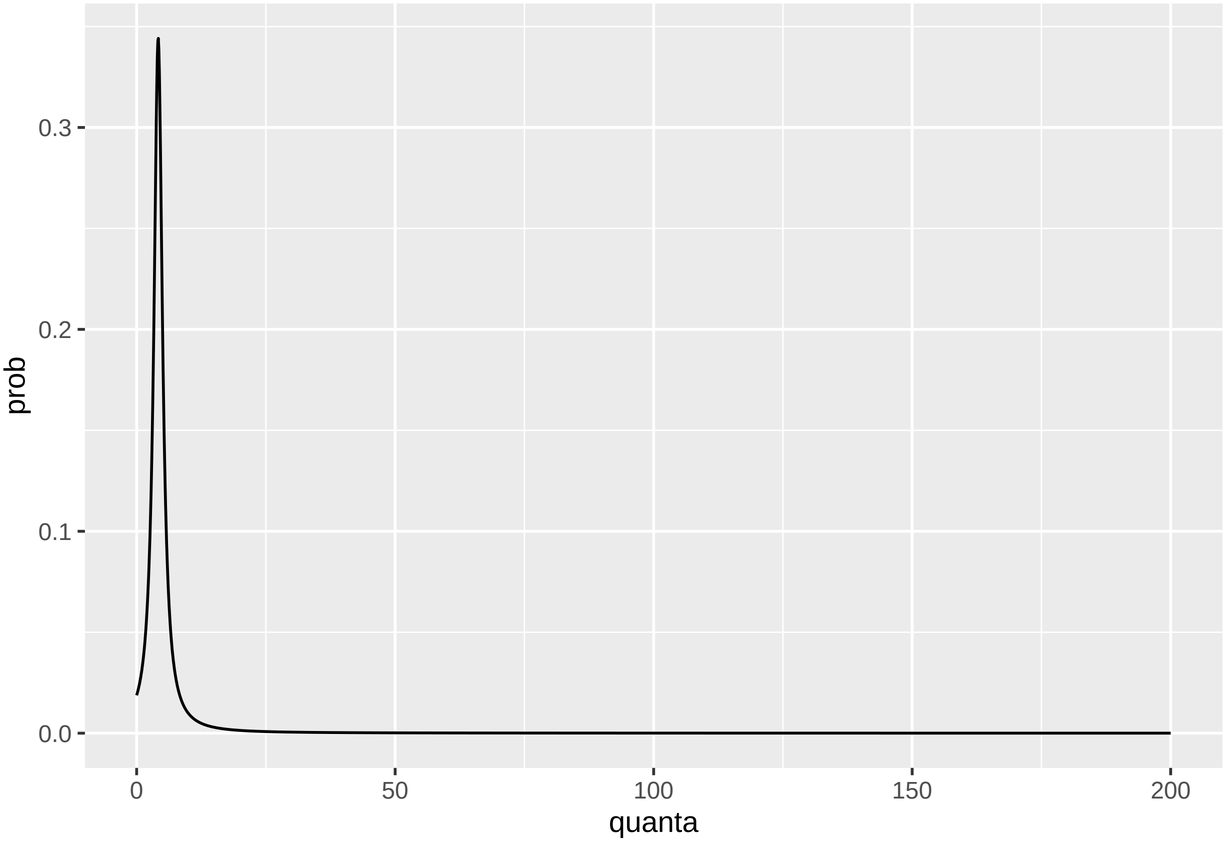
Riley (1962): 130 patients, q-Wert: 1.25 Escombe (2008): 117 patients, q-Wert: 8.2 Nardell (1991) : 1 patients, q-Wert: 12.5 Andrews (2014) : 571 patients, q-Wert: 0.89 Dhamadhakari (2012) : 17 patients, q-Wert: 138/34 (no mask/mask)

```
q <- (1.25*130+8.2*117+12.5+0.89*571+138*17)/(130+117+1+571+138) #mean

#Escombe Table 2
mean_one_inf <- mean(c(12,3,5.5,1.8,18,12)) #mean quanta of pers. which infected one pig
mean_two_inf <- mean(c(2.9,40)) #mean quanta of pers. which infected two pigs
```

```r
q_inf_persons <- c(12,3,2.9,5.5,1.8,18,40,12,226,52,mean_two_inf,rep(mean_one_inf,11))
# reported quanta plus the two missing
q_sample_total_unif <- c(q_inf_persons, runif(117-length(q_inf_persons), min = 0, max =1))
#  rest unif in [0,1], as quanta below 1 isnt enough to infect an indidual
sd <- sd(q_sample_total_unif)

dq <- function(x) {
  dtrunc(x, spec = "st", a = 0, b = 300, mu = q, sigma = 1, nu = 1)
} #function using parameters from option 1

rq_distr <- data.frame(quanta = seq(0, 200, .1)) %>%
  mutate(prob = dq(quanta))

sample_q <- sample(rq_distr$quanta, 1000, replace = TRUE, prob = rq_distr$prob)

plot2 <- ggplot(rq_distr, aes(x = quanta, y = prob)) +
  geom_line()

plot(sample_q)
```

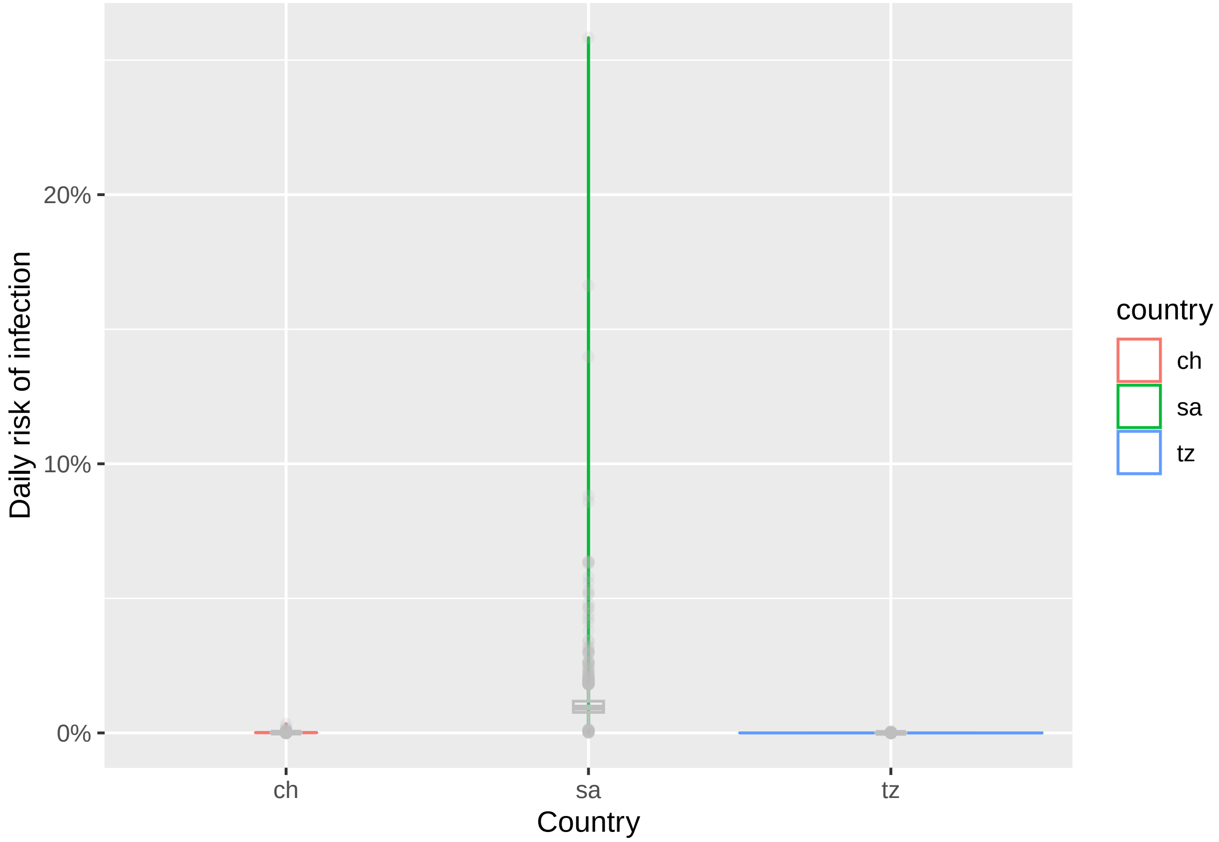


```r
plot2
```

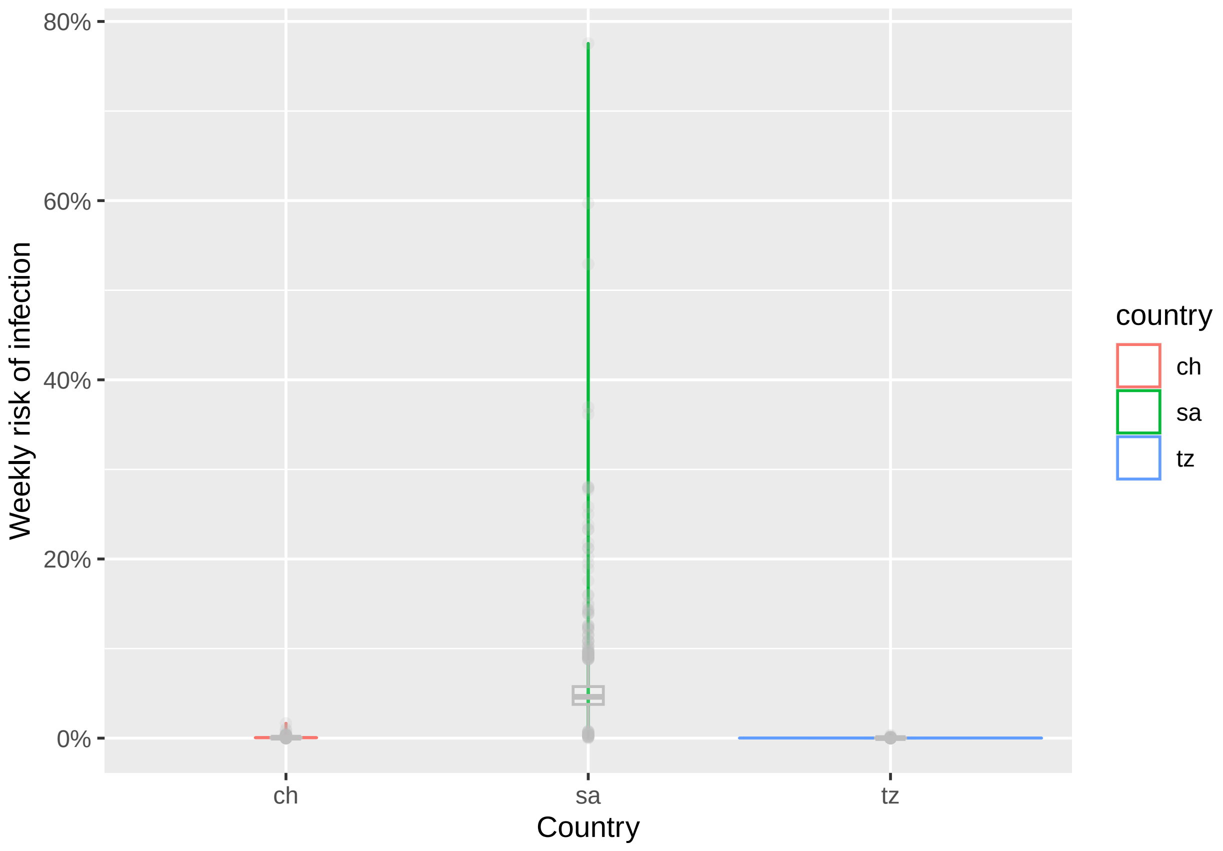


```
class_ch <- 20
class_sa <- 30 #Powerpoint
class_tz <- 50 #Powerpoint

prev_youth_ch <- (1.65 + 1.14 + 0.46 + 5.19 + 5.58 + 4.7)/600000 #decimal; values for age group 10-14 a
prev_youth_sa <- 149/100000 #decimal; for age group 15-24; https://www.thelancet.com/action/showPdf?pii
prev_youth_tz <- 1177/61741120
#0.4 https://ntlp.go.tz/tuberculosis/paediatric-tb/

I_ch <- prev_youth_ch*class_ch #prevalence per class (per year)
I_sa <- prev_youth_sa*class_sa
I_tz <- prev_youth_tz*class_sa

f_bar_ch <- f_bar_compare[1,2, drop=TRUE]
f_bar_sa <- f_bar_compare[2,2, drop=TRUE]
f_bar_tz <- f_bar_compare[3,2, drop=TRUE]

day <- 8
week <- 8*5
month <- 8*5*4
year <- 8*5*4*12

#preparing datasets for plotting
df_ch <- tibble(country = c(rep("ch",1000)), f_bar = c(rep(f_bar_ch,1000)), f_bar_first = c(rep(f_bar_pa
  mutate(P_year = 1 - exp(-(f_bar*I*q*year)/n)) %>%
  mutate(P_month = 1 - exp(-(f_bar*I*q*month)/n)) %>%
  mutate(P_week = 1 - exp(-(f_bar*I*q*week)/n)) %>%
  mutate(P_day = 1 - exp(-(f_bar*I*q*day)/n)) %>%
  mutate(P_year_one = 1 - exp(-(f_bar*0.01*q*year)/n)) %>%
  mutate(P_month_one = 1 - exp(-(f_bar*0.01*q*month)/n)) %>%
  mutate(P_week_one = 1 - exp(-(f_bar*0.01*q*week)/n)) %>%
  mutate(P_day_one = 1 - exp(-(f_bar*0.01*q*day)/n)) %>%
  mutate(P_year_first = 1 - exp(-(f_bar_first*I*q*year)/n)) %>%
  mutate(P_month_first = 1 - exp(-(f_bar_first*I*q*month)/n)) %>%
  mutate(P_week_first = 1 - exp(-(f_bar_first*I*q*week)/n)) %>%
  mutate(P_day_first = 1 - exp(-(f_bar_first*I*q*day)/n)) %>%
  mutate(P_year_second = 1 - exp(-(f_bar_second*I*q*year)/n)) %>%
  mutate(P_month_second = 1 - exp(-(f_bar_second*I*q*month)/n)) %>%
  mutate(P_week_second = 1 - exp(-(f_bar_second*I*q*week)/n)) %>%
  mutate(P_day_second = 1 - exp(-(f_bar_second*I*q*day)/n))

df_sa <- tibble(country = c(rep("sa",1000)), f_bar = c(rep(f_bar_sa,1000)), I = c(rep(I_sa,1000)), I_sta
  mutate(P_year = 1 - exp(-(f_bar*I*q*year)/n)) %>%
  mutate(P_month = 1 - exp(-(f_bar*I*q*month)/n)) %>%
  mutate(P_week = 1 - exp(-(f_bar*I*q*week)/n)) %>%
  mutate(P_day = 1 - exp(-(f_bar*I*q*day)/n)) %>%
  mutate(P_year_one = 1 - exp(-(f_bar*0.01*q*year)/n_ch)) %>%
  mutate(P_month_one = 1 - exp(-(f_bar*0.01*q*month)/n_ch)) %>%
  mutate(P_week_one = 1 - exp(-(f_bar*0.01*q*week)/n_ch)) %>%
  mutate(P_day_one = 1 - exp(-(f_bar*0.01*q*day)/n_ch))

df_tz <- tibble(country = c(rep("tz",1000)), f_bar = c(rep(f_bar_tz,1000)), I = c(rep(I_tz,1000)), I_sta
  mutate(P_year = 1 - exp(-(f_bar*I*q*year)/n)) %>%
  mutate(P_month = 1 - exp(-(f_bar*I*q*month)/n)) %>%
  mutate(P_week = 1 - exp(-(f_bar*I*q*week)/n)) %>%
  mutate(P_day = 1 - exp(-(f_bar*I*q*day)/n)) %>%
  mutate(P_year_one = 1 - exp(-(f_bar*0.01*q*year)/n_ch)) %>%
  mutate(P_month_one = 1 - exp(-(f_bar*0.01*q*month)/n_ch)) %>%
  mutate(P_week_one = 1 - exp(-(f_bar*0.01*q*week)/n_ch)) %>%
  mutate(P_day_one = 1 - exp(-(f_bar*0.01*q*day)/n_ch))

df_complet <- bind_rows(df_ch, df_sa, df_tz)
```

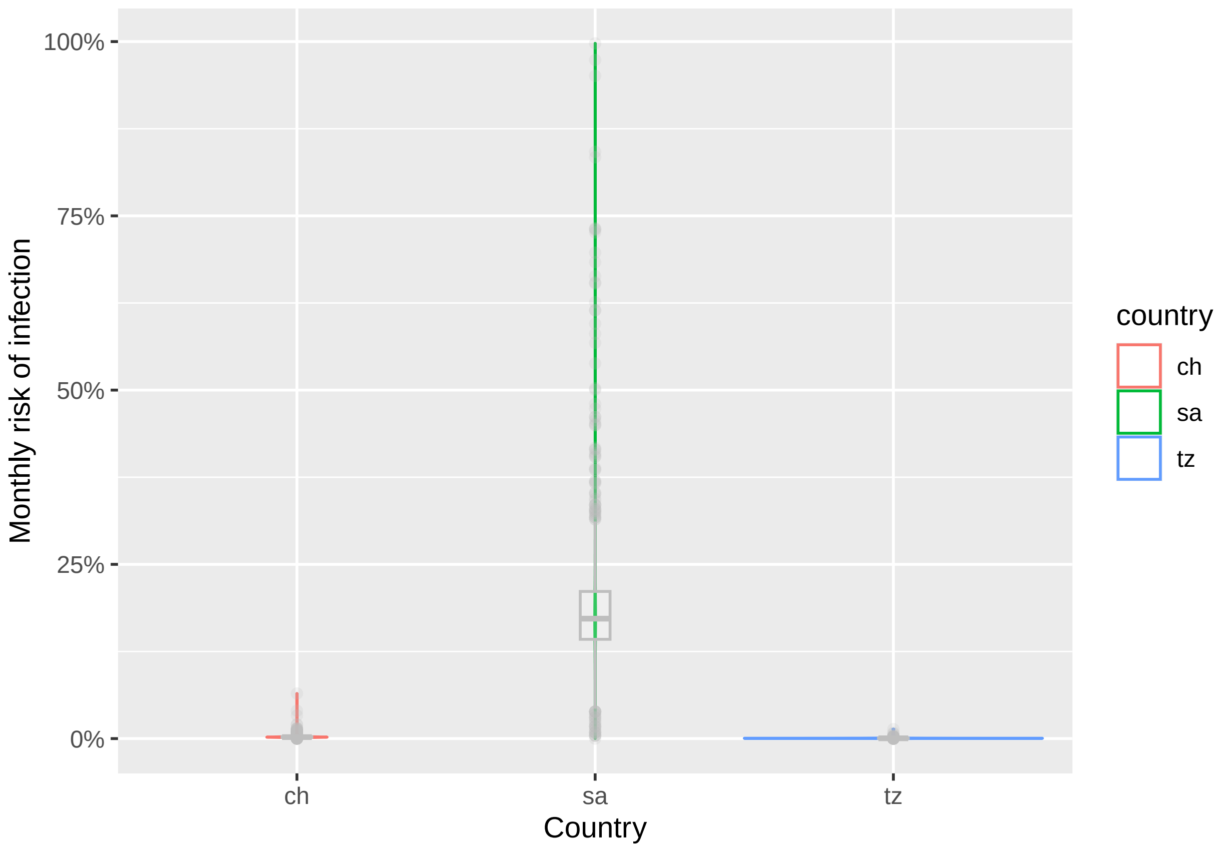
```
plot_base_day <- ggplot(df_complet, aes(x = country, y=P_day, colour = country)) +
  geom_violin(width = 1) +
  geom_boxplot(width=0.1, color="grey", alpha=0.2) +
  scale_y_continuous(labels = scales::percent_format(scale = 100)) +
  xlab("Country") +
  ylab("Daily risk of infection")

plot_base_week <- ggplot(df_complet, aes(x = country, y=P_week, colour = country)) +
  geom_violin(width = 1) +
  geom_boxplot(width=0.1, color="grey", alpha=0.2) +
  scale_y_continuous(labels = scales::percent_format(scale = 100)) +
  xlab("Country") +
  ylab("Weekly risk of infection")

plot_base_month <- ggplot(df_complet, aes(x = country, y=P_month, colour = country)) +
  geom_violin(width = 1) +
  geom_boxplot(width=0.1, color="grey", alpha=0.2) +
  scale_y_continuous(labels = scales::percent_format(scale = 100)) +
  xlab("Country") +
  ylab("Monthly risk of infection")

plot_base_year <- ggplot(df_complet, aes(x = country, y=P_year, colour = country)) +
  geom_violin(width = 1) +
  geom_boxplot(width=0.1, color="grey", alpha=0.2) +
  scale_y_continuous(labels = scales::percent_format(scale = 100)) +
  xlab("Country") +
  ylab("Yearly risk of infection")

plot_base_day
```



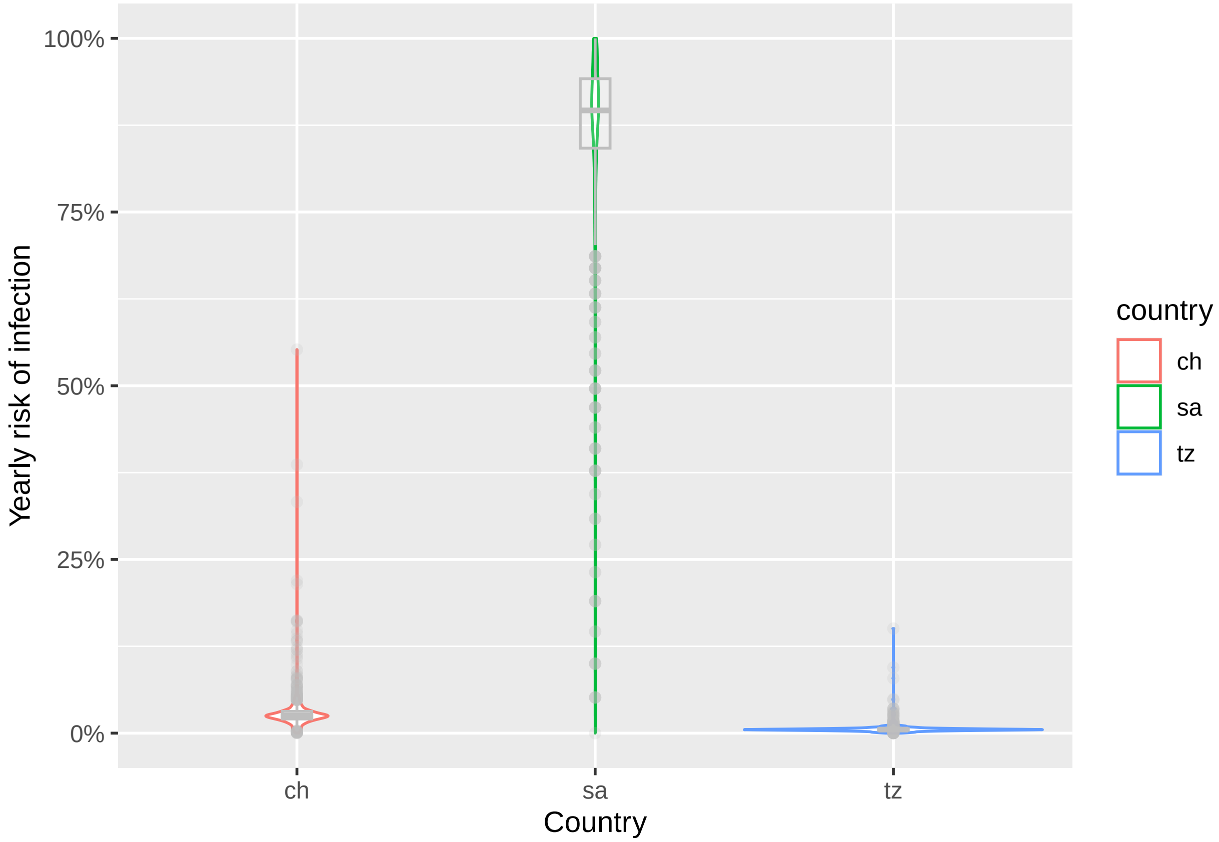
```
plot_base_week
```



```
plot_base_month
```

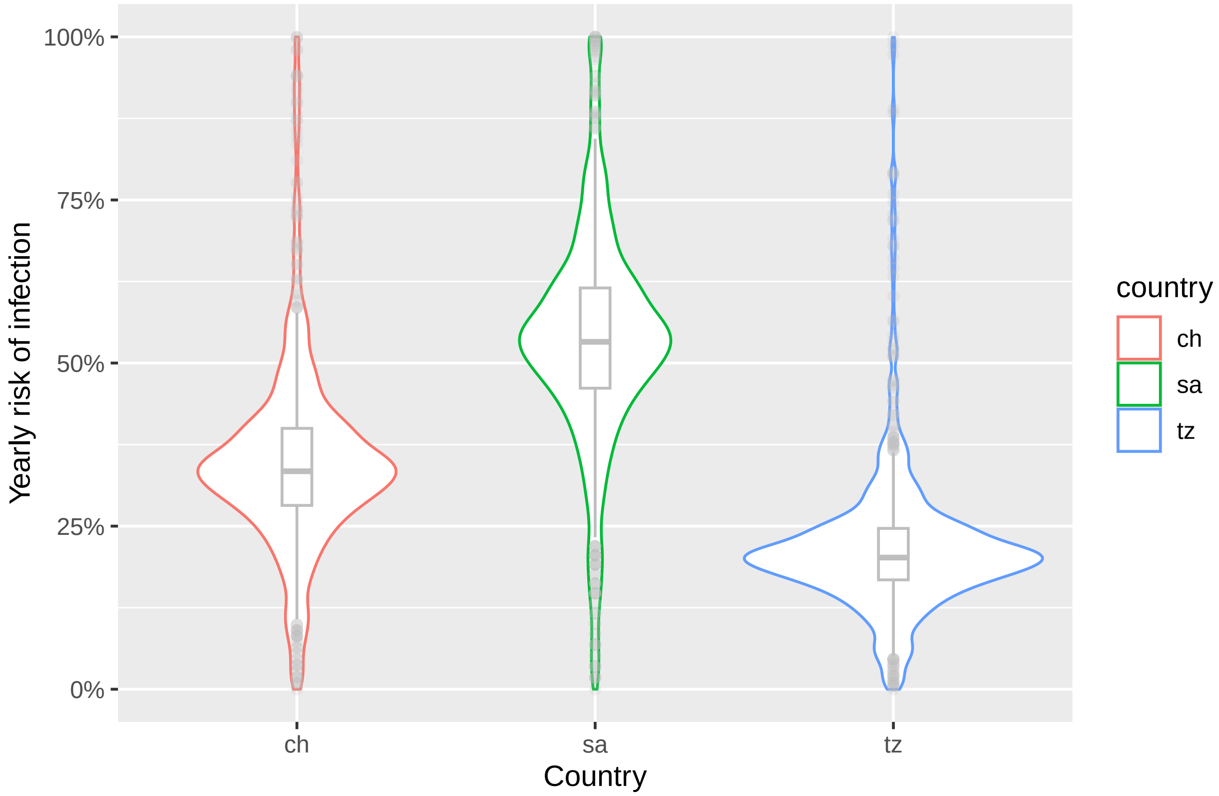


```
plot_base_year
```



Now I will compare the risks of infection, assuming that the prevalence is the same for every country and also assuming that the class size is the same. The prevalence per country is not used. This is to highlight the influence of air quality.

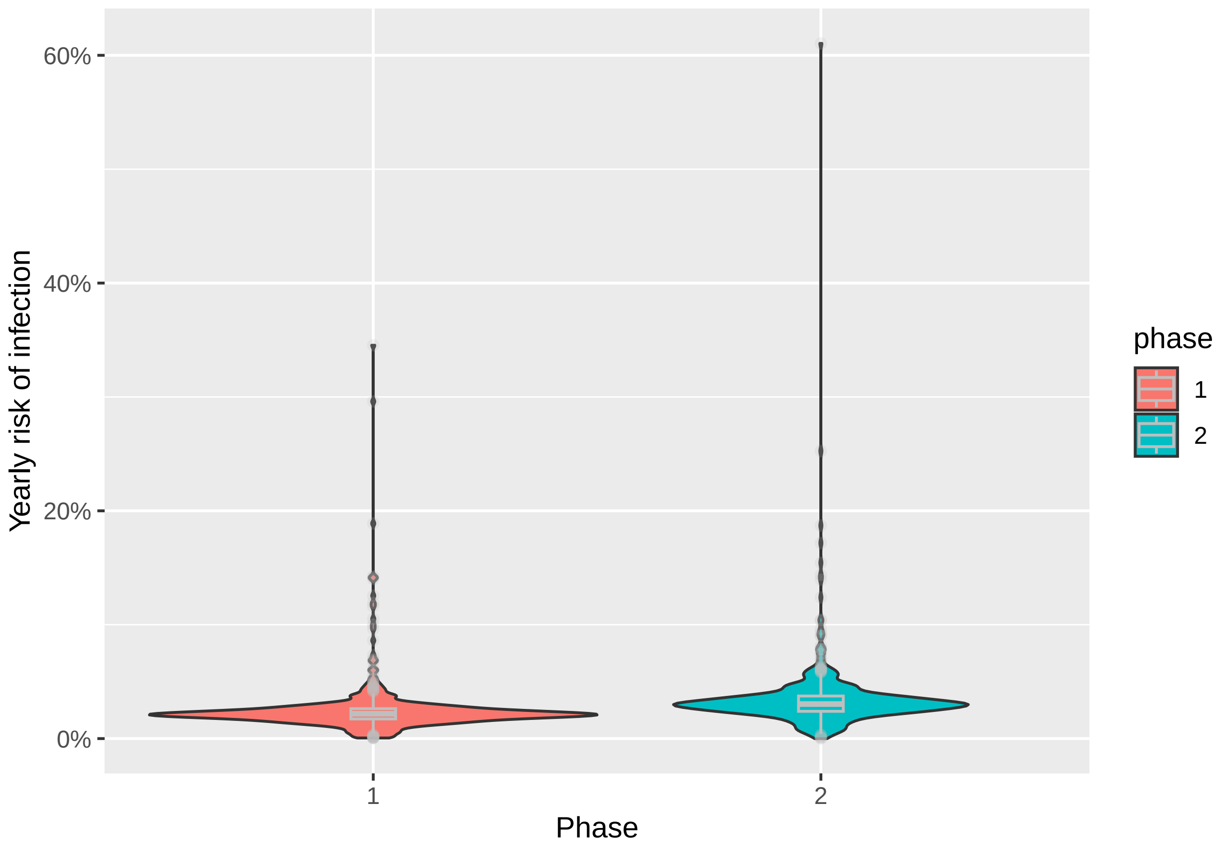
```
plot_base_year_same <- ggplot(df_complet, aes(x = country, y=P_year_one, colour = country)) +
  geom_violin(width = 1) +
  geom_boxplot(width=0.1, color="grey", alpha=0.2) +
  scale_y_continuous(labels = scales::percent_format(scale = 100)) +
  xlab("Country") +
  ylab("Yearly risk of infection") +
  ggtitle("")

plot_base_year_same
```



It can be seen that this makes the risk of infection between the three countries more balanced.

Next, I will consider the influence of time after the last break. However, this analysis only makes sense for Switzerland, since no significant differences were found in the Tanzania data, perhaps the timing for breaks is different there.

```
df_ch_breaks <- tibble(phase = c(rep("1",500), rep("2",500)), f_bar = c(rep(f_bar_pause_first,500), rep
  mutate(P_year = 1 - exp(-(f_bar*I*q*year)/n))

plot_P_year_phase <- ggplot(df_ch_breaks, aes(x = phase, y=P_year, fill = phase)) +
  geom_violin(width = 1) +
  geom_boxplot(width=0.1, color="grey", alpha=0.2) +
  scale_y_continuous(labels = scales::percent_format(scale = 100)) +
  xlab("Phase") +
  ylab("Yearly risk of infection")

plot_P_year_phase
```


60%
40%
20%
0%
Yearly risk of infection
1
2
Phase
phase
1
2