Bat virus underroost shedding model

Benny Borremans

Contents

bennyborremans@bbresearch.org

Last update: 15 Jun 2022.

The table of contents can be clicked to jump straight to specific sections.

Goal = create a model of underroost bat virus shedding.

Sheets placed below roosts collect urine from an estimated number of bats.
Urine samples on the sheet are pooled, and tested for RNA concentration using RT-PCR.
Total sample volume depend on the number and volume of urine droplets on the sheets.
The number and species of bats above the sheet are estimated, but not all bats can always be observed, and bats can move after/before observation.
Samples are stored in one of two buffers (AVL/VTM), or without buffer (NB).
Buffer type affects PCR sensitivity.

The end result (Ct value in a pooled sample) depends on all these factors,
which makes it difficult to estimate shedding prevalence in the population.

The goal of this project is two-fold:
(1) Create a simulation model of the different processes that are believed to be involved, to get better insights and build intuition.
(2) Create a model to estimate shedding prevalence in the population from the available data, capturing as much of the observation process as reasonably possible.

Simulation model

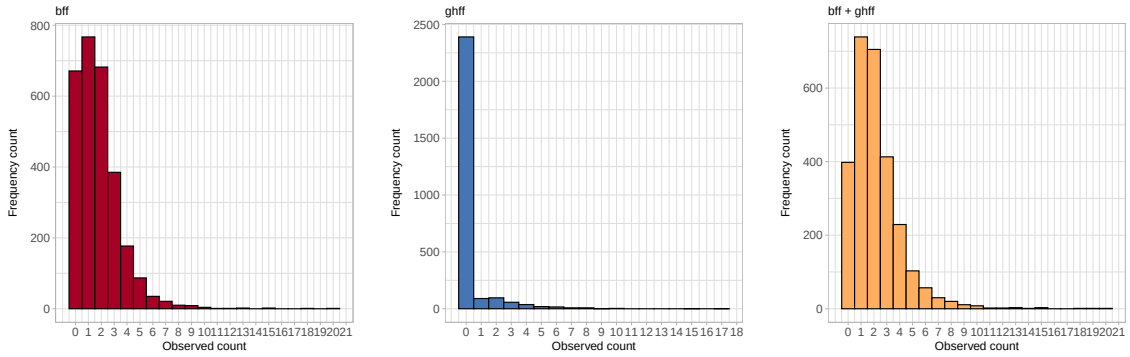
Model number of bats above a sheet

Counts and species

There are counts for 2 species, black flying foxes and grey-headed flying foxes.
Counts can be morning, afternoon, and/or overall.
Observations have a level of confidence in the count and/or species.
These confidence levels are available for bff only (do they cover both bff and ghff, or bff only?).

Using only observations with a high level of confidence.
Using only morning observations (as these seem to be the most common).
Removing observations that are not exact (e.g. "5+").
All sites pooled.

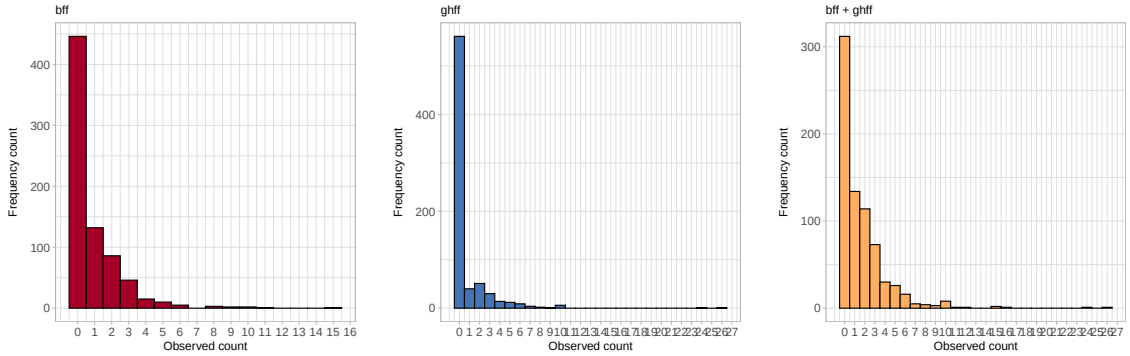
Histograms of counts:



Are higher numbers actually more rare, or just more uncertain?

==> check lower confidence counts.

Histograms of counts:



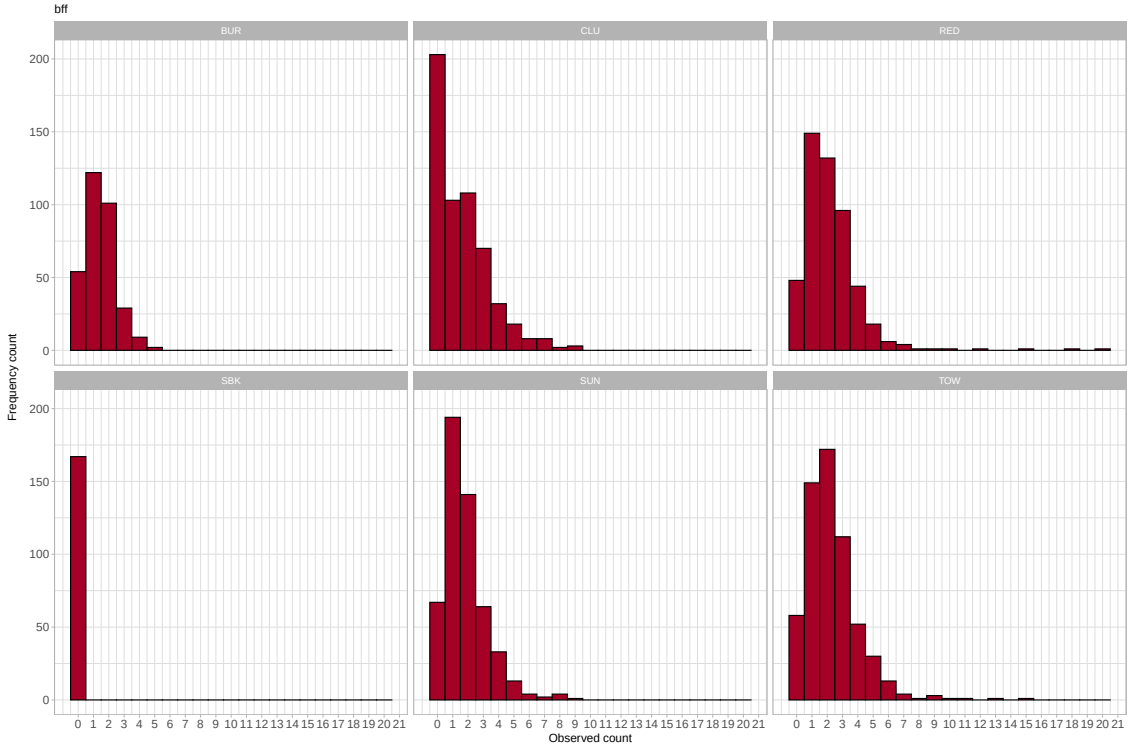
Lower-confidence counts are not higher, except for 2 ghff counts.

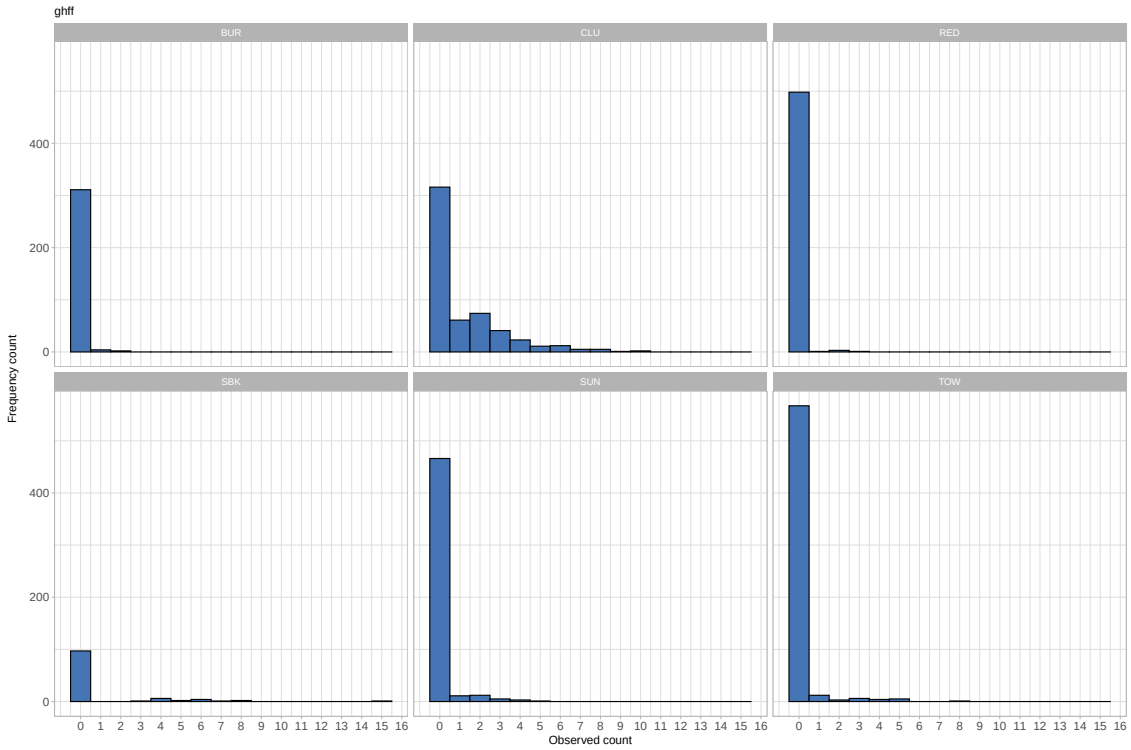
Are higher numbers actually more rare, or just written down as N+?

Not likely, there are only 25 entries with a + sign,
and these are one of: 5+, 10+, 3+, 1+.

Are there differences between sites?

Using only sites with more than 100 observations.

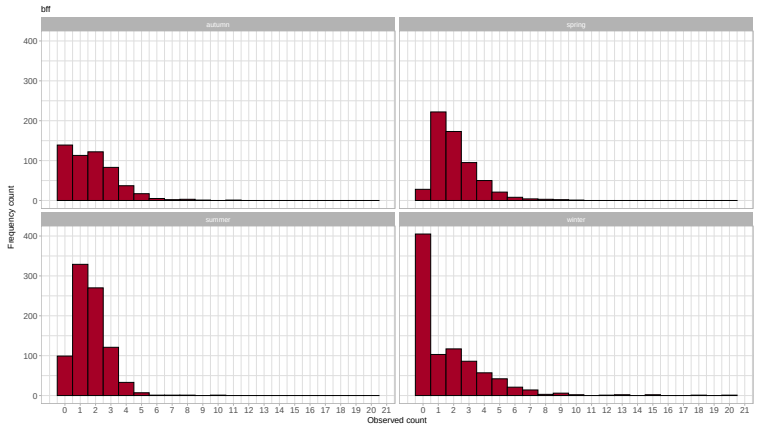




==> all look very similar, except many more 0 counts at CLU.
Any reason for this? Different conditions?
I didn't find anything in the notes, and the "bats" column mostly says "stable".

Any difference between seasons?

Histograms for different seasons.
bff only, most data available, don't need figure overload.



There seems to be an effect of season, that probably should be taken into account.

While spring and summer look like Poisson distributions,
autumn and winter seem closer to a mixture of a Bernoulli and a Poisson (probability of seeing any bats + if there are bats, how many).

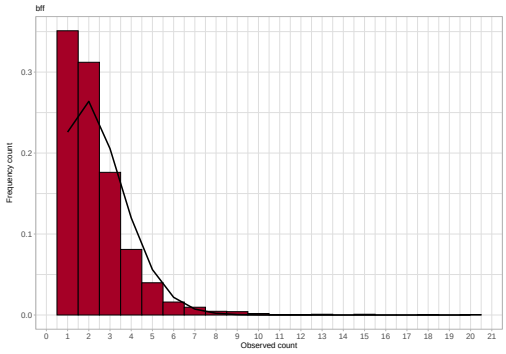
==> Fit model to positive counts only, as simulating 0 bats will not be useful.

Try Poisson distribution

All bff data pooled, across seasons and sites:

Lambda = 2.3359268.

Fitted distribution:

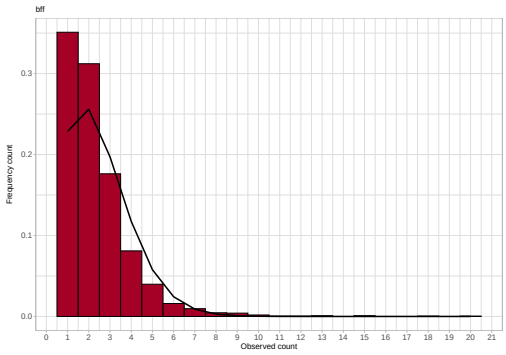


This distribution fits pretty well for the lower counts, but does not allow for the occasional higher numbers.

==> try a distribution with some more variance: negative binomial.

All bff data pooled, across seasons and sites:

Fitted distribution:

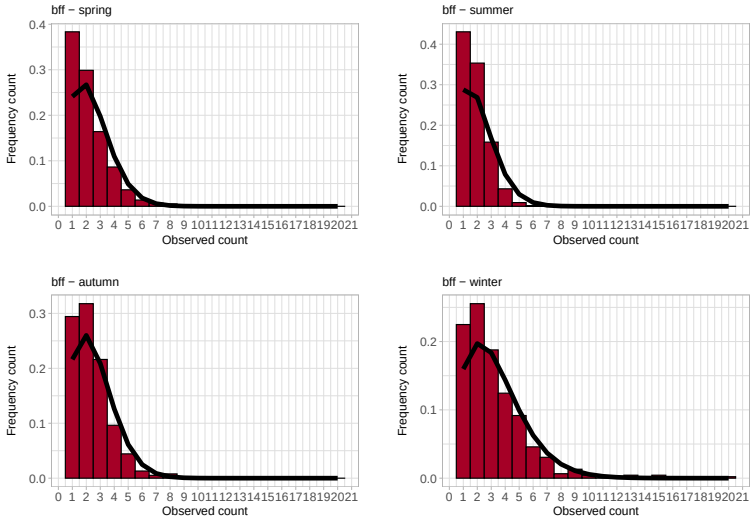


==> almost no difference, but a marginally wider tail for the negative binomial distribution.

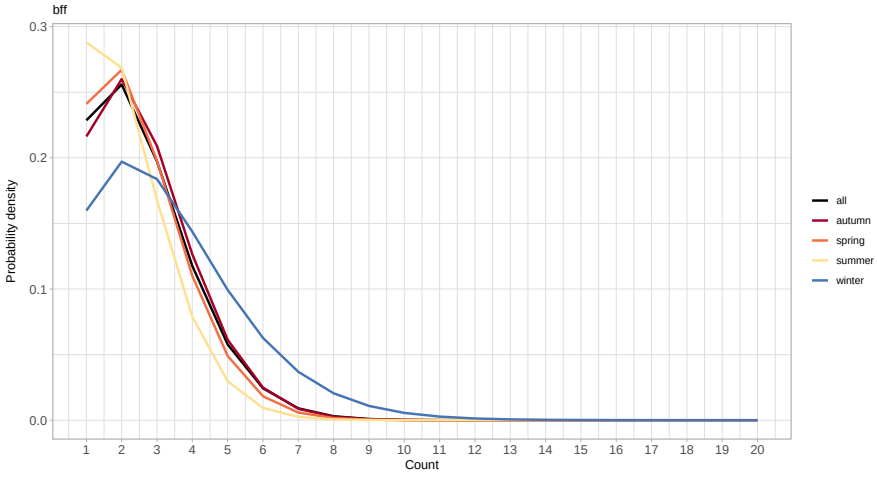
A distribution with a heavier tail would be a bit better, but for the simulation we can use the negative binomial distribution.

All bff data pooled, per season, across all sites:

Fitted distributions:



All combined:



==> fitted negative binomial distributions are very similar across seasons.

Conclusion:
Bat count can be modeled adequately using a Negative Binomial distribution, with:
$N_{bats} \sim NegBinom(size = 29.9, mu = 2.3)$

Counts and sample volume

How are sample volumne and the number of bats related?
More bats = more urine, but how strong is this correlation, and what is its shape?

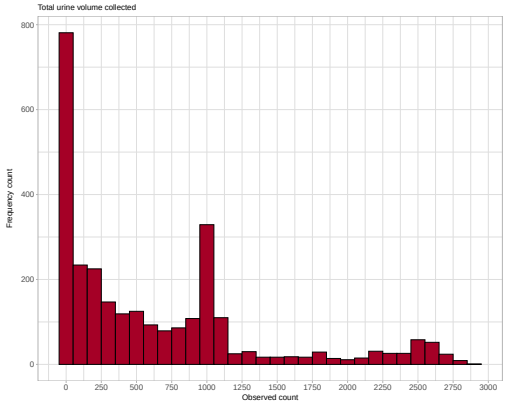
Considerations that need to be made when looking at this:
- There can be evaporation.
- The sample is added to buffer (how is this recorded in the data?)

How much urine can different numbers of bats produce?
What is the variation in collected urine volume, given a certain bat count?

And for each bat count, how does evaporation affect the distribution of volumes?

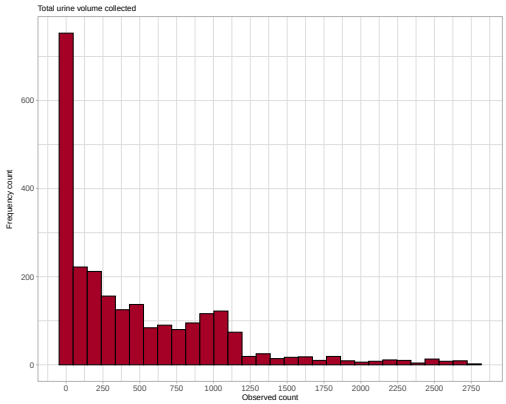
How common is evaporation in each season?

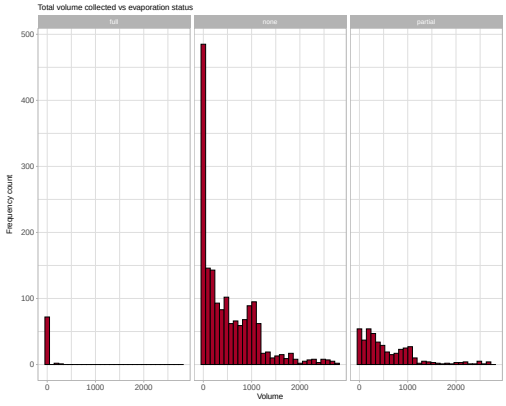
(using only confidence level = 1 data)
There are some volumes indicated as <X (e.g. < 140), which could be included when modeling by allowing censoring,
but to keep things simple these are just removed.



==> there are many zeros,
and a lot of 1000s.
Those 1000s are mostly the result of 500 for each of the vtm tubes.
Does this mean they weren't the maximum volume possible?

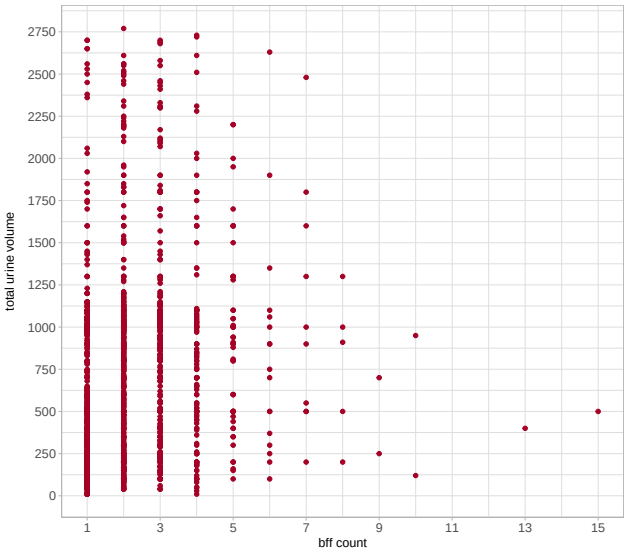
Removing these samples for now:





Correlation between bat counts and total volume.
Excluding counts of 0.
And excluding “full” evaporation, as that is always 0.

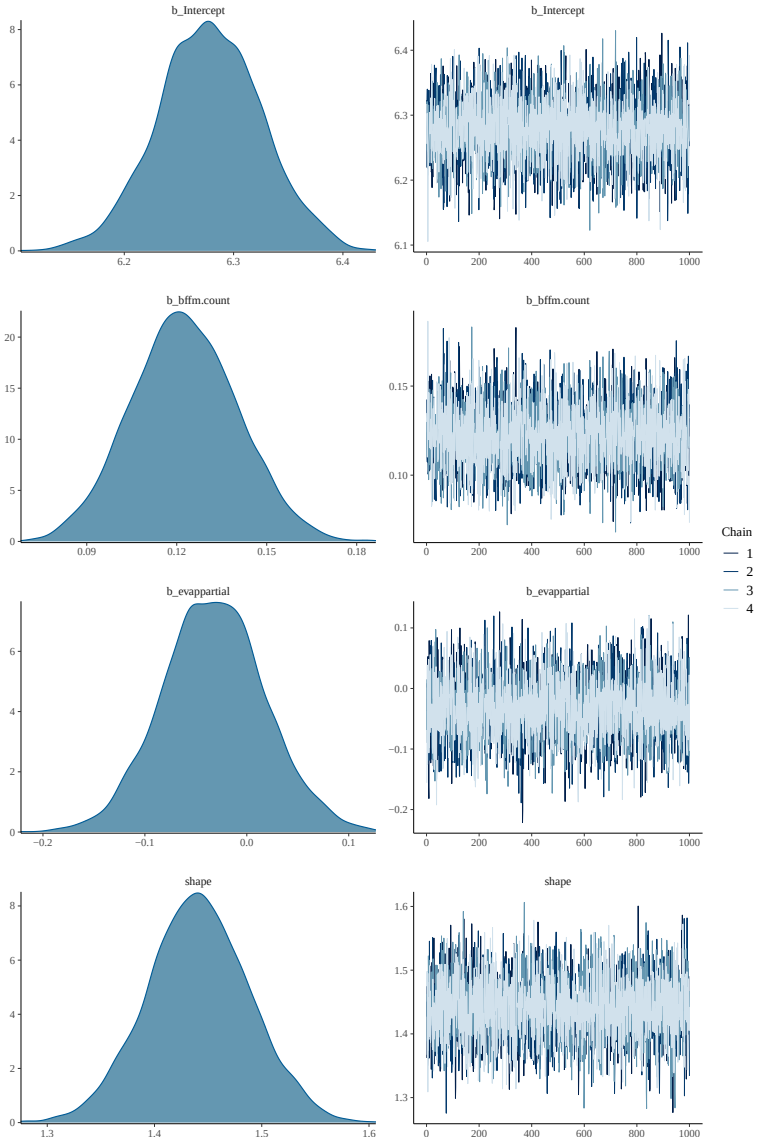
Even then, many total volumes are still 0, why is that?
Excluding total volume = 0 data for now.



Regression model:

total urine volume ~ bat count + evaporation status

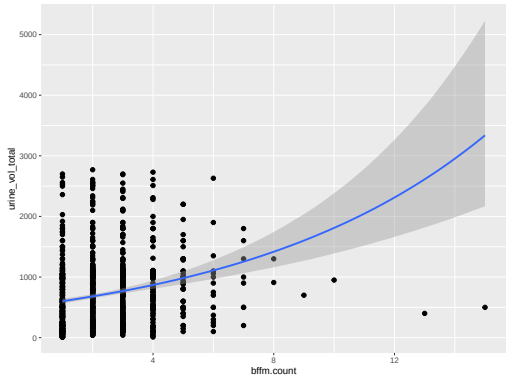
Output (log):

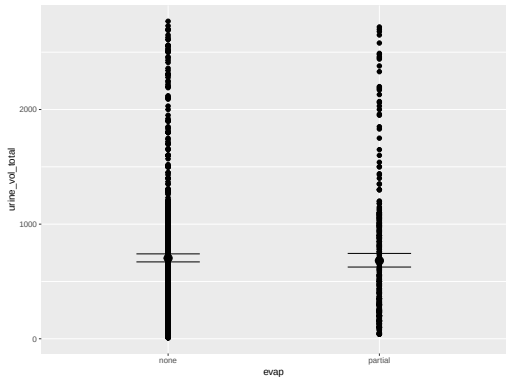


Back-transformed (exp) coefficients:

```
## [1] 532.8124962   1.1304320   0.9680188
```

Fitted functions/coefficients:



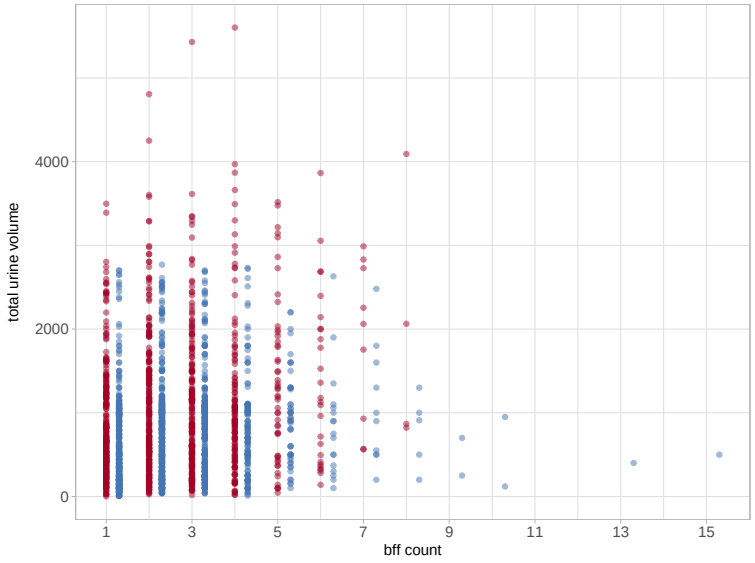


There is a positive correlation between bat count and total urine volume.
There is a minor effect of evaporation.

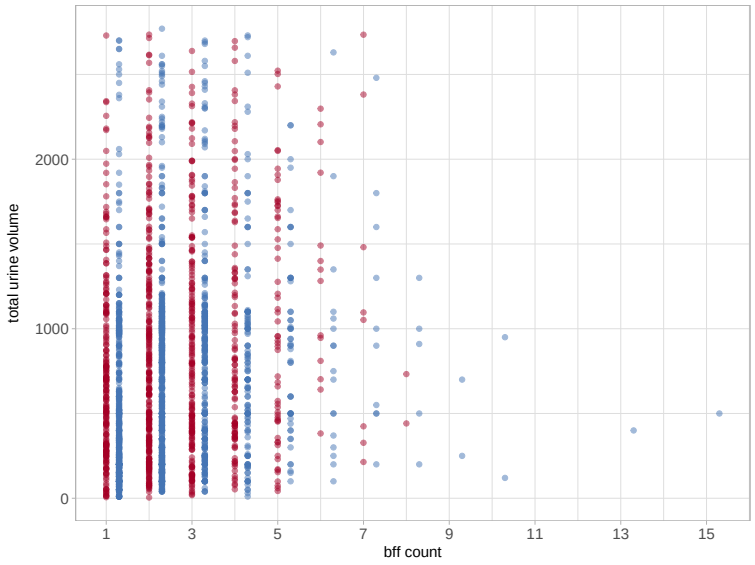
Can we use this model to adequately simulate urine volumes, given a bat count and evaporation status?

Bat counts are simulated using the negative binomial distribution fitted above (excluding 0s).
Evaporation status is simulated by randomly choosing 'none' or 'partial'.
Urine volume is then simulated using a gamma distribution with parameter values randomly selected from the model posterior.

Red = predicted, blue = observed.



Simulation output is decently similar, but the large standard deviation results in the prediction of urine volumes larger than the ones observed.
==> should be excluded:



Prevalence estimation model