SHORT DESCRIPTION OF THE ISSUE

Here I will shortly explain the issue from a statistical point of view, regardless, what is being measured and how. So I will only explain how the data looks like and what I would like to see from the data. A more detailed explanation is below the short description.

In my data, I have several different groups. Main division is in 12 different groups, since there are 12 drugs being compared. For each of these 12 drugs, there is another division in two groups, the natural behavior and the disease behavior. Within each of these two groups, there are three different experiment conditions. In the natural group, there are light, dark and alternations between the light and dark, while in the disease group, all experiments are in dark, but there are three different disease inducing substances, Apo high, Apo low and PTZ. Further, in each of these groups, there is a control group and three other groups, differing by the dosage of a novel drug being tested. And lastly, even further, in each of these groups there are 12 random subjects.

In summary, data looks like this:

For 12 different drugs being tested:																							
NATURAL												DISEASE											
LIGHT				DARK				LIGHT/DARK				APO HIGH				APO LOW				PTZ			
ctrl	d1	d2	d3	ctrl	d1	d2	d3	ctrl	d1	d2	d3	ctrl	d1	d2	d3	ctrl	d1	d2	d3	ctrl	d1	d2	d3
12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj	12 subj

So 12x2x3x4 cases for 12 subjects.

In each case, something is measured for each of the 12 subjects. Measurement is continuous for 5 minutes, followed with a 1 second pause. Measurements are done 13 times, resulting in 65 minutes of measurement with 1 second pause between them. Therefore, my exact time resolution is limited to the 13*5 minutes per case, as the measurement within the 5 minute has no time resolution, quantifications are only accumulated within the 5 minutes.

In summary, each subject has 13 longitudinal measurements, but these measurements actually contain accumulation of quantifications within 5 minutes, resulting in that within each of those 13 longitudinal measurements there are different total numbers of quantifications.

For example, the dataset per case with 12 subjects can look like this:

Subject	TimeFrame	Measurement1	…..	MeasurementN
1	1	.		.
1	1	.		.
1	2	.		.
1	2	.		.
.	.	.		.
.	.	.		.
.	.	.		.
1	13	.		.
2	1	.		.

2	1	.	.
2	1	.	.
2	1	.	.
2	2	.	.
.	.	.	.
.	.	.	.
.	.	.	.
2	13	.	.
.	.	.	.
.	.	.	.
.	.	.	.
12	13	.	.

Where the quantifications in the measurements 2-N are dependent on the quantifications in the measurement 1, through different points of view, explained in detail below.
At the same time, the quantifications in measurement 1 could be dependent on the time frame, so consequently quantifications in the measurements 2-N are influenced by time frame through quantifications in measurement 1, while the time frame could also individually influence the quantifications in measurements 2-N.
At the same, the number of quantifications in each measurement per time frame is the same between the measurements, but not the same between the time frames and the time frames influence the number of quantifications.

The data is pretty messy, not normally distributed and regarding time and measurement 1 vs measurements 2-N, it is very heteroscedastic. None of the relationships seems to linear, but some do seem to be exponential.

I would like to analyze how time directly influences the measurements 2-N, so then I would probably need to take into consideration the influence of measurement 1 itself, the effect of time on the measurement 1 and the different number of quantifications per time frame. And here is where I am having trouble, as I don't have sufficient knowledge and experience with statistics and I am not even sure if these can be modeled with statistical methods.

Below is a more detailed explanation, what is being measured and how, as well as how the data looks like in the end.

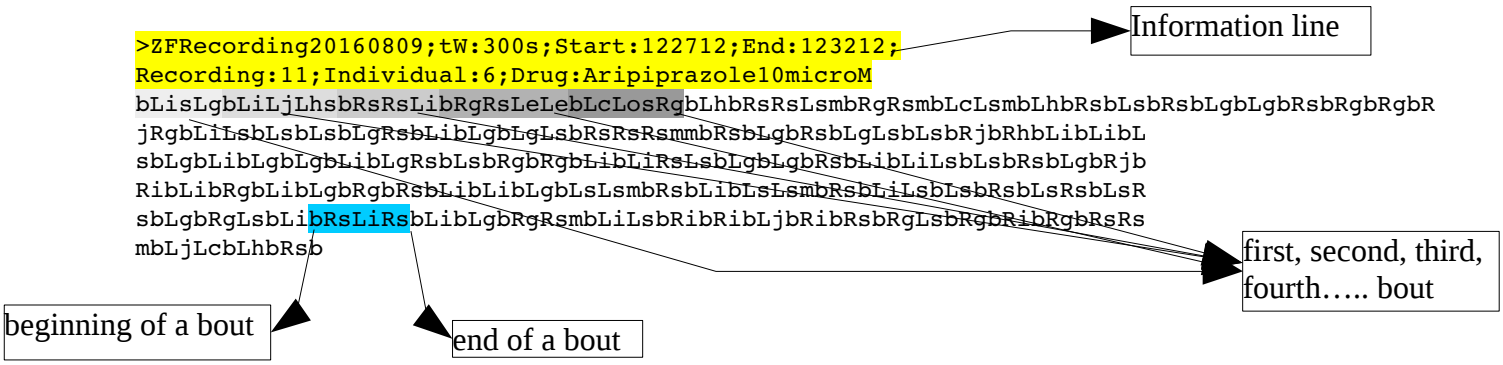
LONGER DESCRIPTION OF THE ZEBRAFISH ACTION SEQUENCE ANALYSIS AND THE ISSUES PRESENT

INTRODUCTION

The data comes from zebrafish movement. As mentioned before, in each case, there are 12 random fish, which are being recorded continuously for 5 minutes, 13 times with 1 second pause between the 5 minute recordings. Zebrafish are being recorded in a small well, while a tracking software would constantly track the zebrafish. From the output of the tracking software, action /movement sequence is produced, based on the turns/angles the zebrafish had used to move. Each 5 minute recording would

produce one action sequence per fish.
There are 8 different turn types defined and marked as s,j,c,o,e,g,h,i, a turn with no angle is also considered here(move forward). There are 3 extra auxiliary letters. Letter b separates between the bouts(more about what a bout is below) and represents a still moment between the movement, unfortunately, there is no information how long the fish was still. Letters R and L inform about the direction of the turn.

At the moment, the action sequence is constructed from bouts, which then consist of different turns. A bout is considered continuous movement of the fish. The fish is usually still, in place, then moves, a bout is recorded and then the fish is still again, and so on for 5 minutes. The bout consists of turns the fish had made, the longer the fish moved, the more turns there will be. The more times the fish moved, the greater is the number of bouts per action sequence.
Action sequences are stored in fasta files, example of one action sequence:



Several descriptive statistics can be obtained from these action sequences.
Total bout count, which is the number of b's -1.
Bout lengths, turn proportions per bout, etc.

We had observed that taking the mean descriptive statistics per action sequence is not appropriate, as we loose a lot of information, so I tried to analyze sets of descriptive statistics per action sequence, in other words, per time frame. So I constructed the dataset mentioned in the short description, where the exact variables are:

Subject TimeFrame BoutLength STurnProportion JTurnProportion CTurnProportion ….. ITurnProportion

DATA

A dataset is constructed per case, a case within the 12 drugs being tested, within either natural or disease and then within 3 conditions(light, dark, light/dark for natural and for disease either apolow, apohigh or ptz) and then within 4 experiments(control and 3 different dosages).
I can combine all the controls within the conditions, so I get a large control dataset for natural light, natural dark, natural light/dark, disease apolow, disease apohigh and disease ptz.
Then I have datasets for 12 drugs, 2*3 conditions, 3 dosages = 216 datasets.

The different conditions in natural and disease, will influence the descriptive statistics. For example, in dark, fish feel safer, so they move more in the beginning, which results in a high number of bouts in the first time frames, but after a while they will fall a sleep, so the number of bouts will go down. Under

light they will feel more stressed, scared, so they will move less, but will not fall asleep, so the number of bouts is more or less constant through all time frames. With light dark alternations, the fish don't have time to fall a sleep, so the dark periods will always be more active and the alternations between the number of bouts per time frame are substantial.
The disease inducing substance ApoLow, will sedate the fish, so they will be very inactive, while ApoHigh will first cause hyperactivity and then sedation. PTZ will cause hyperactivity, but in a special way, where the bouts become fewer, but much longer(consist of more turns).

ISSUES

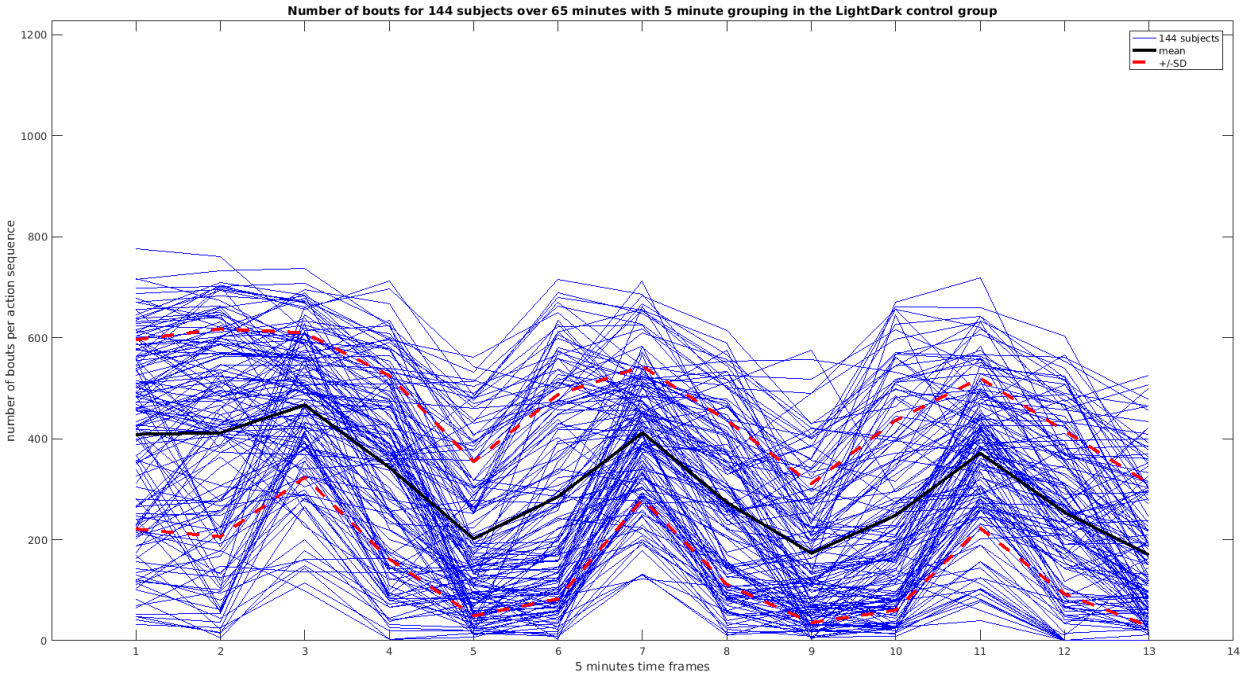
There is no time resolution within the 5 minute recording. I only have the time resolution that is 13 times 5 minute, so I get 13 time frames. Within the time frame, there can be any number of bouts and all the quantifications are made per bout. So within the time frame, quantifications are accumulated and will result in a different total number of quantifications per time frame.

If I want to analyze how the condition itself, influences the turn proportions, I face problems with mixed effects.
Bout length itself will influence the turn proportion, if the bout consists of only 1 turn, then the proportion can be 0 or 1, if the bout length is 2 turns, then the proportion can be 0,0.5 or 1. As the bouts get longer, the proportions for all 8 different turn types will stabilize, while the actually surrounding will have influence, since the fish are in a small well, so they can not move forward for a long time.
Since the progression of the condition is seen through the 13 time frames, I am trying to see how exposure in time influences the turn proportion.
But the exposure in time, per 13 time frames, can also influence the bout length and the bout length influences the turn proportion.
The exposure in time, per 13 time frames, influences the number of bouts per time frame and that has a significant influence on the turn proportions. For example, lots of short bouts will exhibit different proportions than lots of longer bouts.
So I am having difficulties in how to asses the direct influence of the condition on the turn proportion. For example, I would like to asses whether exposure to light or dark influences the turns fish preform. And same for substances ApoLow, ApoHigh and PTZ. But I do not know how to separate the effects of bout length and number of bouts per time frame.

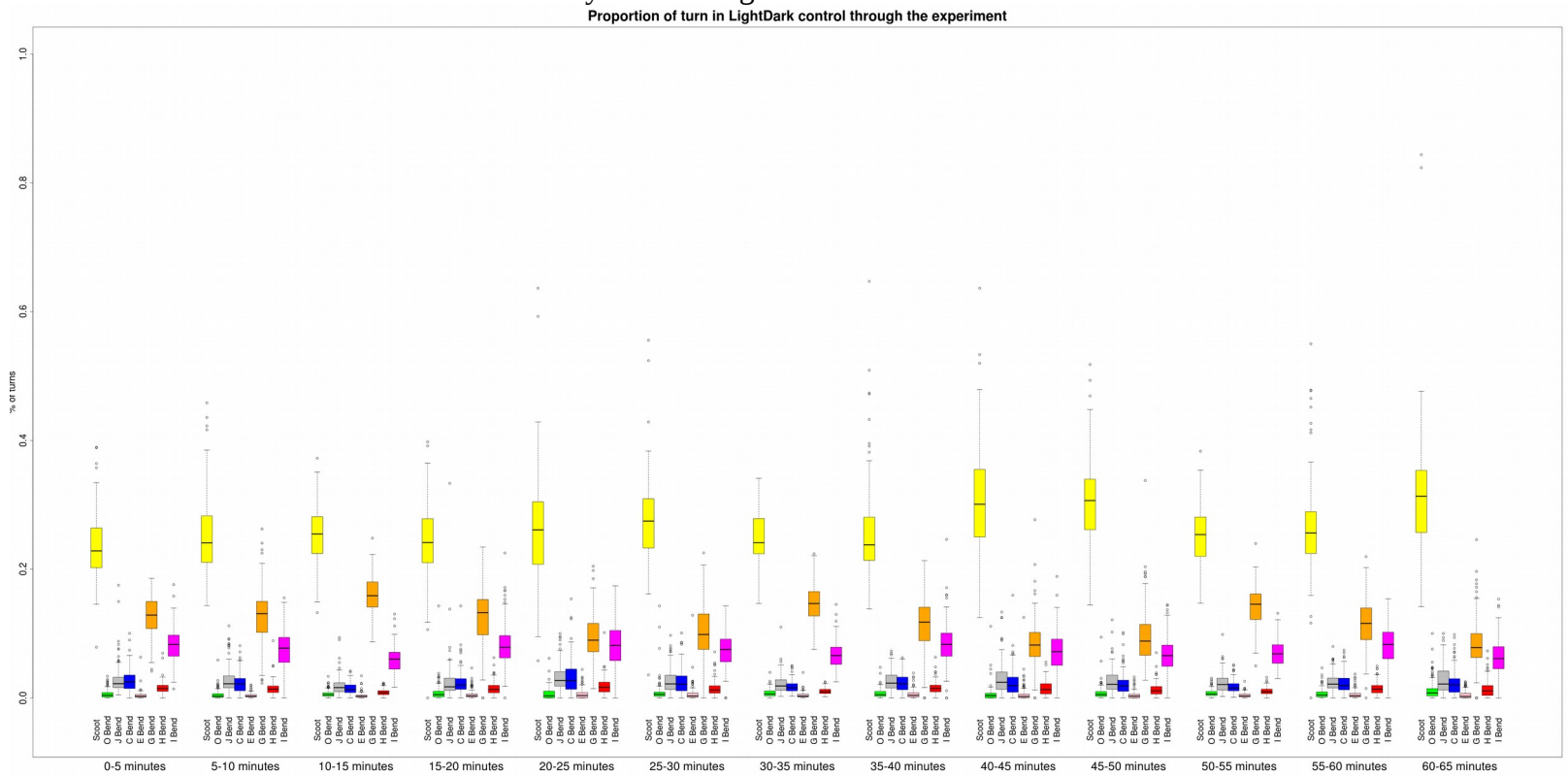
If I want to further asses how the different drugs being tested influence the turn proportion, I want to have a clear picture and model of the control. Then it will be easier to compare the 12 drugs being tested and how well they leave the natural behavior intact, while curing the disease. I might simply be able to do cluster analysis and calculate the differences/distances between the clusters, but nevertheless I feel like specific modeling should also be possible with the right statistical approaches.

As I mentioned before, the data is not pretty.
I will add some plots for example:

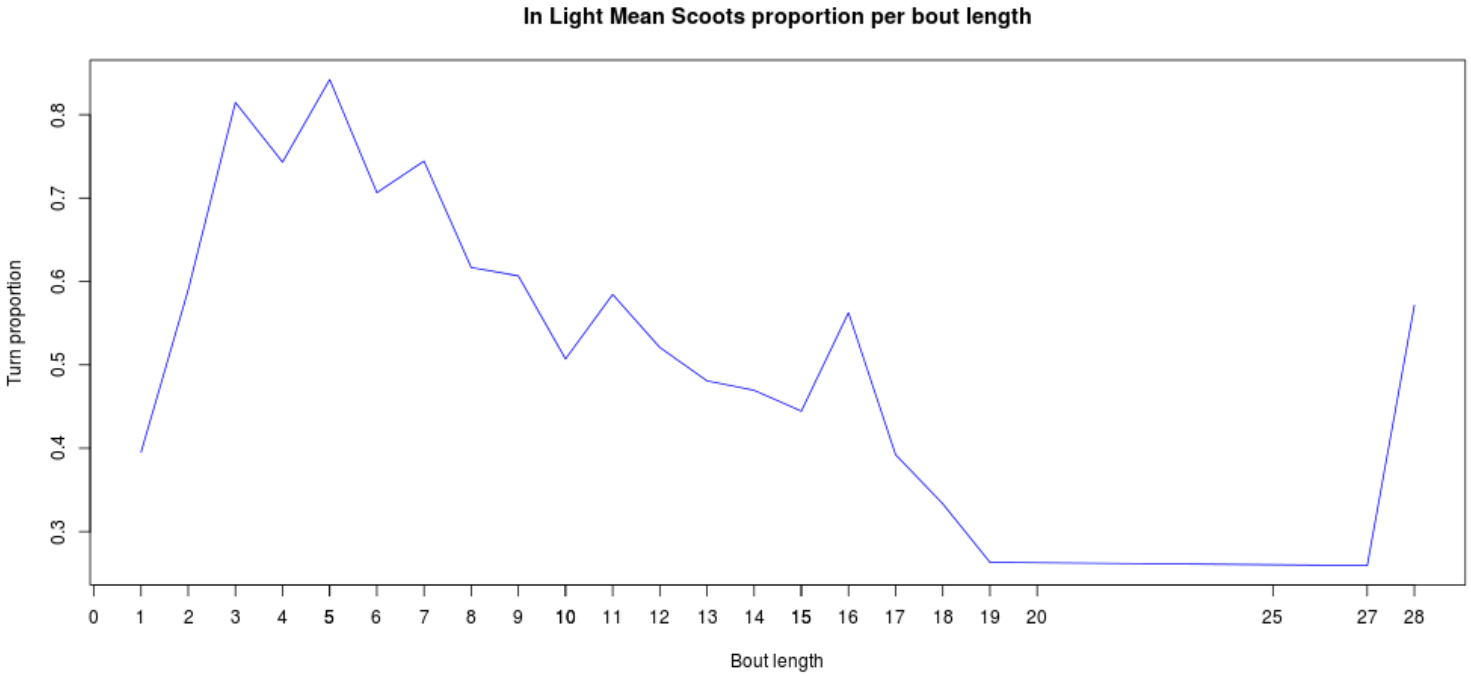
Number of bouts, within the pooled control group of the natural light/dark alternating condition, clearly shows the consequential fluctuation of the number of bouts:



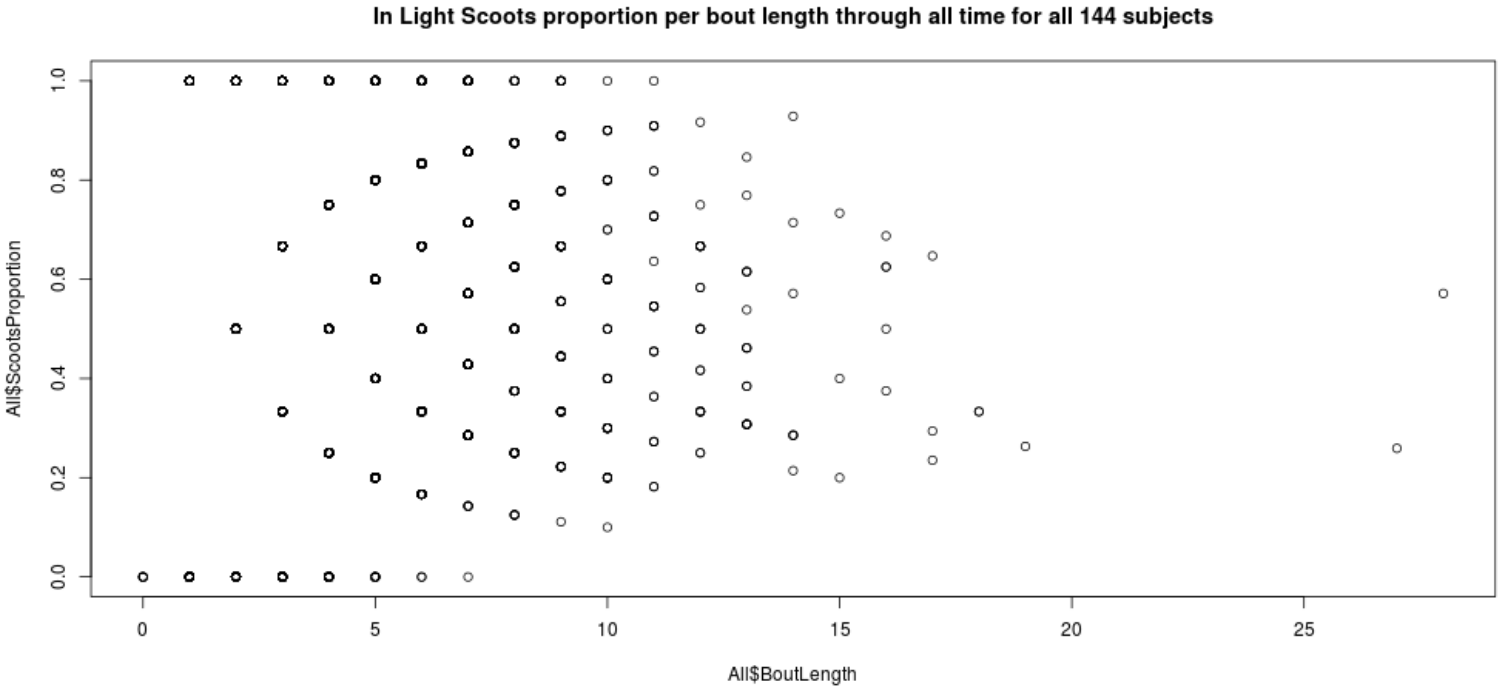
Boxplot of the turn proportions per time frame within the same pooled control group of the natural light/dark alternating condition, where it is not certain, how much of the fluctuations is caused by the different number of bouts or directly from the light/dark alternations:



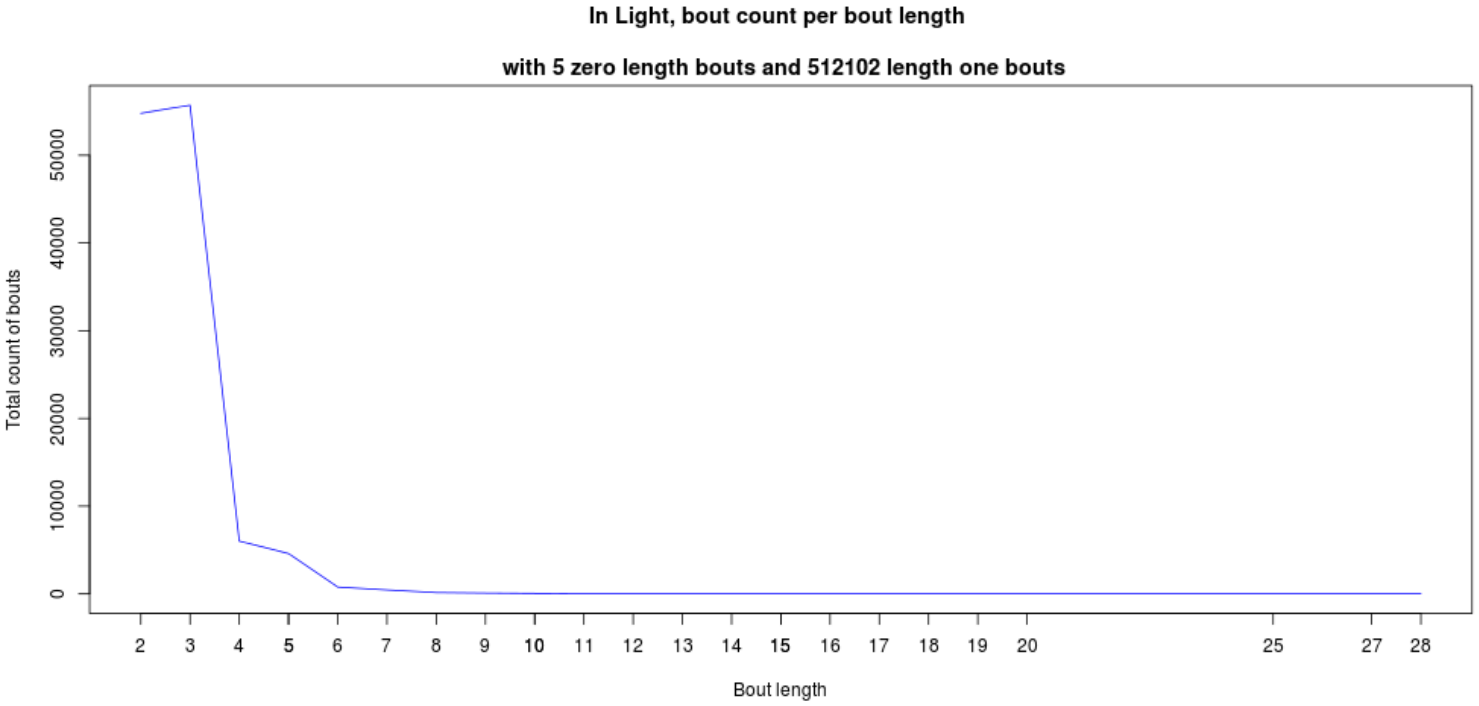
Plotting the mean of “s” turn proportions(scoots) per bout length, pooling all time in a pooled control group of the natural light condition:



While the actual scatter plot for the same thing looks like this:



And the number of bouts per bout length when pooling all time in a pooled control group of the natural light condition looks like this:



In light, these numbers will not change drastically through the time frames, since it is constant exposure to light, the number of bouts will go down gradually by around 10% only, due to fish habituating to the environment. But in other conditions, these numbers change from time frame to time frame.

IDEAS

Could I model turn proportions per bout length through time, adjusting bout length influence only for time frames, while taking into consideration the different amount of quantifications in each time frame, and then look at the residuals, to see if there is any influence left from the time frames them selves?

Or should I look directly at the influence of the time frames, adjusting for the bout length, which is to be adjusted for the time frame, while considering the different number of quantifications per time frame?

Is any of this even possible, since the turn proportions and bout lengths are not normally distributed, have unequal variance between the time frames and subjects, while the relationships are not linear? It might really be too much variation from different sources of variation. What would be the best way to separate the dataset? Should I try to model each of the 13 time frames separately and then compare them as groups instead of longitudinal measurements? I am thinking about this since in some cases there is so much variation from different sources like, habituation to environment, disease inducing substance influence, drug influence, random effects due to a small set of different subjects...

I was advised to check the limma package in R, where there is a batch effect removal function where certain components are removed, but others are kept. Could I look at my problem in a similar way? If I try to remove the effect of bout length in different batches and keeping the effect of the condition, even if the condition influences the bout length and the size of batch, and the batches are actually accumulations of quantifications in longitudinal measurements.

Any advise, example of a similar issue etc. are very welcome.