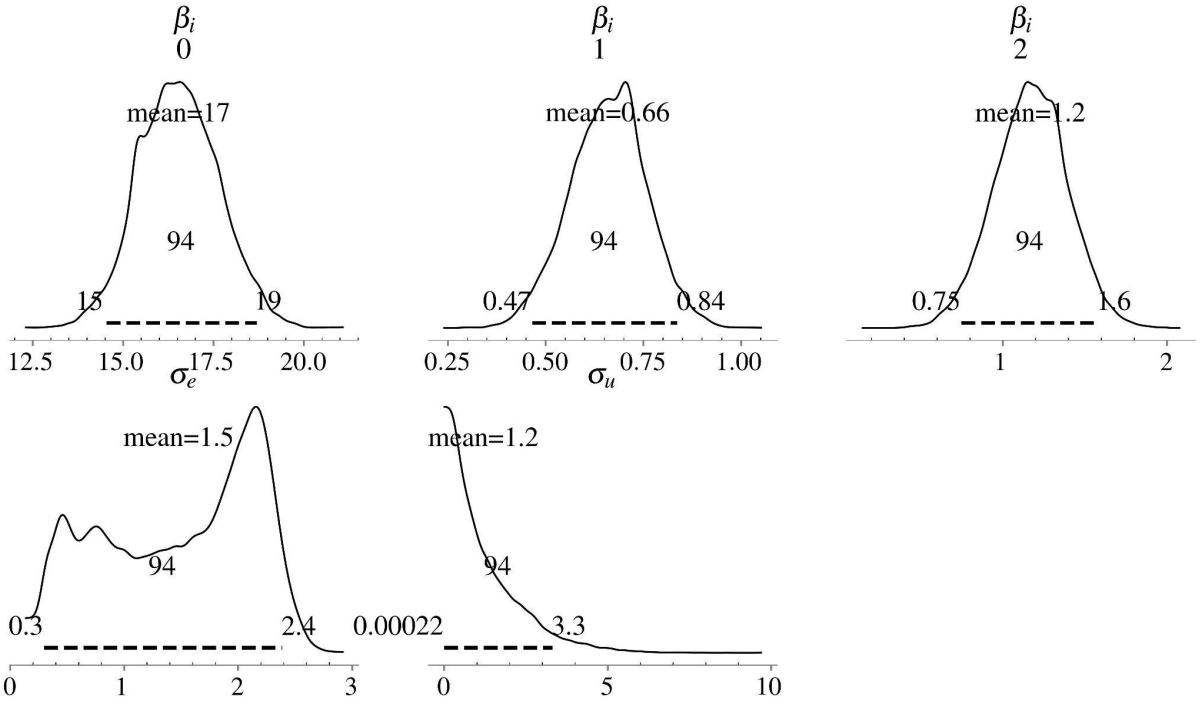
Bayesian Statistics

1. Orthodontic Distance. A longitudinal study was conducted to understand the effect of age and sex on the orthodontic distance (y). Measurements on 27 children are given in the file ortho.csv. There are a total of 16 boys and 11 girls, which are identified in the dataset using the column Subject. Consider the following random effects model:

$$y_{ij} \mid \beta_0, \beta_1, \beta_2, u_i, \sigma_e^2 \sim\sim^{ind.} \quad N\left(\beta_0 + \beta_1 age_{ij} + \beta_2 \,\text{sex}\, x_i + u_i, \sigma_e^2\right),$$
$$u_i \mid \sigma_u^2 \sim \text{ iid } N\left(0, \sigma_u^2\right),$$

a) Fit the model ignoring the random effects. Plot the posterior densities of the four parameters β0β0, β1β1, β2β2, σϵσϵ. What differences do you see from the previous analysis using random effects (compare posterior means and credible intervals of the four parameters?

b) Posterior distributions for the requested parameters were calculated using PyMC3. Results are below:

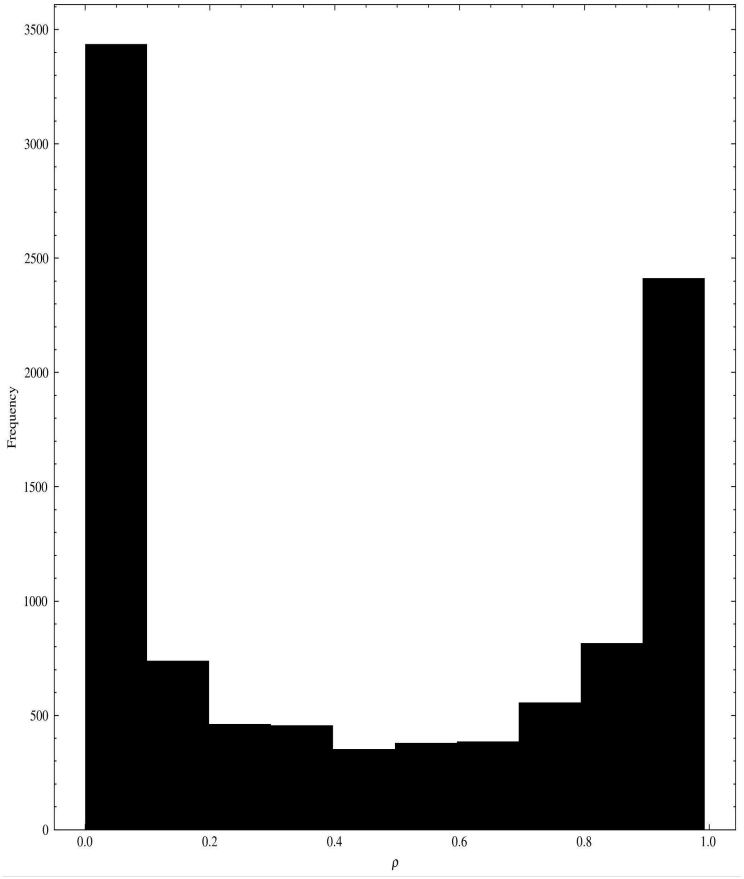


From this plot one may infer the following:

- The fact that β_0 is far greater than the other parameters imply that there is a baseline Orthodontic distance which is independent from both age and sex and bears greater importance.
- The effect of age and sex are almost indistinguishable since both parameters contain each other in their confidence intervals.

- Random effects didn't add anything new to this analysis as shown in the posterior of their standard deviation almost didn't move from the prior implying almost no random effects. The fact that their prior didn't update might be evidence their inclusion didn't bring new knowledge to the model.

c) Intra-class correlation coefficient histogram:



This Histogram shows that the intra-class correlation was most of the time 0 which doesn't support the hypothesis that it was actually significantly distinct from 0. Giving a confidence interval for this variable would be misleading since it shows there is complete intra-class correlation or none most of the time (Boundaries being 0 and 1).

d) The previous model was fitted without the random effects and tables presenting the results are presented below.
With Random Effects:

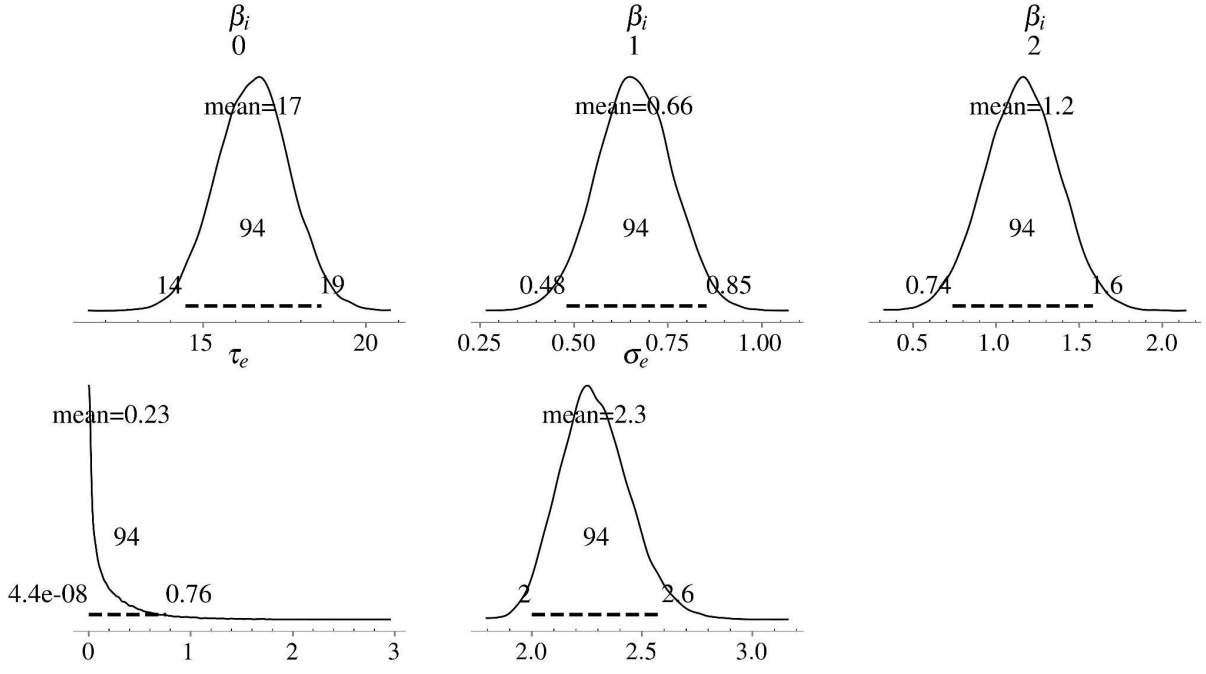
	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
$\beta_i[0]$	16.521	1.161	14.399	18.739	0.043	0.03	822.0	346.0	1.01
$\beta_i[1]$	0.663	0.104	0.472	0.859	0.004	0.003	754.0	308.0	1.01
$\beta_i[2]$	1.163	0.222	0.738	1.577	0.002	0.001	11661.0	22719.0	1.0
τ_e	0.931	2.194	0.0	3.67	0.166	0.117	462.0	202.0	1.01
τ_u	2.263	4.316	0.119	9.326	0.181	0.128	298.0	426.0	1.02
σ_e	1.594	0.658	0.376	2.48	0.042	0.03	242.0	66.0	1.04

Without Random effects:

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
$\beta_i[0]$	16.543	1.112	14.455	18.658	0.006	0.004	40327.0	49289.0	1.0
$\beta_i[1]$	0.66	0.099	0.474	0.848	0.0	0.0	40280.0	48362.0	1.0
$\beta_i[2]$	1.161	0.224	0.743	1.583	0.001	0.001	76292.0	61983.0	1.0
τ_e	0.225	0.293	0.0	0.756	0.001	0.001	40744.0	27520.0	1.0
σ_e	2.286	0.16	1.992	2.592	0.001	0.0	66606.0	62648.0	1.0

Confidence intervals for the regression parameters haven't been affected by the presence of random effects. As it was posed above, random effects don't seem to incorporate new knowledge into the model that weren't explained by the previous parameters.

Distribution for the parameters after the random effects are turned off appear below:



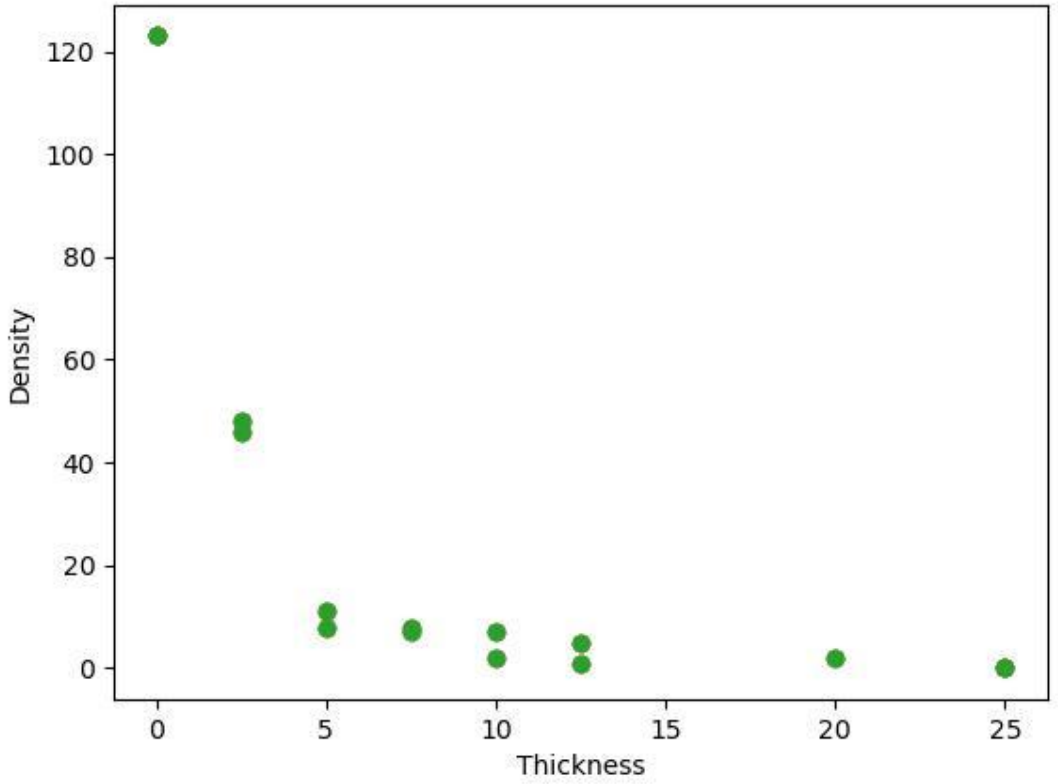
2. **Nanowire density**. Consider the problem of predicting the density of nanowires y with respect to the thickness of polymer films x in a solution-based growth process (see Figure 1). Eight experiments were conducted with two replicates (except for one run). The data are in the file nanowire.csv. The density of nanowires is assumed to follow a Poisson distribution with mean:

$$\mu(x) = \theta_1 \exp\left(-\theta_2 x^2\right) + \theta_3\left\{1 - \exp\left(-\theta_2 x^2\right)\right\}\Phi(-x/\theta_4)$$

where $\Phi(\cdot)$ is the standard normal CDF - note that there is a phi () function in BUGS for this, and in pyme you may use the invprobit() function. Assume the following prior distribution for the parameters:

$$\begin{aligned} \log\theta_1, \log\theta_3, \log\theta_4 &\sim^{iid} N(0, \sigma^2 = 10) \\ \theta_2 &\sim U(0,1). \end{aligned}$$

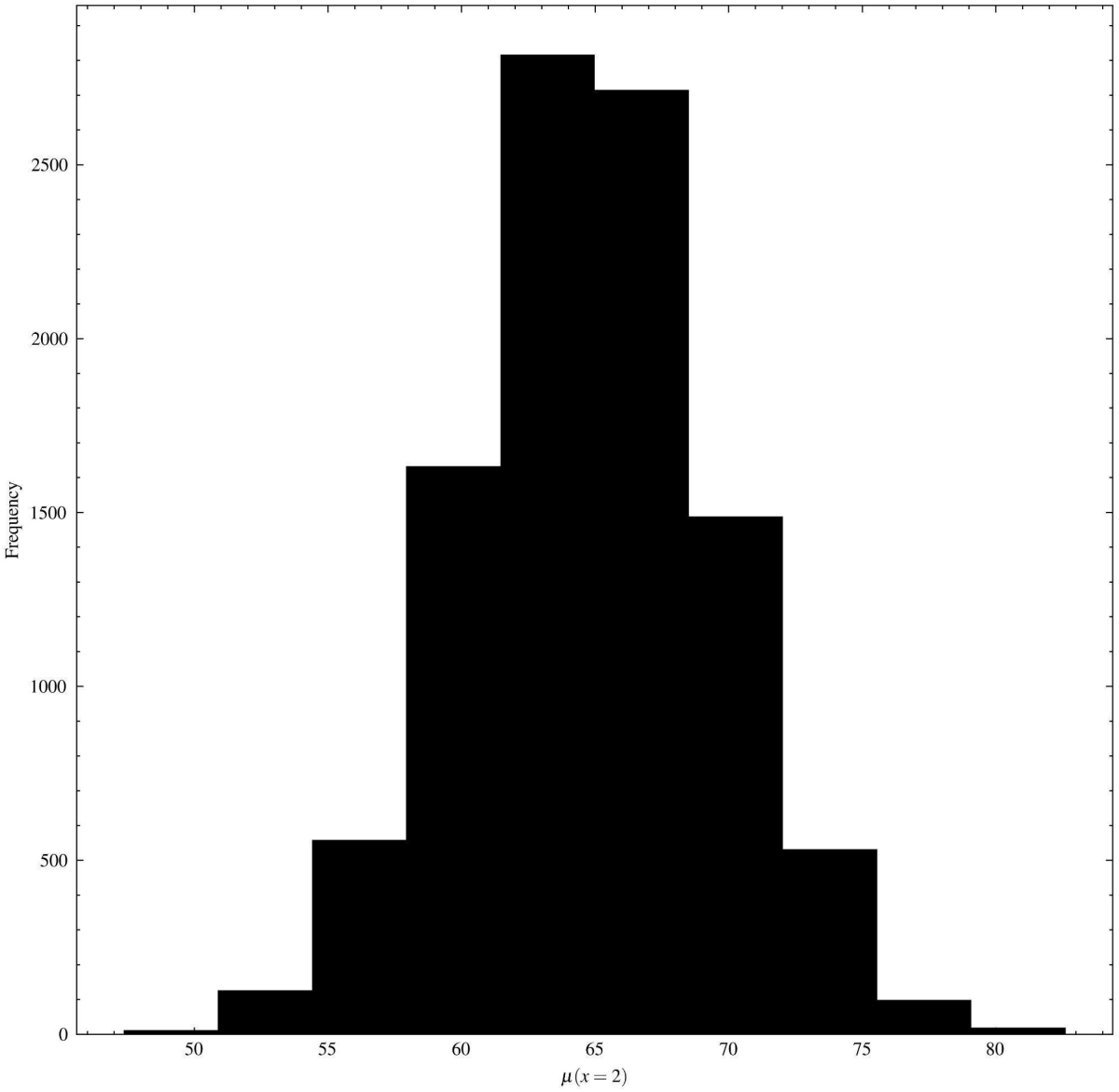
A plot of the data:



a) The model was implemented in PyMC3 and the resulting summaries with the according intervals are shown below:

	mean	sd	hdi 3%	hdi 97%	mcse mean	mcse sd	ess bulk	ess tail	r hat
θ_1	123.019	7.853	108.59	138.029	0.084	0.059	8702.0	8259.0	1.0
θ_3	27.608	7.848	14.02	42.228	0.107	0.076	5476.0	5767.0	1.0
θ_4	11.473	2.881	7.318	15.875	0.038	0.027	5774.0	6080.0	1.0
θ_2	0.186	0.027	0.135	0.235	0.0	0.0	7396.0	7356.0	1.0
$\mu[0]$	123.019	7.853	108.59	138.029	0.084	0.059	8702.0	8259.0	1.0
$\mu[1]$	123.019	7.853	108.59	138.029	0.084	0.059	8702.0	8259.0	1.0
$\mu[2]$	46.518	4.615	37.614	54.968	0.043	0.031	11262.0	10155.0	1.0
$\mu[3]$	46.518	4.615	37.614	54.968	0.043	0.031	11262.0	10155.0	1.0
$\mu[4]$	10.091	1.599	7.141	13.083	0.019	0.013	7437.0	7660.0	1.0
$\mu[5]$	10.091	1.599	7.141	13.083	0.019	0.013	7437.0	7660.0	1.0
$\mu[6]$	6.619	1.099	4.621	8.719	0.012	0.009	7881.0	8930.0	1.0
$\mu[7]$	6.619	1.099	4.621	8.719	0.012	0.009	7881.0	8930.0	1.0
$\mu[8]$	4.808	0.753	3.458	6.261	0.007	0.005	13260.0	11340.0	1.0
$\mu[9]$	4.808	0.753	3.458	6.261	0.007	0.005	13260.0	11340.0	1.0
$\mu[10]$	3.382	0.67	2.089	4.584	0.006	0.004	12352.0	10053.0	1.0
$\mu[11]$	3.382	0.67	2.089	4.584	0.006	0.004	12352.0	10053.0	1.0
$\mu[12]$	1.001	0.535	0.151	1.972	0.006	0.005	6903.0	7419.0	1.0
$\mu[13]$	0.406	0.36	0.005	1.055	0.005	0.003	6355.0	7024.0	1.0
$\mu[14]$	0.406	0.36	0.005	1.055	0.005	0.003	6355.0	7024.0	1.0

b) The posterior distribution when x=2 was obtained using the Deterministic wrapper:



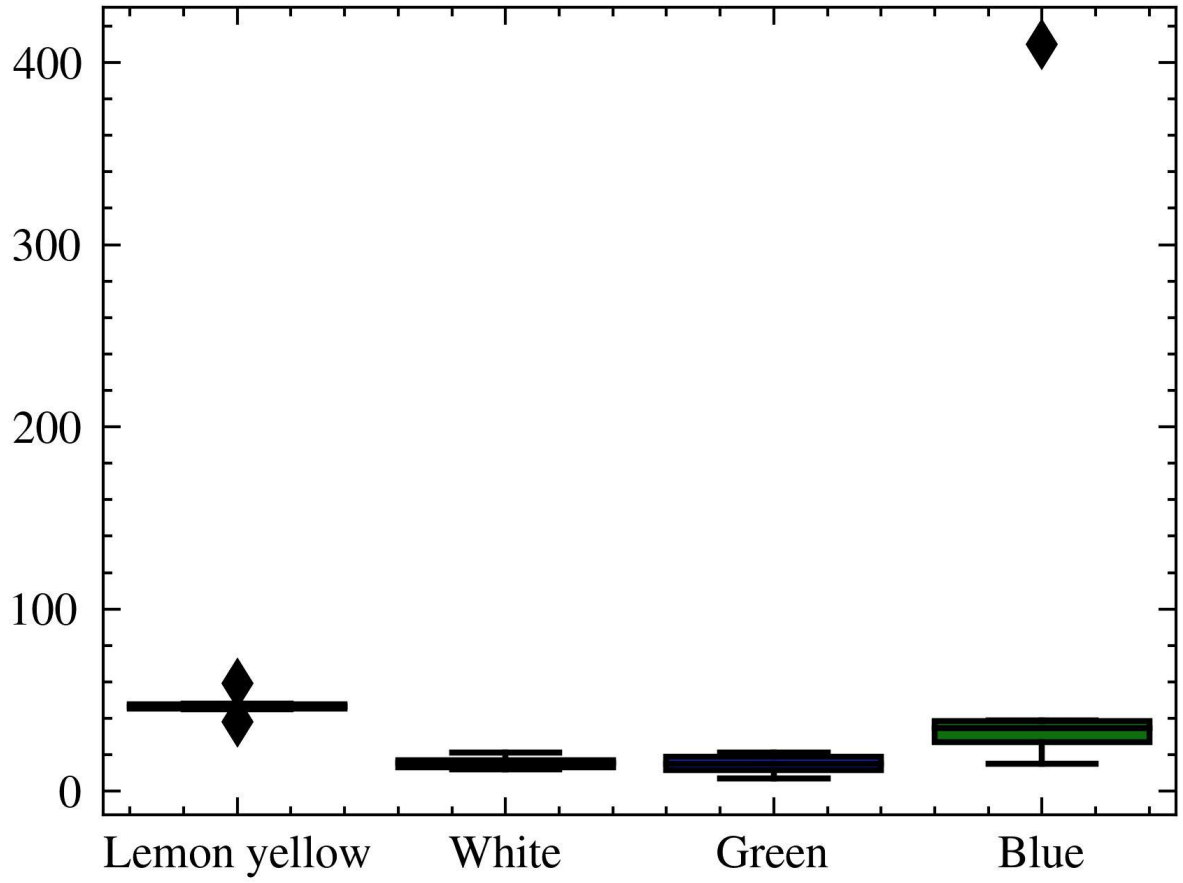
Since all parameters are random variables then it stands that the response value at a single point i.e., X=2, is a probability distribution. The histogram represents draws from this distribution. It's possible to derive from this distribution that there is appreciable scatter in the response for the model at X=2. This is mostly because of the sharp gradient present at that point and the fact that the estimate is based on the surrounding 2 data points.

3) **Color Attraction for Oulema melanopus.** Some colors are more attractive to insects than others. Wilson and Shade $(1967)^{1}$ conducted an experiment aimed at determining the best color for attracting cereal leaf beetles (Oulema melanopus). Six boards in each of four selected colors (lemon yellow, white, green, and blue) were placed in a field of oats during summer time. The following table (modified

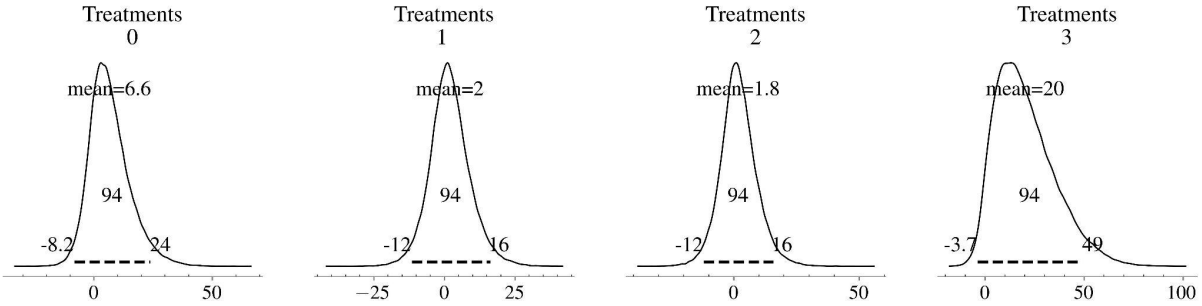
from Wilson and Shade, 1967) gives data on the number of cereal leaf beetles trapped:

Board color	Insects trapped					
Lemon yellow	45	59	48	46	38	47
White	21	12	14	17	13	17
Green	16	11	20	21	14	7
Blue	37	32	15	25	39	41

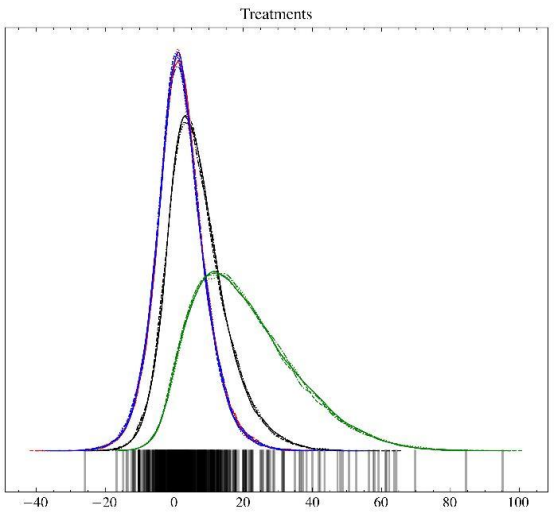
a) Bayesian ANOVA analysis was done using PyMC3. The STZ constraint was done by imposing a zero mean to the treatments as advised by Krutchske. Results are presented below:
First, the boxplots for the different colors:

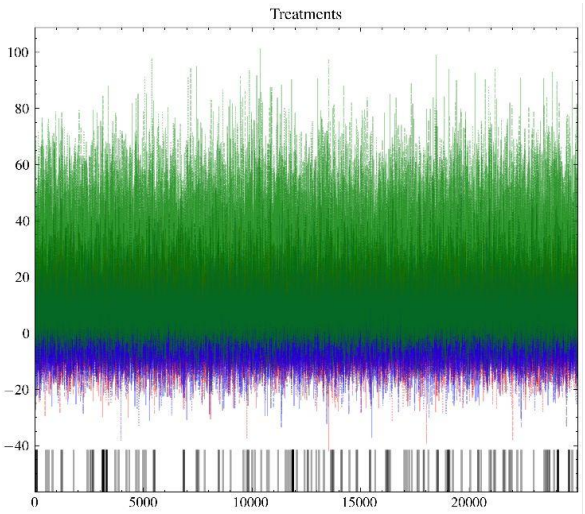


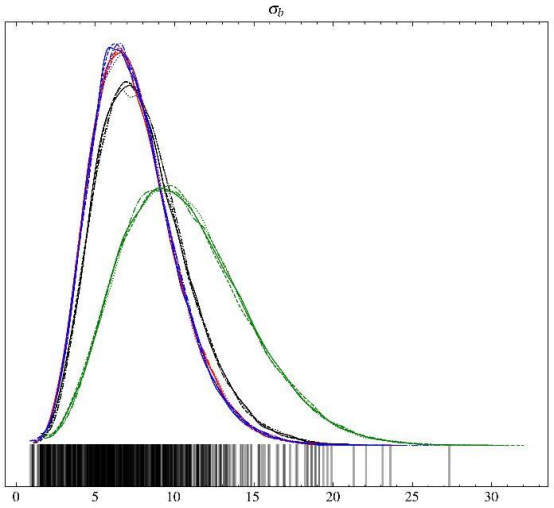
The posterior plots for the treatments:

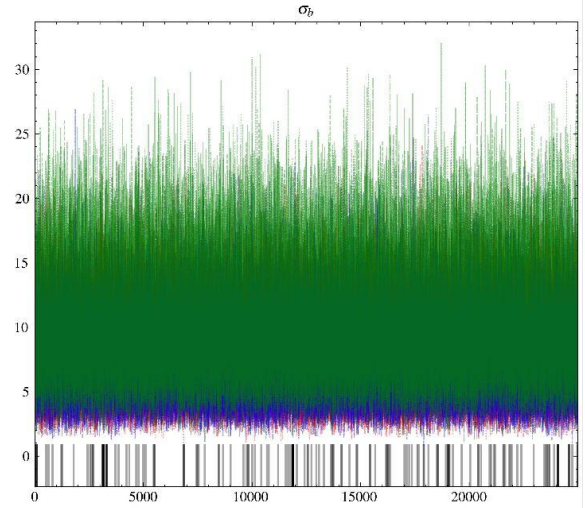


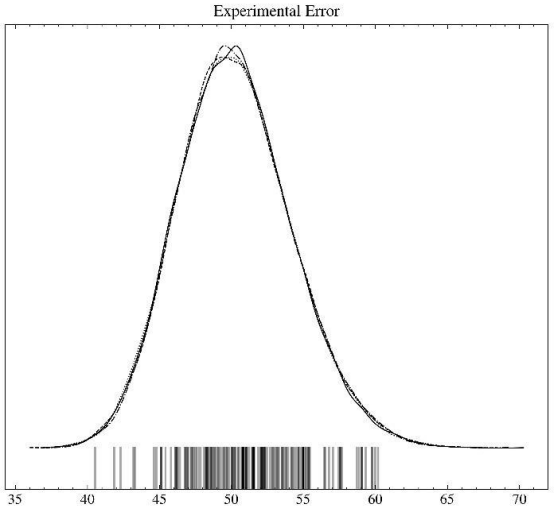
For each class a standard deviation for the treatment and 0 mean were imposed. This is shown by the different KDE plots that appear in the Treatments and sigma_b plots. The posterior distribution and sampling process for each variable are presented. This plot is mainly here for the diagnosis of the model. We can see that the sampling process was uniform i.e., for all parameters, no region of low density appears (implying that the sampler was caught there). Another detail to be noticed is that the experimental error is comparable to the value of the treatments. This implies that the experiment itself had troubles.

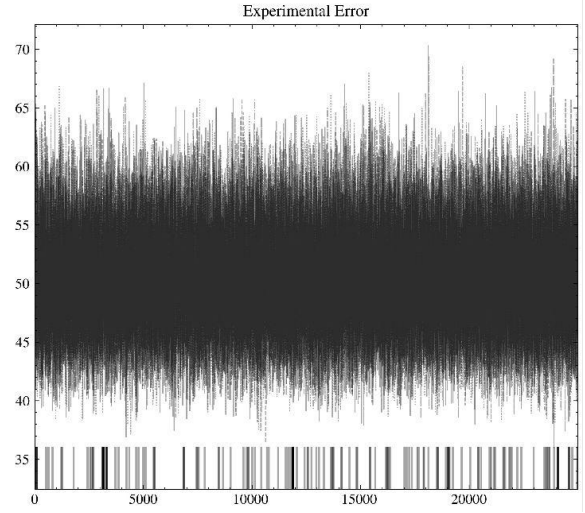

Treatments
−40
−20
0
20
40
60
80
100


Treatments
100
80
60
40
20
0
−20
−40
0
5000
10000
15000
20000

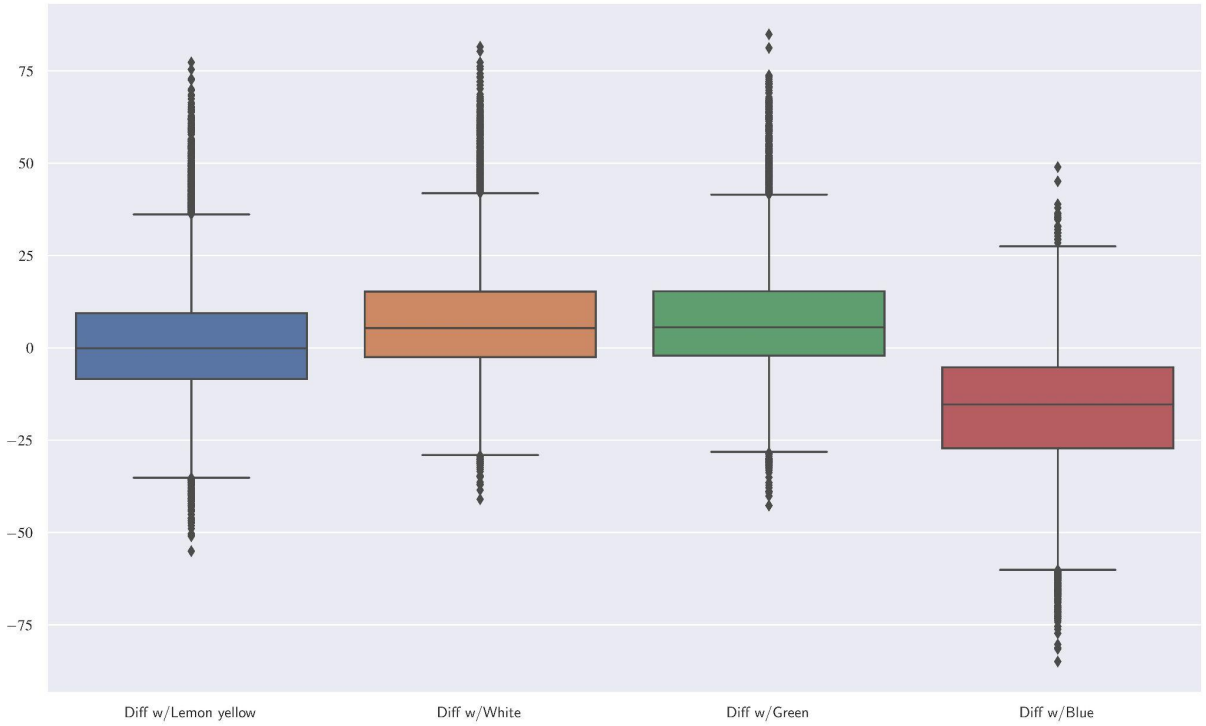

σ_b
0
5
10
15
20
25
30


σ_b
30
25
20
15
10
5
0
0
5000
10000
15000
20000


Experimental Error
35
40
45
50
55
60
65
70


Experimental Error
70
65
60
55
50
45
40
35
0
5000
10000
15000
20000

b) From this Anova analysis we can conclude there is no clear evidence of statistical significance for the bugs actually preferring a color. To test this the difference between treatments for each color was calculated and its boxplots done.



The box plots were obtained by subtracting the treatment random variable under study from all the other ones. The resulting deterministic random variables were then sampled and flattened and box plots derived from them for visualization purposes.

The meaning of the graph is then that for each color, we want to assess whether its class' treatment is included in the rest of the color's treatment's confidence intervals. Since every box contains 0 in its confidence interval one may assume there is no statistical significance for that color in particular. The most outstanding color is blue but with this amount of data its statistical significance can't be concluded.