Insight Into Experimental Data
DDM160

Individual Report_Challenge 1

Virginia Patricia Rispoli
S152855
23_03_2017
Windows user

0.THE DATA SET

The data set has been downloaded from an SPSS tutorial platform (http://calcnet.mth.cmich.edu/org/spss/Prjs_DataSets.htm).
The data consists of drug information collected on 50 patients used to perform frequency and descriptive statistics.
The variables in the data set are:
Subject: Patient;
Treatment: Two levels: 0 for Placebo and 1 for treatment group;
Age: age of patient;
Gender: Male(1) or Female(2);
Before_exp_BP: Blood pressure before experiment;
After_exp_BP: Blood pressure after experiment.

	A	B	C	D	E	F
1	Subject	Treatment	Age	Gender	Before_exp_BP	After_exp_BP
2	D1	1	65	1	103.3	80.5
3	D2	1	59	1	93.6	85.9
4	D3	1	60	2	92	85.2
5	D4	1	54	1	93	87.8
6	D5	1	65	1	95.4	85.3
7	D6	1	57	2	109.6	94.2
8	D7	1	69	2	97.9	83.9
9	D8	1	62	2	96	85
10	D9	1	49	1	98.4	86.3
11	D10	1	45	1	98.4	90
12	D11	1	65	1	95.5	85.2
13	D12	1	62	2	91.7	87.9
14	D13	1	64	2	98.6	84.6
15	D14	1	68	1	98	83.8
16	D15	1	70	2	96.4	85.5
17	D16	1	66	2	104.4	93
18	D17	1	65	1	111.7	85.4

Figure 1. The dataset NewDrug.CSV

The research question: is the effect on the blood pressure with the medicine (Treatment 1) bigger than the effect of the placebo treatment (Treatment 0)?
It has been decided to analyze the effect of the two different treatments (column B), namely how the blood pressure does change after treatment 0, placebo and treatment 1, medicine.

To upload the dataset to the program I have set 2 conditions, and from the Open tab, I have uploaded the file as a CSV file with attributes (field separation: semicolon).
In this case, the measurements for different conditions are not in separated columns, so I have put the column Treatment as the first condition and the After_ex_BP as the dependent value.
Additionally, I have asked the program to treat the data as continuous since the dataset presents measures of blood pressure.

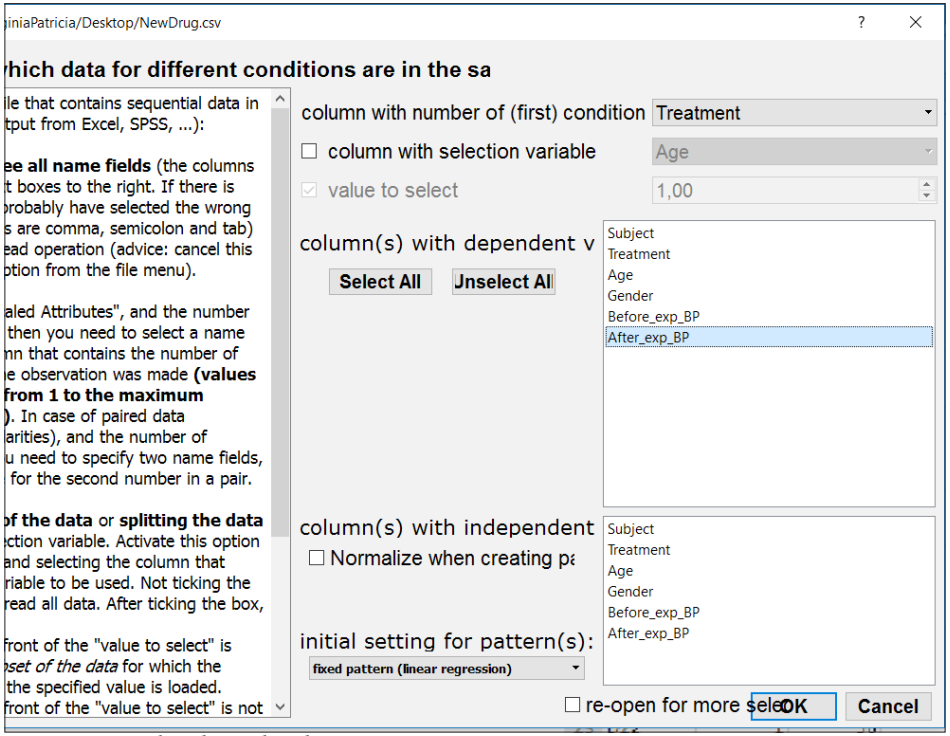


Figure 2. Uploading the dataset on ILLMO.

1.HISTOGRAMS

After uploading the dataset on ILLMO, specifying the number of conditions as 2, I have chosen to look at the data on histograms. A histogram is a graphical representation of the distribution of the data, and it is used to give a sense of the density of the underlying distribution of the data.

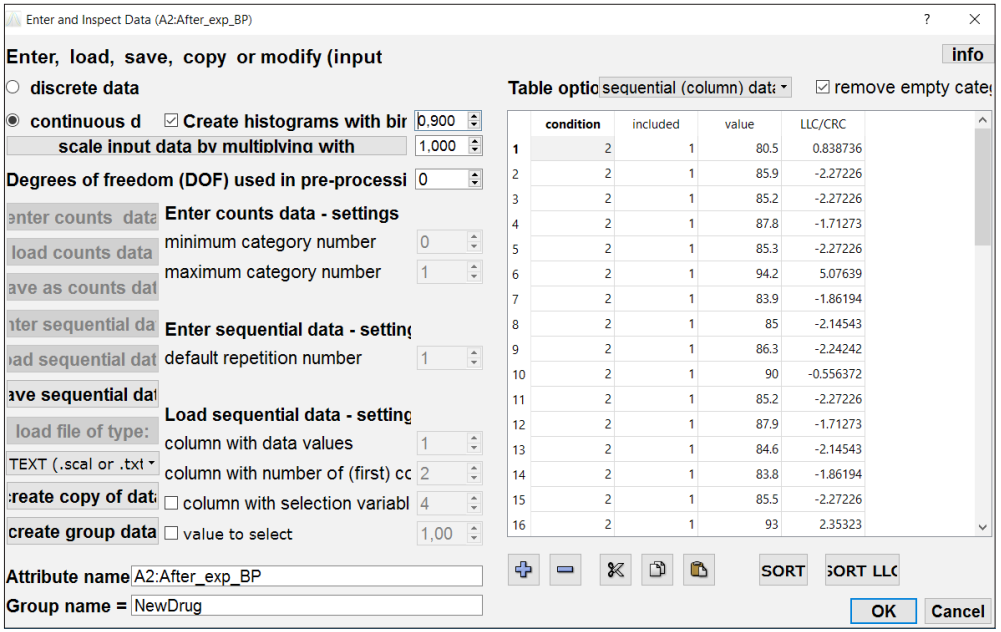


Figure 3. Quantizing the data set NewDrug.CSV

In order to facilitate the comparison of histograms between pairs of conditions, the histogram for the selected condition is highlighted by drawing it in a different color (red) than the histogram for the reference condition (black), as shown in Figure 4.

Since this is a continuous dataset, the option of regular histogram is not interesting for the analysis as the cumulative one is. Visually assessing such line density is quite difficult and the rendering of regular histograms is not very informative. By default, ILLMO does not generate regular histograms for continuous data, but this can be overruled ticking the checkbox in front of the "Create histograms with bin size" in the data dialog, (see Figure 3) and so quantizing the data with a value of 0.900, provided that it is substantially smaller than 1. Changing the bin size to a value of 0,900 also increases the LLC_Q to a value of 34.75, which is a value that expresses a better fit of the module compared to the LLC value without quantization (LLC= -64,24).

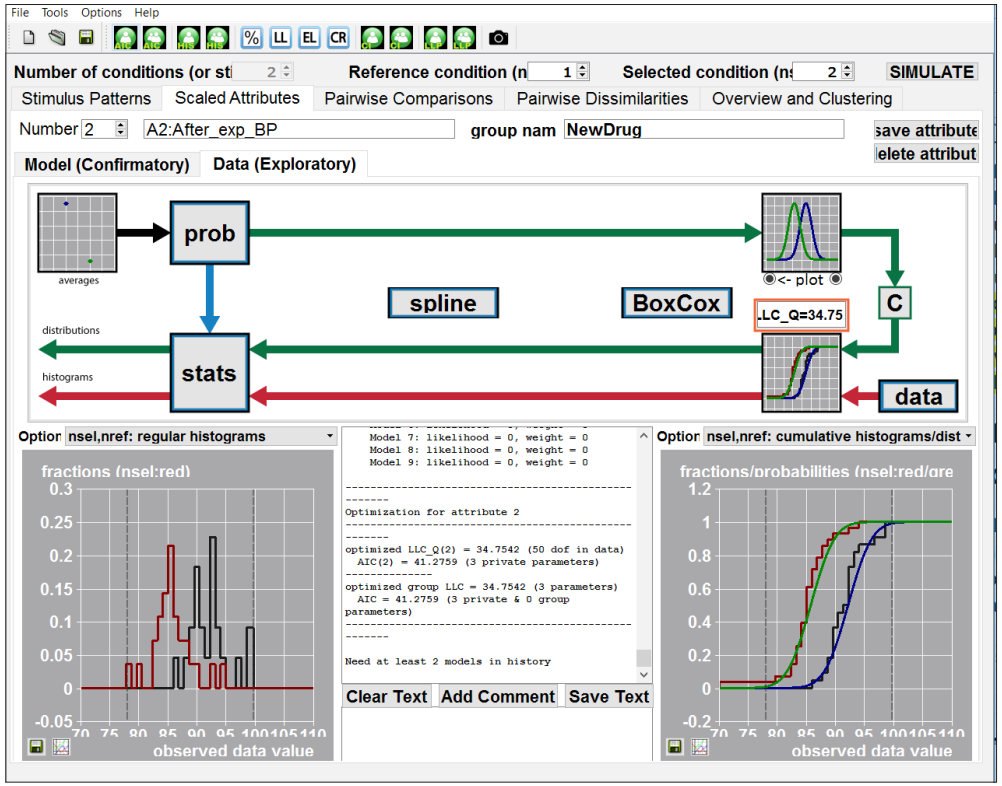


Figure 4. Continuous histograms in ILLMO for the reference (in black) and selected (in red) condition (data: NewDrug.CSV).

The LLC_Q value expresses the log-likelihood criterion value after the quantization of the data.
While the histograms for the 2 conditions appear to be slightly

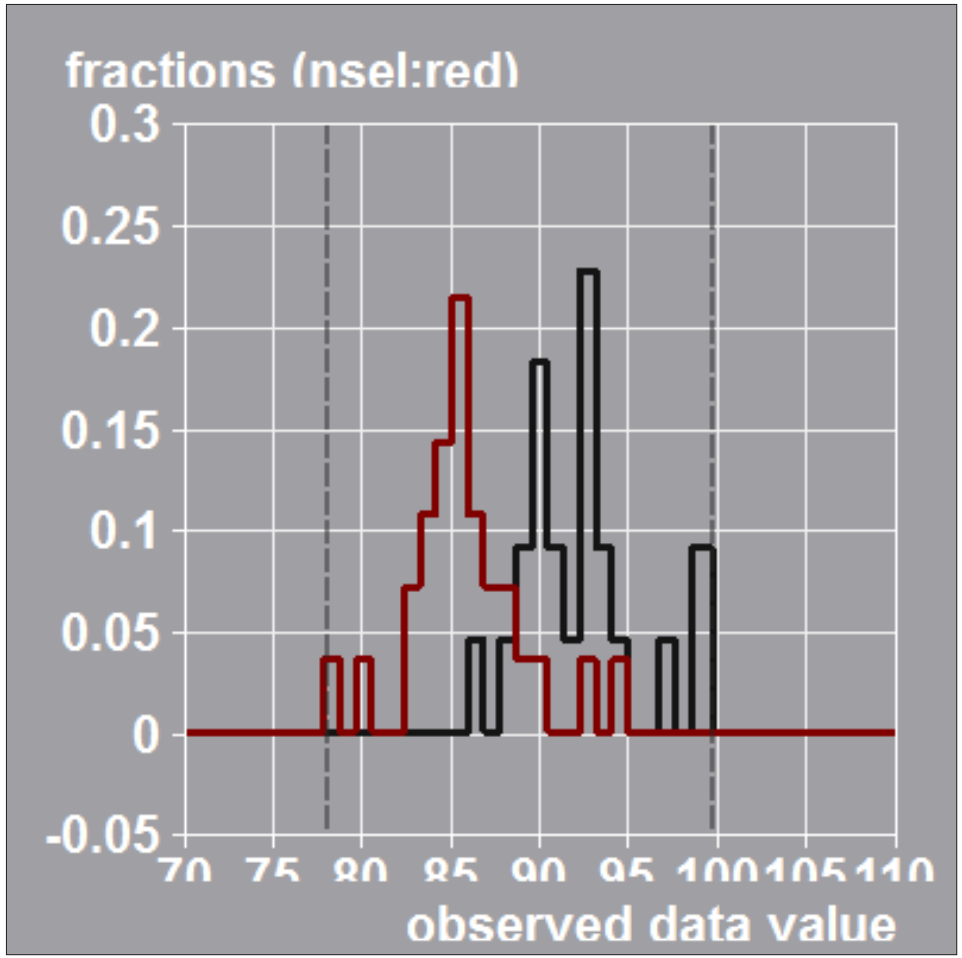


Figure 5. Regular histograms with bin size 0.9 for continuous data (data: NewDrug.CSV).

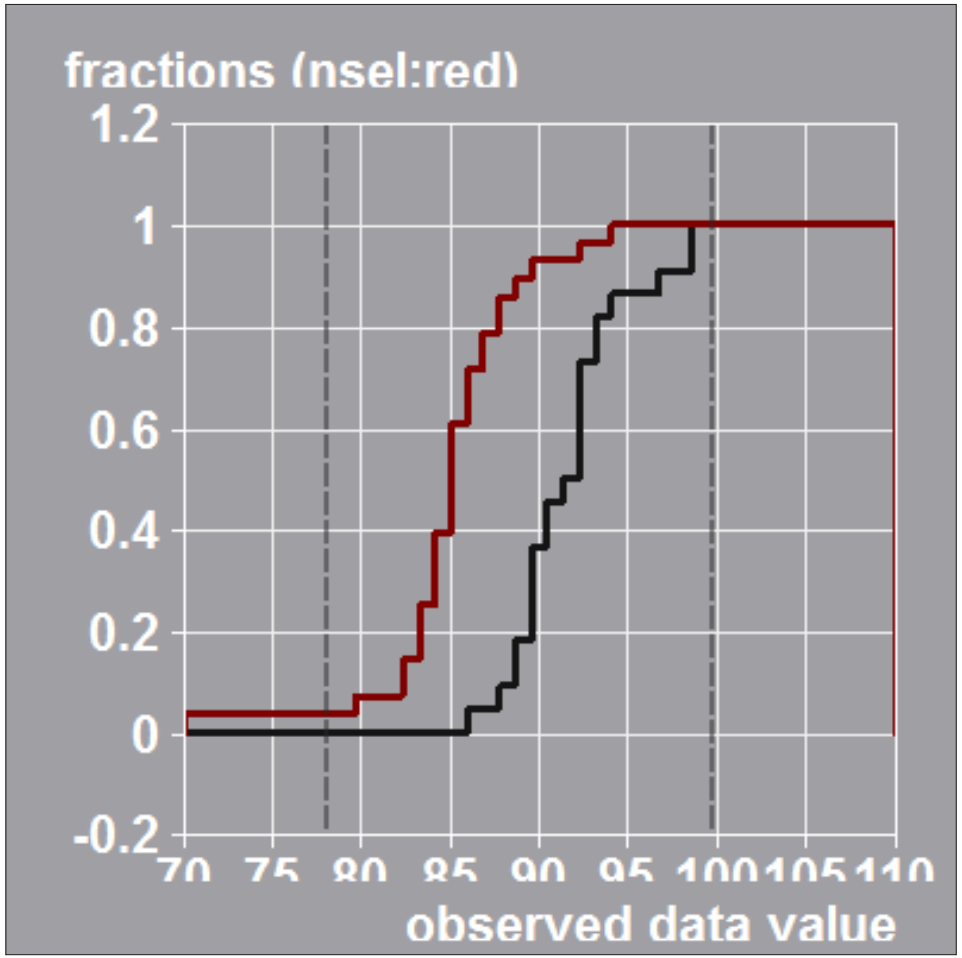


Figure 6. Cumulative histograms with bin size 0.9 for continuous data (data: NewDrug.CSV).

different, later in the report statistical modeling will be used to establish whether or not they are significantly different. Also, for this reason, the cumulative histogram is a better solution since it does show better that the two conditions are different, and it is easier to interpret the difference in the data, thing that is difficult to assess only from the regular histograms.
Already from the cumulative histograms, it is possible to notice a difference in the two variables analyzed.

For this analyzed example, we have a data set with two conditions, with the same standard deviation (σ= 3,2051) and two different averages (μ1= 92,1864 and μ2= 85,7786).

2. DISTIBUTION & MULTIMODEL COMPARISON

The default choice from the program is the Gaussian model, with an LLC_Q=34,75 in this case, and I can assess which is the model that fits my data the best with a multi-model comparison.
The Gaussian model for this data, after the quantization of 0,9 done to the dataset, has an LLC_Q=34,75, and with the comparison, I will assess if there is a model that fits better my data based on the LLC value and on the AIC value.
The multi-model comparison will help the choice for the best fitting model to the data.
To assess if a model does fit the dataset, we use two different parameters that a model needs to accomplish a good fit to the

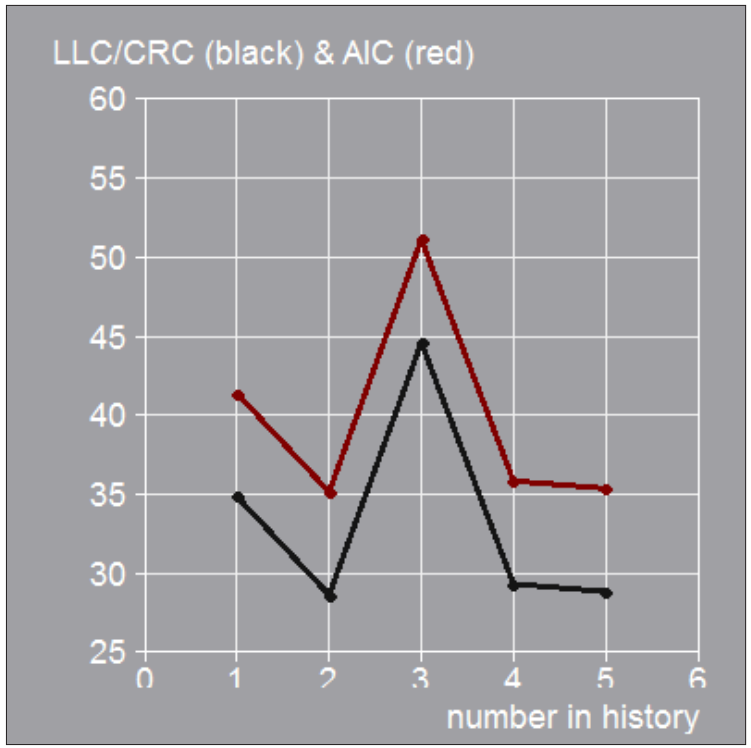
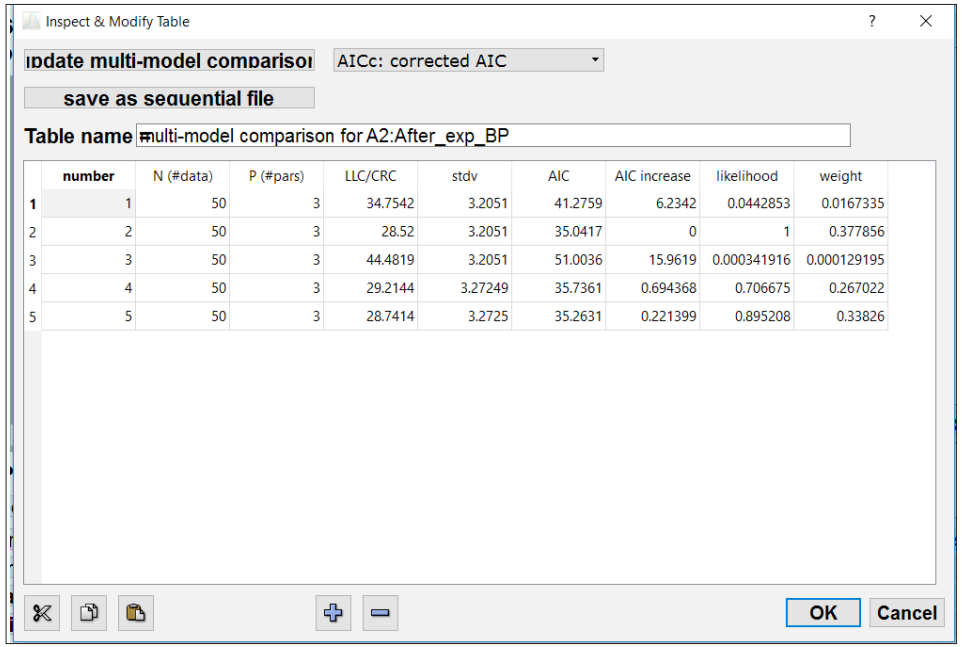


Figure 7. The history model of the models used to find the the one that better fits the dataset.
In the model we can see the curve of the LLC value and the curve of the AIC value, and we aim at the module that will has the two values closer to 0.



	number	N (#data)	P (#pars)	LLC/CRC	stdv	AIC	AIC increase	likelihood	weight
1	1	50	3	34.7542	3.2051	41.2759	6.2342	0.0442853	0.0167335
2	2	50	3	28.52	3.2051	35.0417	0	1	0.377856
3	3	50	3	44.4819	3.2051	51.0036	15.9619	0.000341916	0.000129195
4	4	50	3	29.2144	3.27249	35.7361	0.694368	0.706675	0.267022
5	5	50	3	28.7414	3.2725	35.2631	0.221399	0.895208	0.33826

Figure 8. Multimodel comparison history showing the variatios of LLC value and AIC value for each model analyzed.

observed data, the log-likelyhood criterion, that is the value for how well one or more probability distributions, which we refer to as the model, fit to an equal number of observed histograms. The goodness of fit is however only one aspect of the quality of a model, so we also analyze the Akaike Information Criterion (AIC), which combines the llc of a model with the number of parameters in a model into a single quality score.
In this case, different models have been analyzed, such as:
1. Normal Gaussian (LLC=34.7542, AIC=41.2759)
2. Laplace (LLC=28.539, AIC= 35.0607)
3. Wide Gaussian (LLC= 44.4819, AIC= 51.0037)
4. Student T (LLC=29.2143, AIC= 35.736)
5. Laplace (LLC=28.539, AIC= 35.0607).
The conclusion is that based on LLC value and AIC the best fit has been found in the Laplace model, and on the analysis done to the

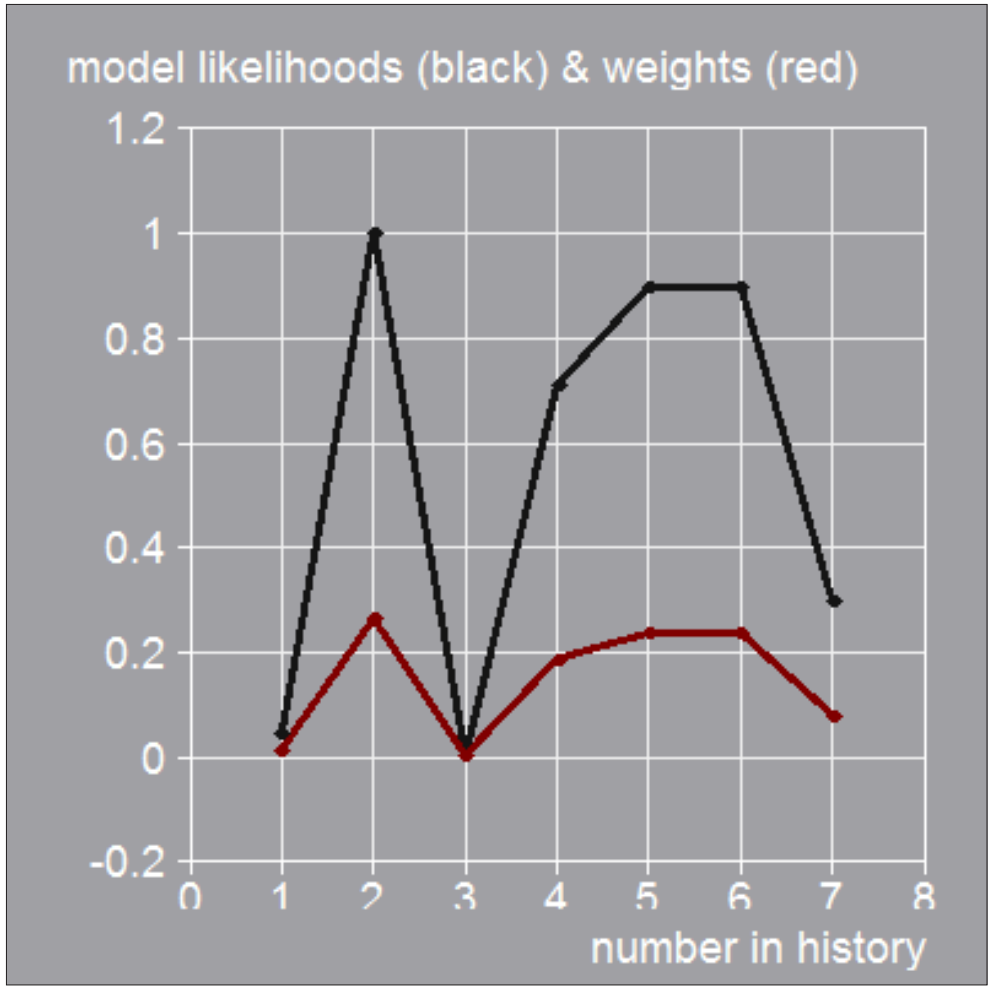


Figure 9. In this case changing the standard deviation from constant across the conditions to varies between the conditions does not increase the fit of the model, and instead it will increase the DOF form 6 to 8 generating more parameters.

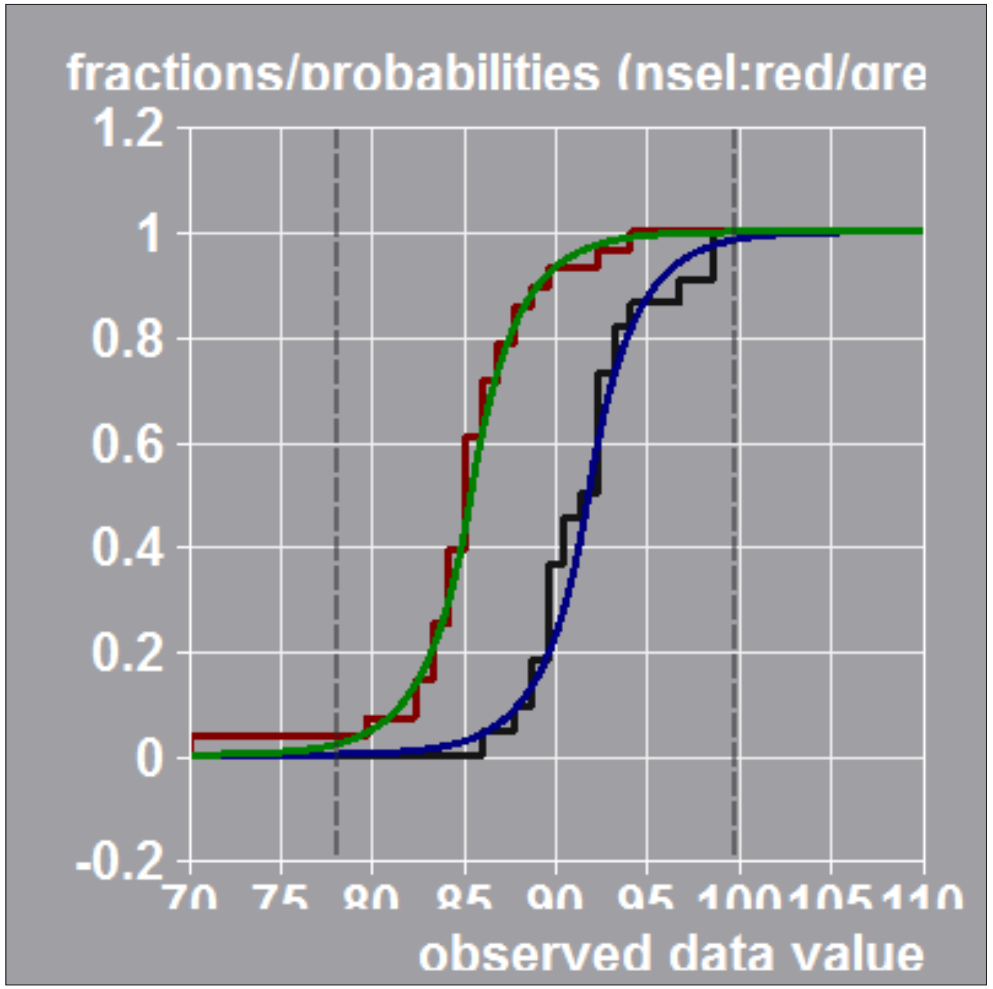


Figure 10. The cumulative histograms with the Laplace distribution for the data: NewDrug.CSV.

effect size, that i will show later in the report.
From the history model in Figure 7 and the model shown in Figure 8, we can define that the Laplace model is the one that fits better the dataset, with an LLC value of 28,74 and an AIC value of 35.2632 (Figure 10).

To optimize the model and see if it possible to have a better fit, I have changed the standard deviation from constant across conditions to varies between conditions. In this case, as it can be seen from Figure 9, this change does not provide a better fit, and instead, it increases the Degree of Freedom from 6 to 8, having more parameters in the analysis.

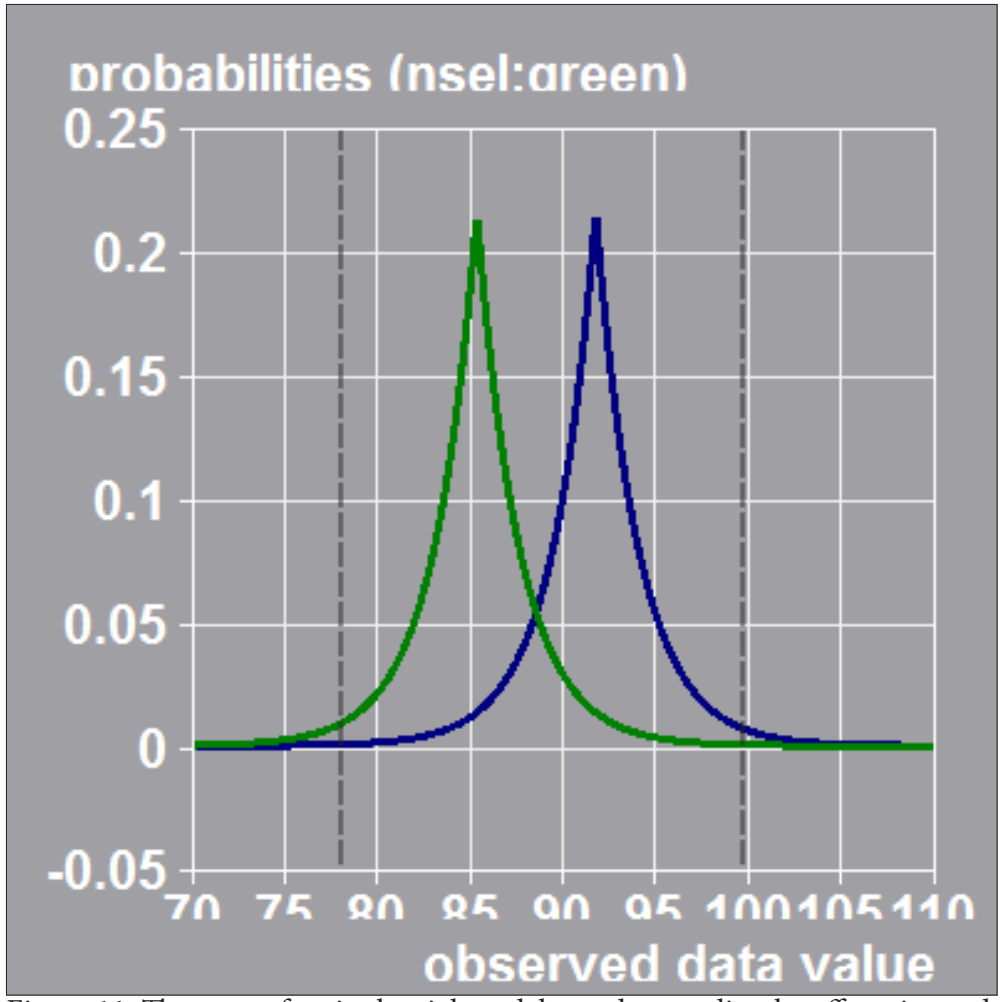


Figure 11. Thurstone for single trial model, good to predict the effect size and to show the difference in the avarages, that is important and correlatd to my effect in this case.

Even though the Gaussian distribution would have been an easier choice for my dataset, and also easier to generalize the selection of the model to other similar datasets, I have decided to maintain the Laplace distribution after plotting the Thurstone of the single trial model, since it is better to predict that the effect size is significant because there is not a big overlapping between the curves. Moreover, the Thurstone does show that there is a visible difference between the two averages, important to analyze the effect size.

Once the model has been selected, we can proceed with statistical inference analysis.

3.STATISTICAL INFERENCE

Log-likelihood function and profile

The LLF, likelihood function, is a function of the parameters of a statistical model given data. Likelihood functions play a key role in statistical inference, especially methods of estimating a parameter from a set of statistics.
The LLF shows the increase in LLC for average values in the neighborhood of the optimal value.

Confidence intervals

The confidence interval is the range of values surrounding the estimated value for a parameter that is expected to contain, with a probability of 95%, the mean value for that parameter. is an observed interval that potentially includes the unobservable true parameter of interest.

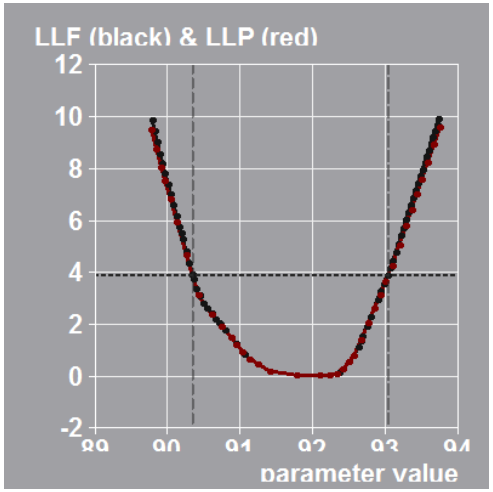


Figure 12. LLF and LLP for the avarage of condition 1. CI(95.00%) for LL/CR average 1, average (1) = 91.7764, CI(1) = [90.3463,93.0415] (LLP)

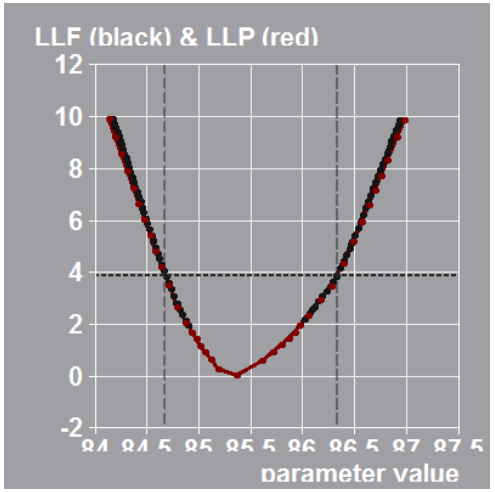


Figure 13. LLF and LLP for the avarage of condition 1. CI(95.00%) for LL/CR average 2, average (2) = 85.365, CI(2) = [84.6723,86.3273] (LLP)

In this case, I have plotted the two LLP and LLF for the different averages, as shown in Figure 12 and 13.
In this case we have for the first condition:
μ1= 91.7764, with 95% CI= [90.3463,93.0415]
and for the second condition
μ2= 85.365, with 95% CI= [84.6723,86.3273].

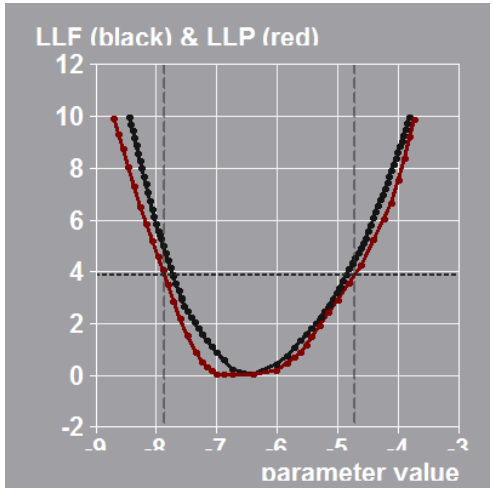


Figure 14. Graph showing the LLF (in black) and the LLP (in red) for the difference in averages in two conditions .

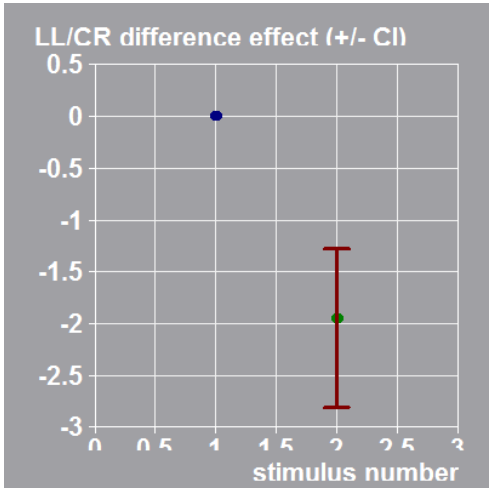


Figure 15. Graph showing that there is a significant difference, and so an effect size since 0 is not comprehended.

Effect size

From the graphs in Figure 14 and 15 we can tell that we can look for an effect in our data.
The effect size of this dataset is
JND = 1.38535 (effect > 1 JND)
area above ROC: 0.925266 (effect > 1 JND)
which translates into a large effect size (see Figure 16).
To complete the analysis of the effect size we have to analyze also the ROC curve, the Receiver Operator Curve(see Figure 18).
From this analysis, we can conclude that the effect of the medicines on the blood pressure is higher than the effect of the placebo treatment.

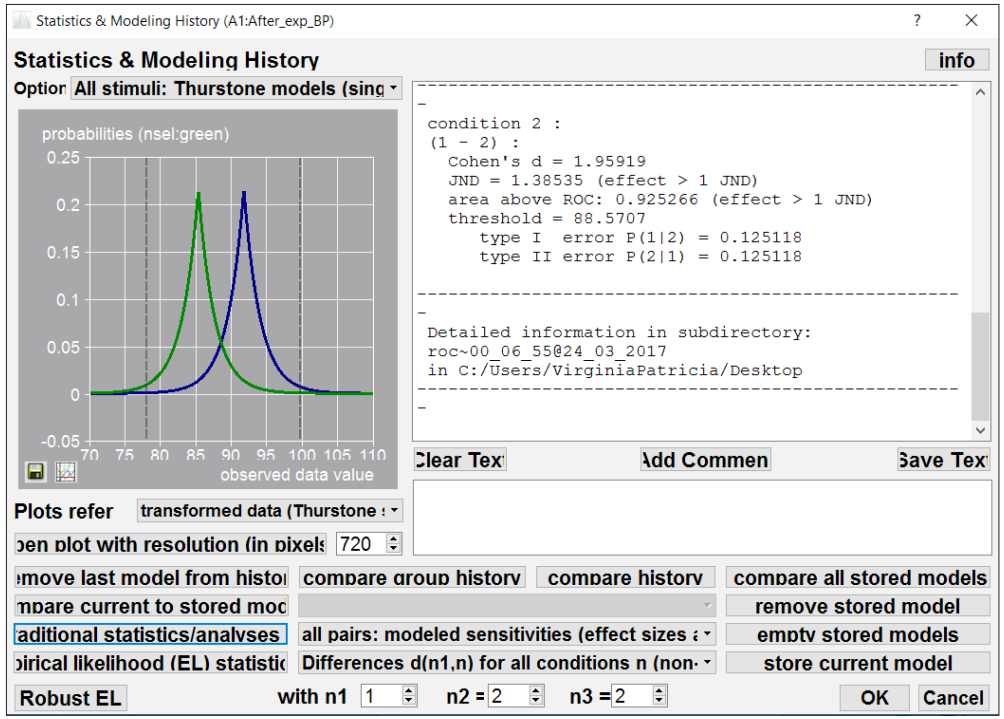


Figure 16. Effect size for the data set NewDrug.CSV

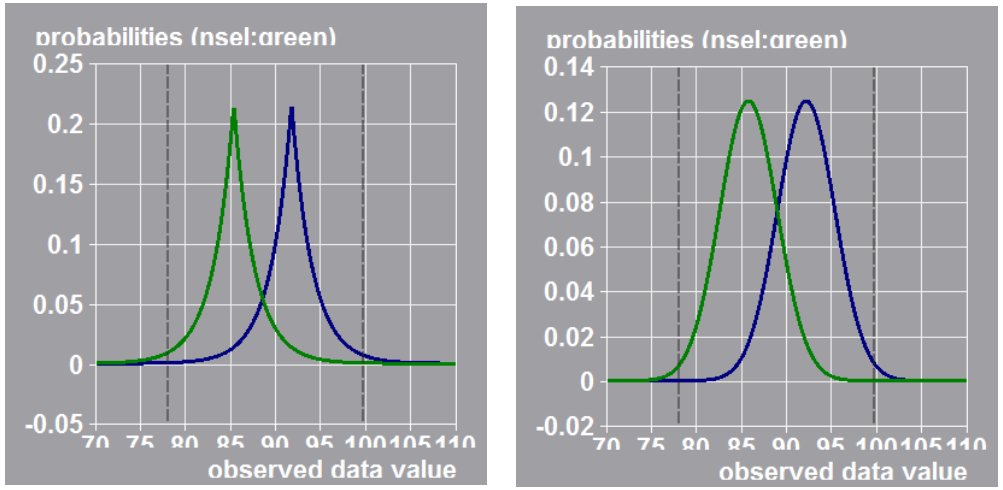


Figure17. The laplace model does show a bigger effect size and there is a smaller overlapping than the Gaussian distribution so it is a better selection more valuable for data.

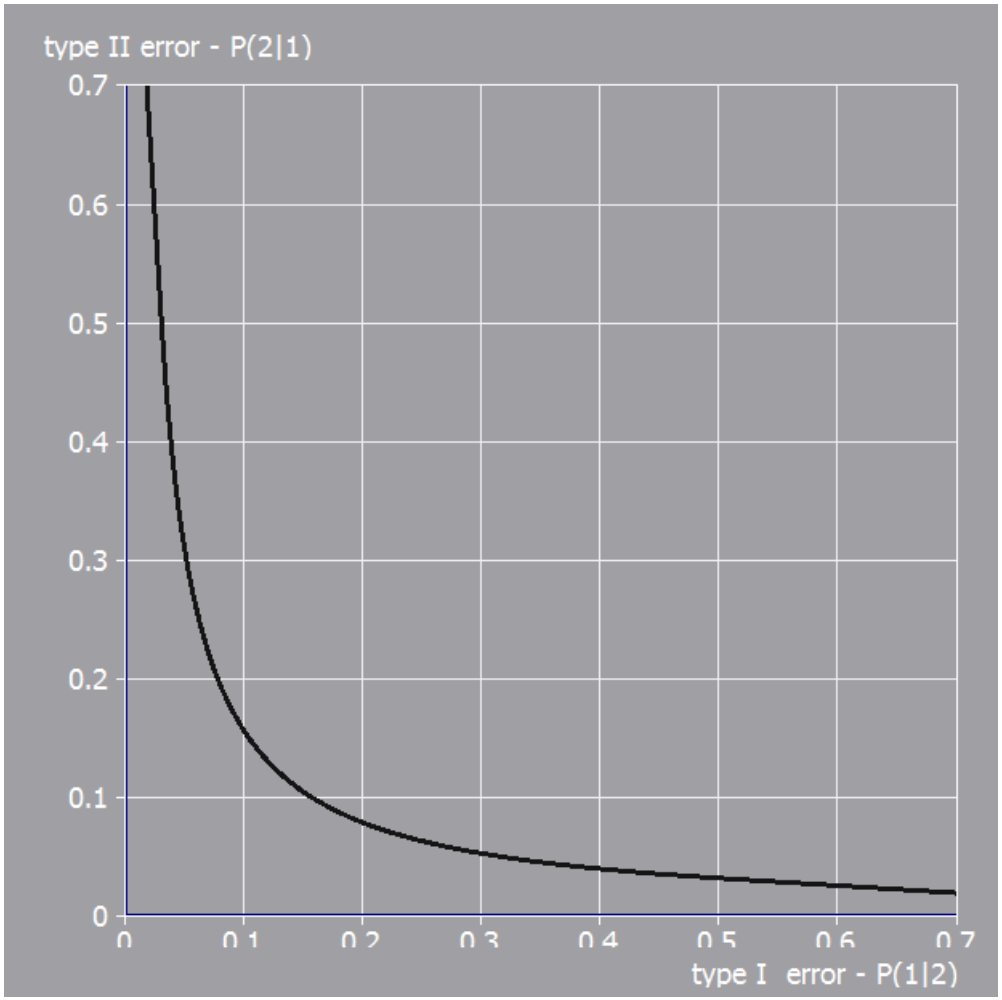


Figure 18. The Receiver Operator Curve (ROC) visualizes the errors that can occur when trying to establish the condition in which an observation was generated based on the observation itself. The point on the ROC that is closest to the origin will be the reported error probabilities. In this example the minimum error for individual trials is 0.125.