Step1 Team Project Multivariate Analysis

Adrian White, Cesar Conejo, Xavier Bryant

12/6/2020

Introduction data set

We have selected the CRASH-2 data set provided by Vanderbilt School of Biostatistics for our project. It describes the outcome of a randomized controlled trial and economic valuation of the effects of tranexamic acid on death, vascular occlusive events, and transfusion requirement in bleeding trauma patients. Tranexamic acid reduces bleeding in trauma patients undergoing surgery but is an expensive treatment option. The trial's objective was to assess the effects and cost-effectiveness of an early administration of this medication.

Participants of the study were adults with, or at risk of, significant bleeding within 8 hours of injury. Sample randomization was determined by the allocation of an eight-digit sequence randomly generated by a computer. Patients and staff were masked to the treatment allocation of the tranexamic acid.

We have adjusted the original data set to remove some variables that were not relevant to our investigation. We have removed variables regarding the exact surgical procedures administered to patients, various IDs, and details on the patient outcome. We removed the health outcome columns because of complications regarding missing data, where the boolean structure of the columns relating to specific outcomes, like stroke or pulmonary embolism, left a large number of cases with missing values. Instead, we added a boolean variable for a general outcome of survival to assess the efficacy of the procedure, rather than looking at particular health outcomes in post-surgery for living patients.

We will be using variables regarding the sex, age, and injury of the patient as well as certain biometrics, like blood pressure, respiratory and heart rates, details on surgical blood transfusion, and a boolean variable on the survival of the patient. Our selection provides us with a balance of continuous and categorical variables, many of which are boolean, with minimal complications due to missing data.

Summary variables in the data set

The variables in this dataset are the following:

1. sex: (Boolean) The sex of the patient (Male/Female)
2. age : (Numerical) Age of the patient(Years)
3. injurytime: (Numerical) Hours since injury (Hours)
4. injurytype: (Categorical) Type of injury {Blunt, Penetrating, Blunt and Penetrating}
5. sbp: (Numerical) Systolic Blood Pressure (mmHg)
6. rr: (Numerical) Respiratory Rate (rate per minute)
7. cc: (Numerical) Central Capillary Refille Time (seconds)
8. hr: (Numerical) Heart Rate (rate per minute)
9. ndaysicu: (Numerical) Number of days in ICU (days)
10. ncell: (Numerical) Number of Units of Red Call Products Transfused.

11. Death: (Boolean) Indicator if the patient survived after the procedure

A summary of the data type is the following:

variable	type_variable	sub_type_variable
sex	Qualitative	Nominal
age	Quantitative	Continuous
injurytime	Quantitative	Continuous
injurytype	Qualitative	Nominal
sbp	Quantitative	Continuous
rr	Quantitative	Continuous
cc	Quantitative	Continuous
hr	Quantitative	Continuous
ndaysicu	Quantitative	Discrete
ncell	Quantitative	Continuous
death	Qualitative	Nominal

A review of the structure of the dataset is the following:

A summary of the values in the data set are:

Finally, the list of different values by column is the following:

Table 2: Count of distinct values of each variable

sex	age	injurytime	injurytype	sbp	rr	cc	hr	ndaysicu	ncell	death
2	81	78	3	153	58	16	154	47	47	2

Visual Analysis

First, we will review the distribution of the variables involve in the dataset

In the case of age, the figure 1 reflects how this variable appears to be largely weighted to the left, with lower ages featuring more frequently than those that are greater, possibly reflecting that younger people often take more risk and work higher at-risk occupations, raising their chance of experiencing trauma involving bleeding.

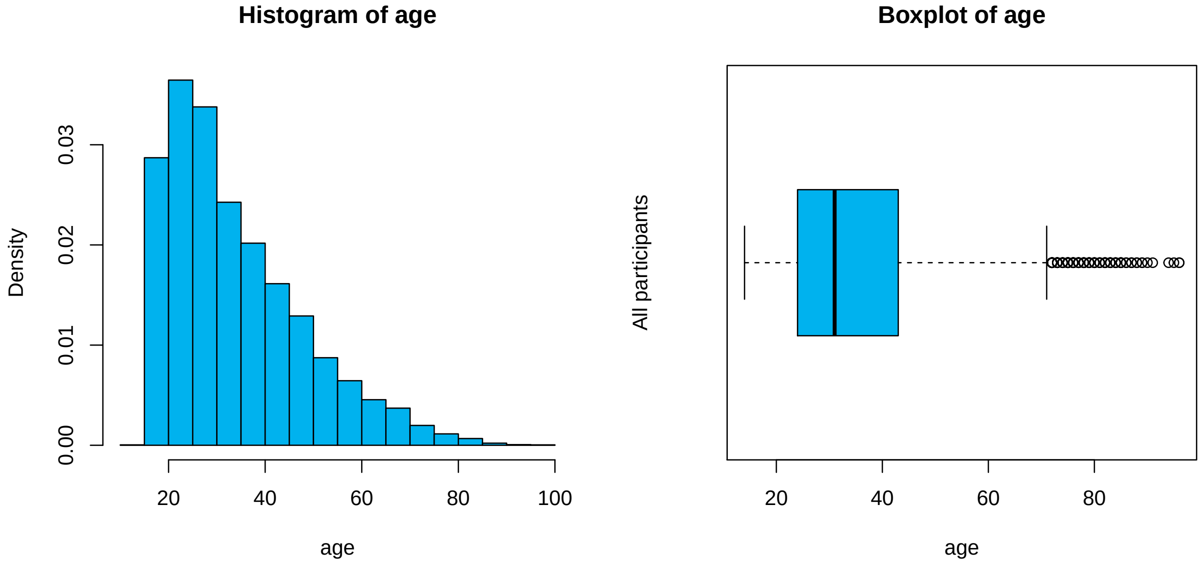


Figure 1: Distribution of age variable

The figure 2 show the distribution of the variable Injury time. We can see how this variable is highly positive skewed with almost all values falling below ten minutes since the injury was experienced. This is likely due to the fact that in cases of serious injury victims are brought to the hospital quite quickly. Then, we apply a `log` transformation to this variables as we can see in figure 3.

For *sbp* (Systolic Blood Pressure), the distribution is a fairly centrally balanced distribution around 90 mmHg. This is logical as a sample of biological characteristics observed in a population is likely to have most people around the mean and then a reasonably tight distribution of those who differ, similar to that of other biological features like height. Furthermore, most people are fairly young in the sample and therefore would have rates that deviant less from the norm, at a healthy level. The distribution is given by figure 4.

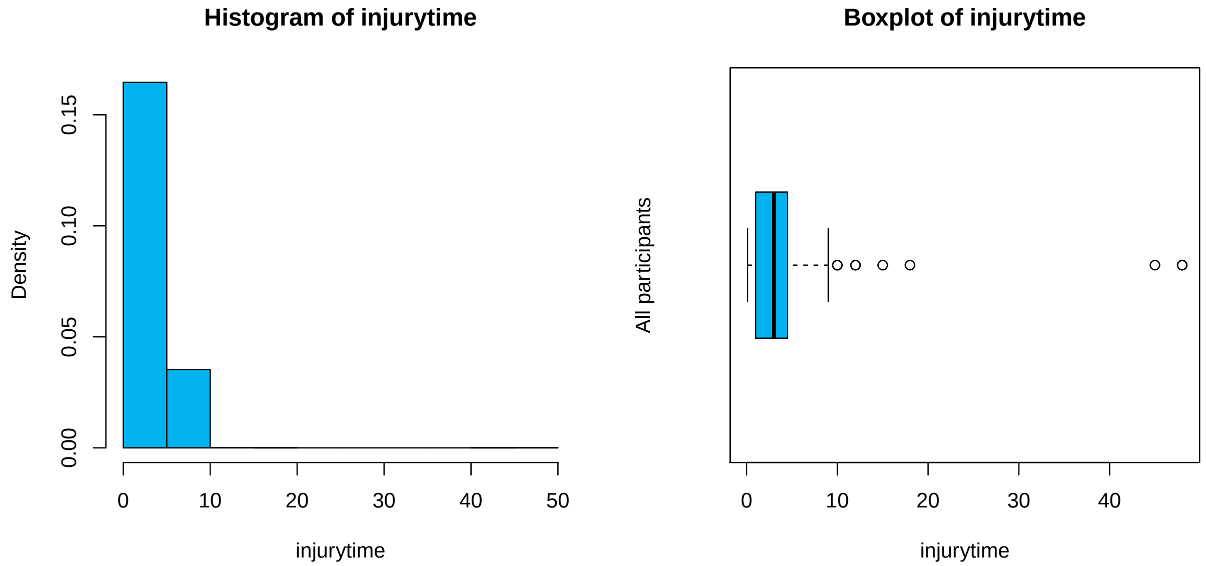
In the case of *rr* (Respiratory Rate) appears, similar to sbp, resembling a moderately balanced distribution around 22 respirations per minute, although is weighted more to the right. The distribution of this variable is showed in figure 5. Taking a log transformation, we have the new distribution in figure 6.

In the case of *hr* (Heart rate), figure 7 show that this distribution seems fairly balanced at around 110, similar to the variables above, like *sbp* and *rr*.

For *cc* (Central capillary) refill has 75% of the observations below of 4 as we can appreciate in figure 8. However, the distribution is right-skewed. As a result, we apply a log transformation that is given in figure 9.

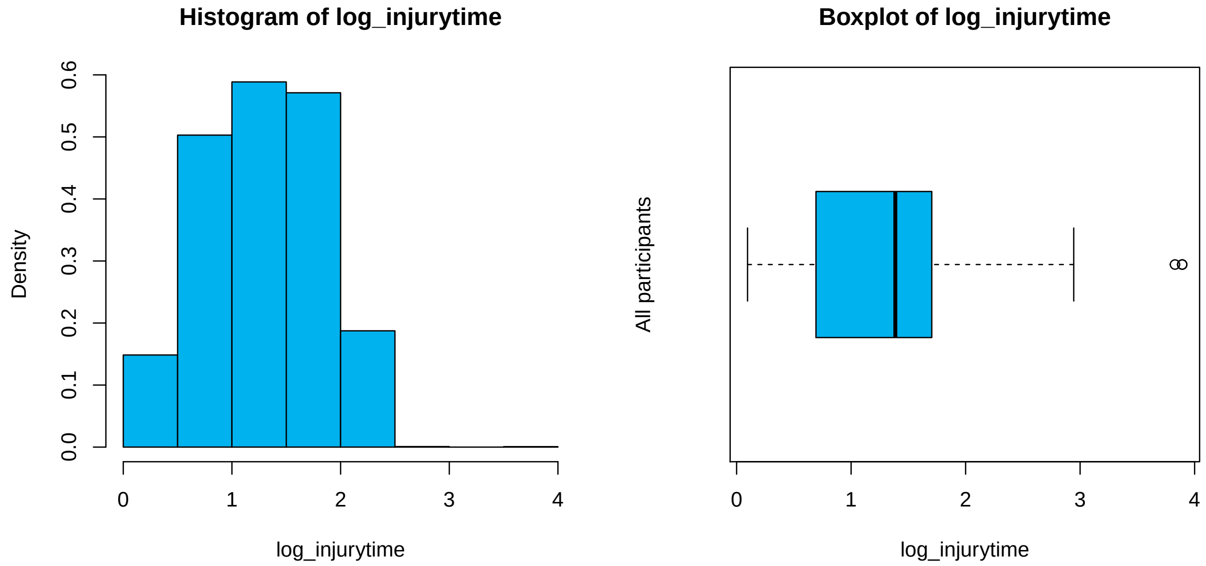
The figure 10 shows the distribution of the *ndaysicu:* variable. In this case, the distribution is heavily weighted to the left and right-skewed. Most patients it seems, with injuries at high risk of bleeding, do not often need to remain in the hospital for long. The transformed distribution is given in the figure 11.

Finally, for the *ncell*distribution is weighted to the left with a median of 3 as we can appreciate in the figure 12. The conclusion of this it that with many patients, only needing a small number of or zero units of red


Histogram of injurytime

Boxplot of injurytime

Figure 2: Distribution of injurytime variable


Histogram of log_injurytime

Boxplot of log_injurytime

Figure 3: Distribution of the log injurytime variable

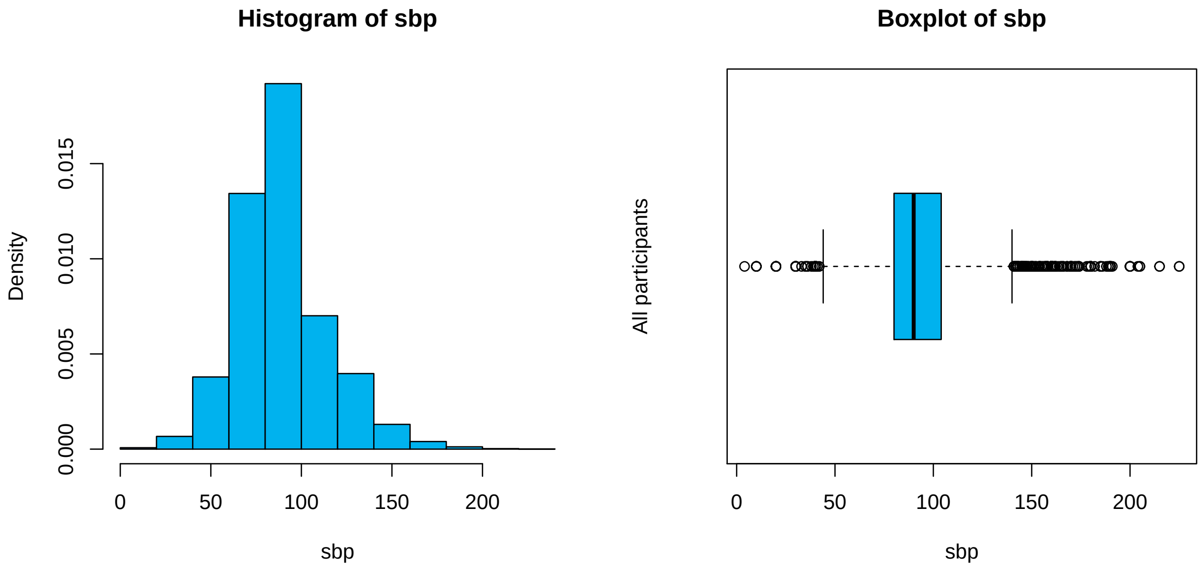


Figure 4: Distribution of sbp (Systolic Blood Pressure)

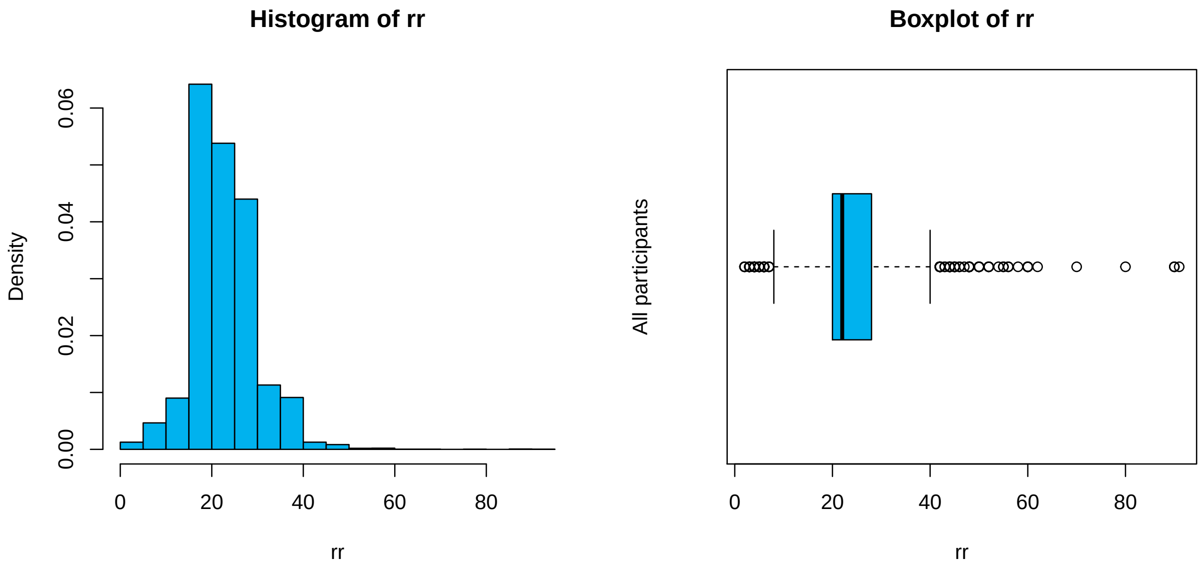


Figure 5: Distribution of rr (Respiratory Rate)

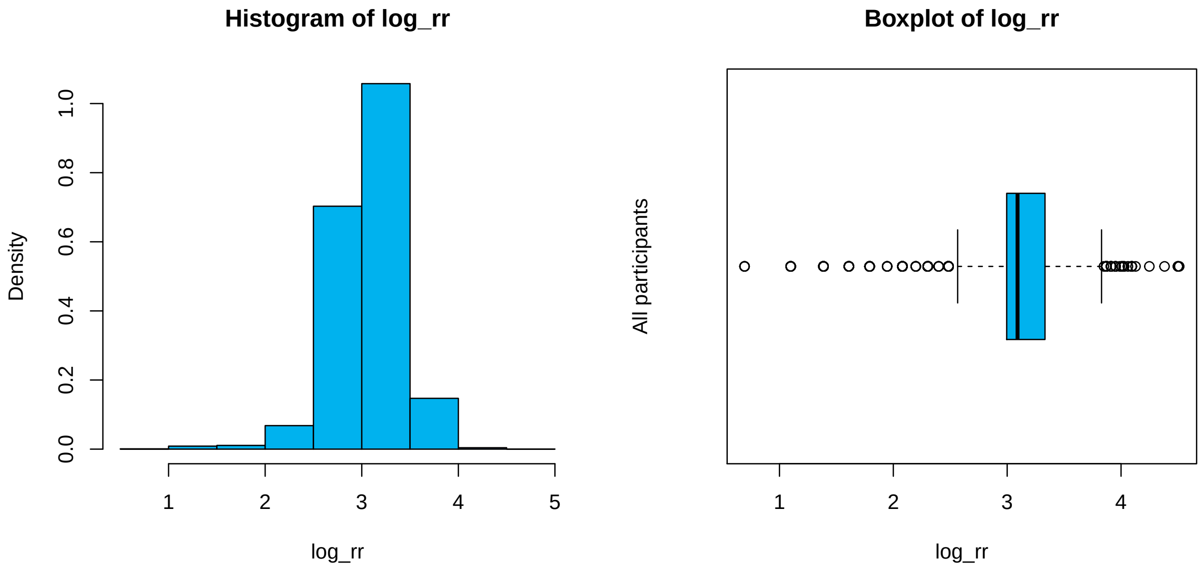


Figure 6: Distribution of log transformation of rr (Respiratory Rate)

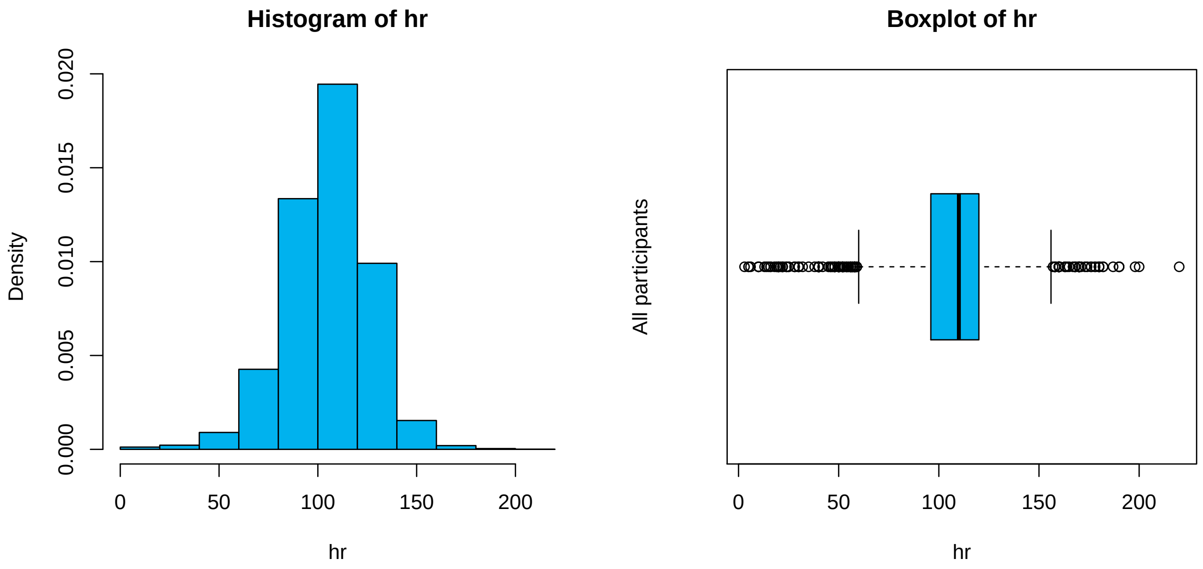


Figure 7: Distribution of hh (hearth Rate)

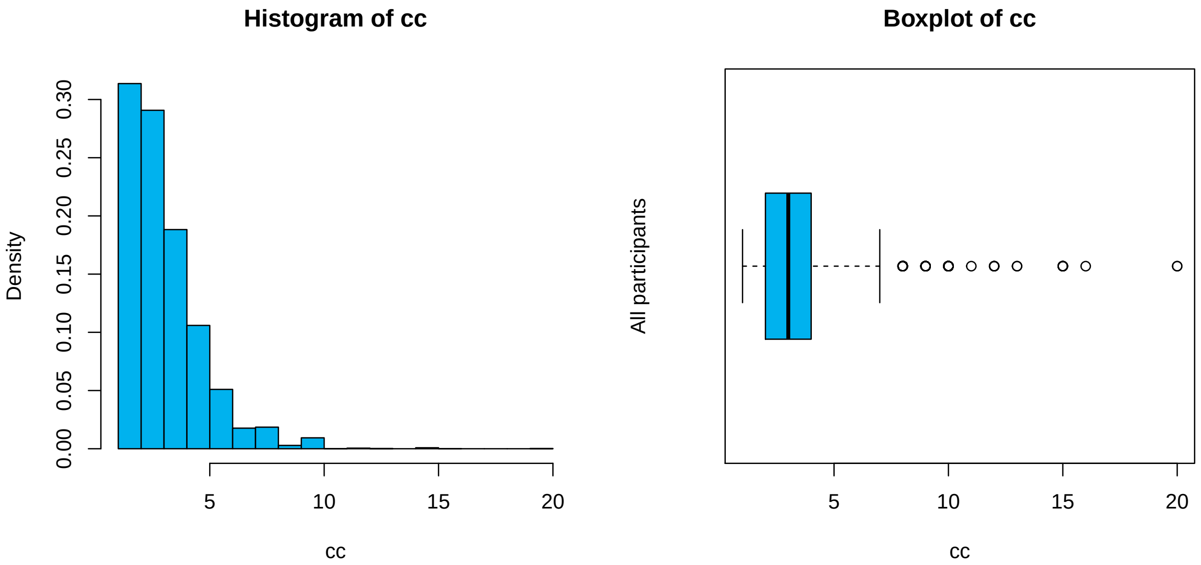


Figure 8: Distribution of cc (Central capillary)

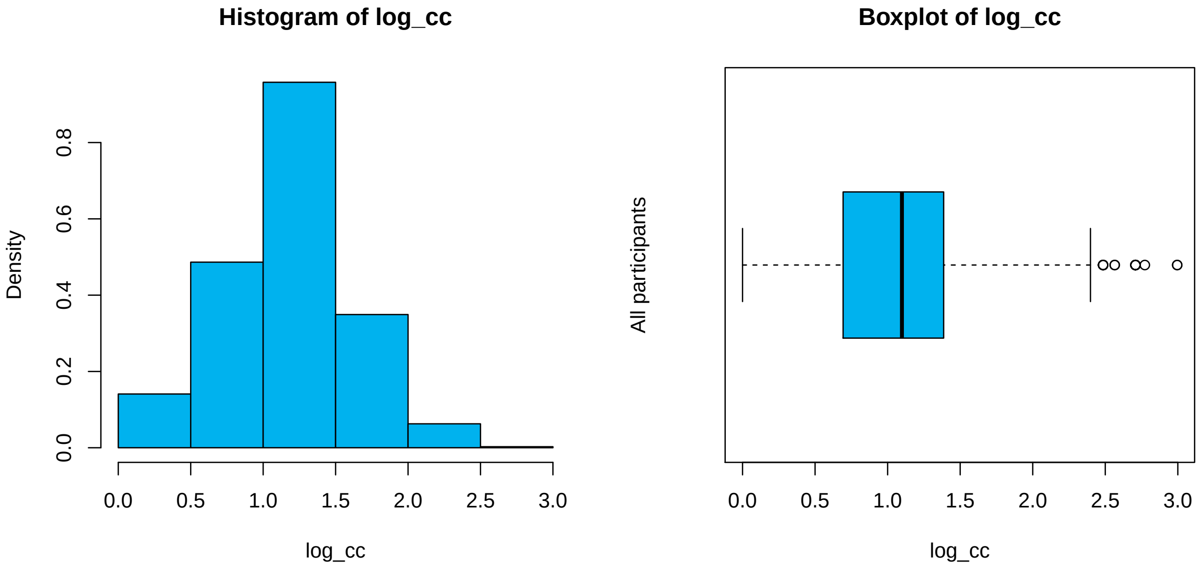


Figure 9: Distribution of log transformation of cc (Central capillary)

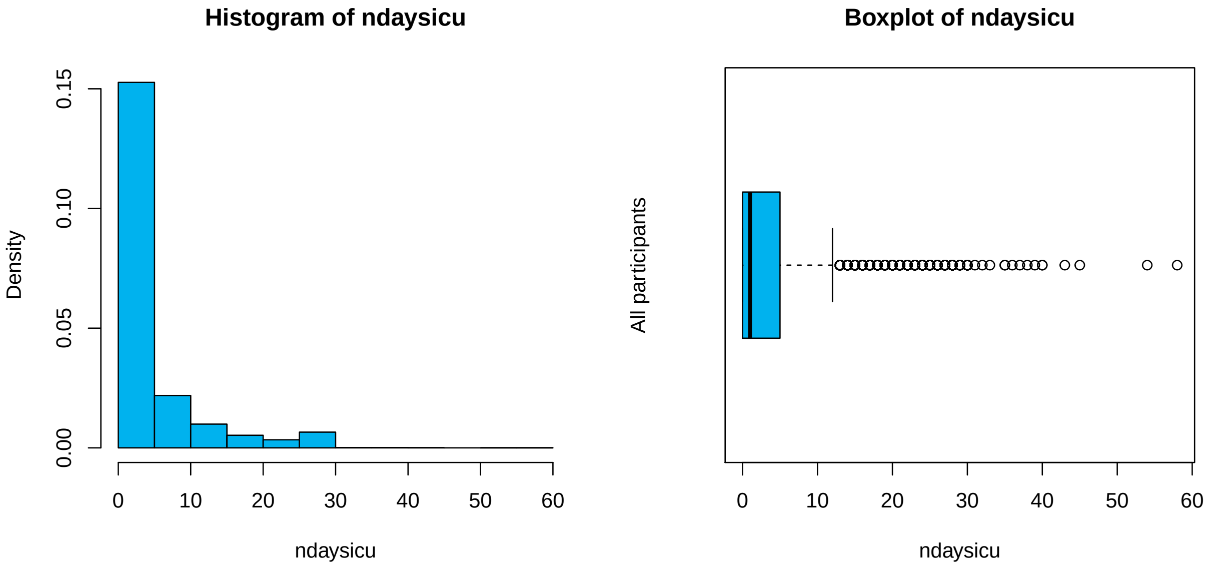


Figure 10: Distribution of ndaysicu

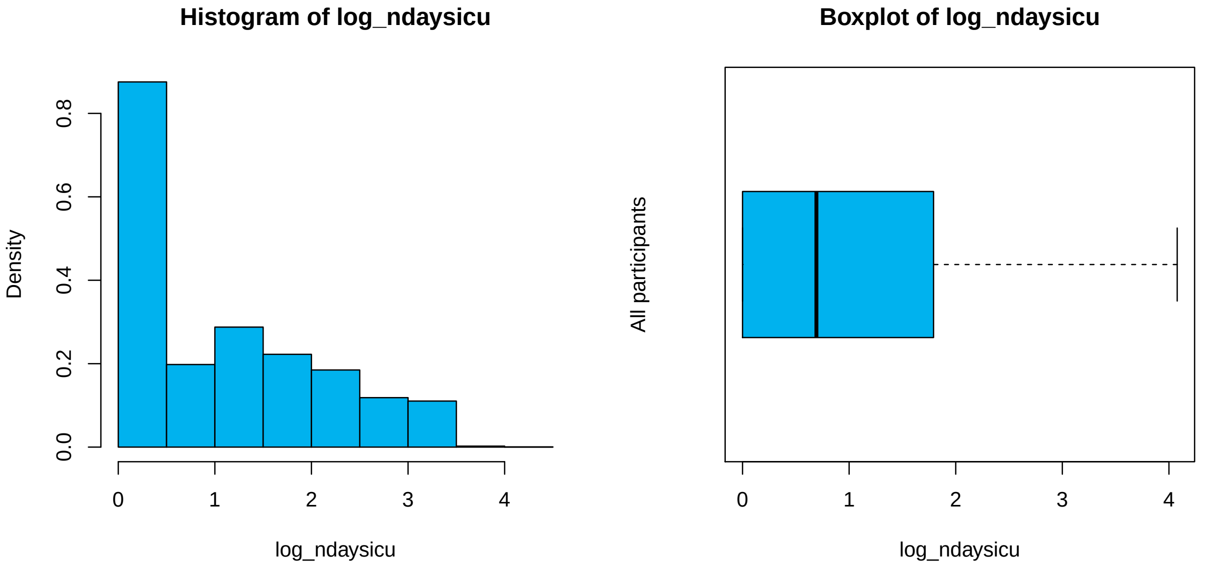


Figure 11: Distribution of log ndaysicu

cell products transfused. Due to this variable is highly right skewed, we apply the log transformation of the figure 13.

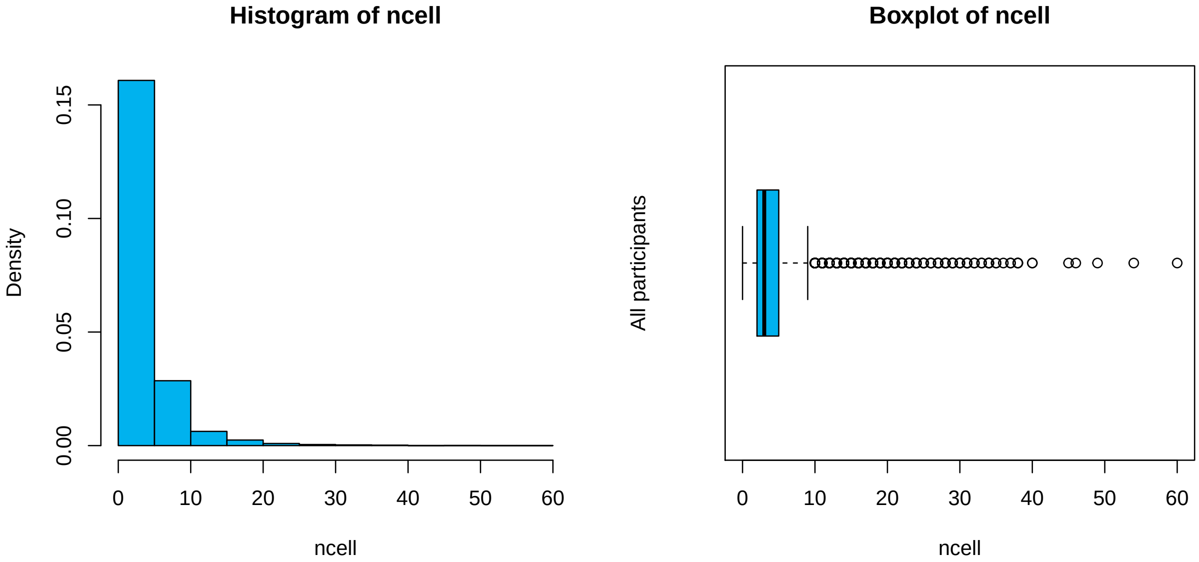


Figure 12: Distribution of ncell

For the categorical variables, we will focus on the distributions of deaths. The figure 14, shows that approximately for each death, 4 people survive. In the context of this problem, if we have an unbalanced proportion of people that survive can be considered as a sign that the drugs works.

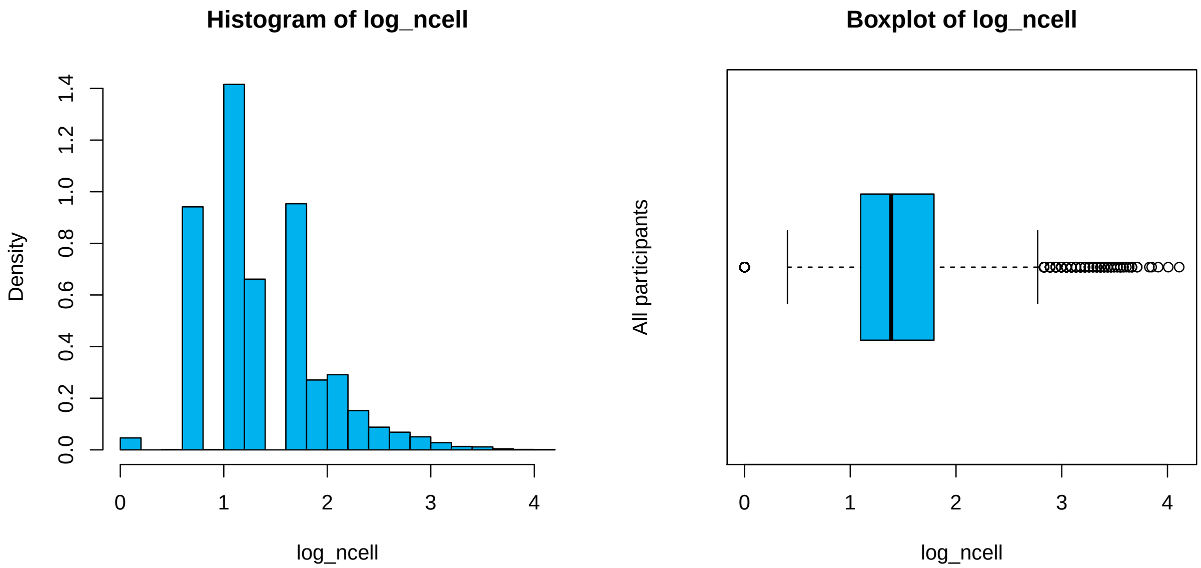


Figure 13: Distribution of log ncell

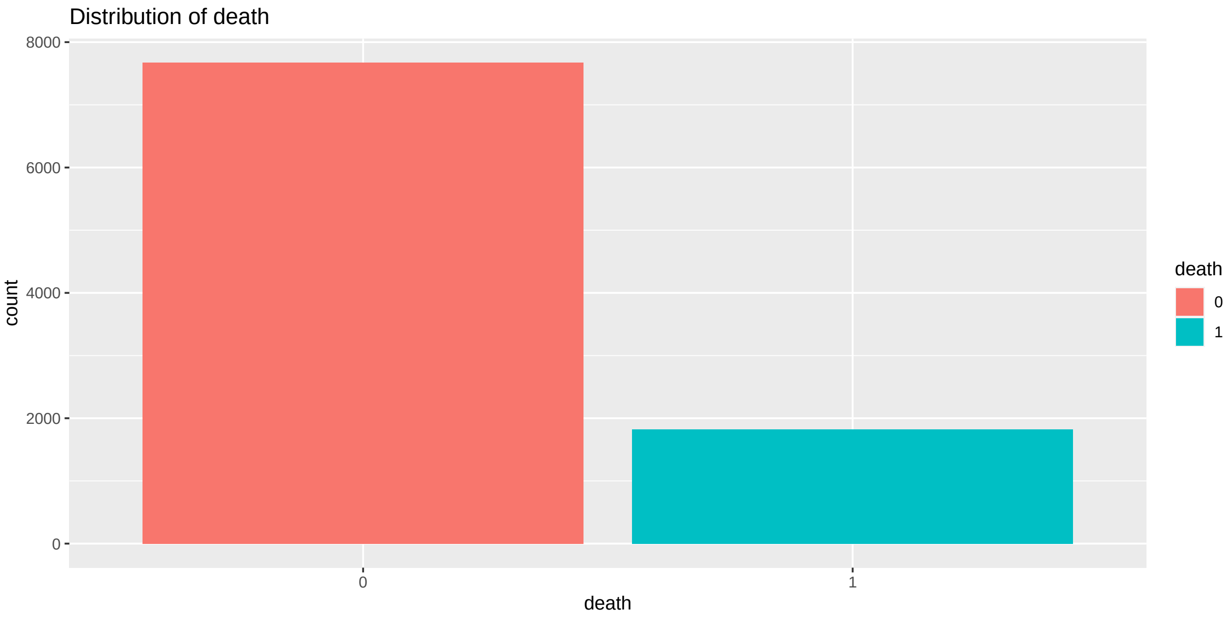


Figure 14: Distribution of deaths

On the other hand, we can study some relations of the quantitative variables in terms of the categorical variable death. Figure 15

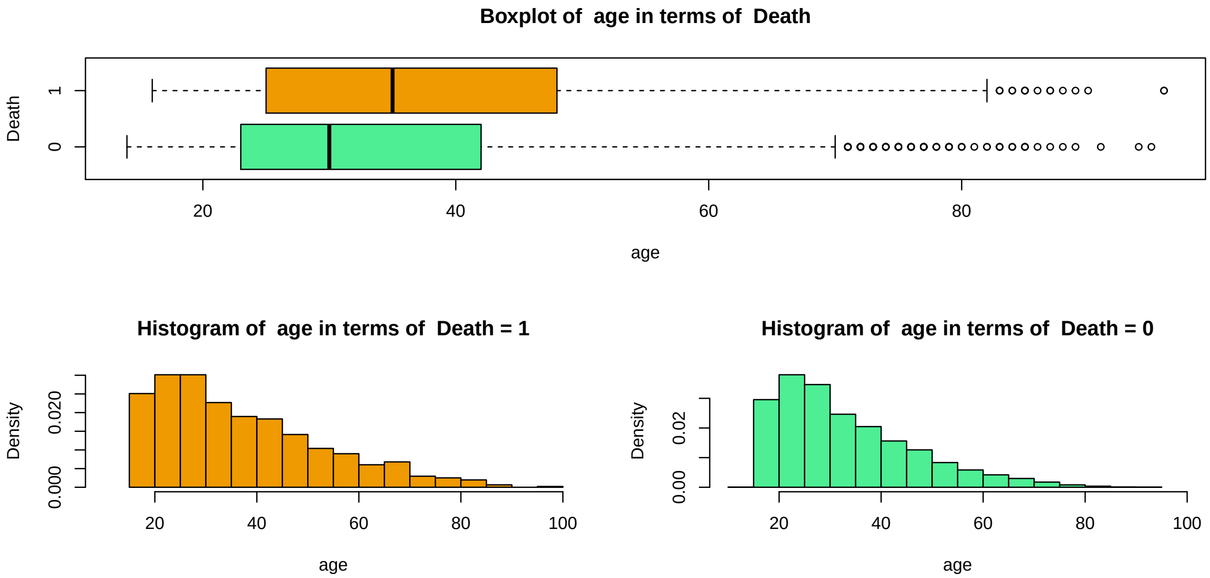


Figure 15: Distribution of age in terms of death

Figure 16

Figure 17

Figure 16

Figure 19

Figure 20

Figure 21

Figure 22

Finally, the scatter plot. . .

The PCP plot. . .

The Andrews' plot. . .

Sample Estimators

Principal Component Analysis

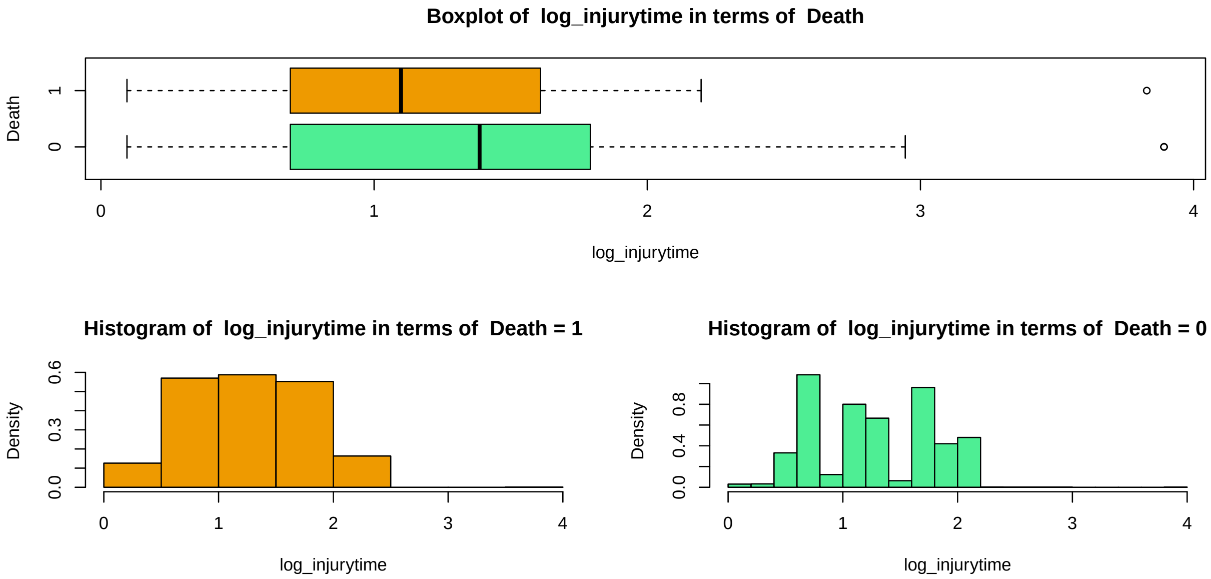


Figure 16: Distribution of the log injurytime in terms of death

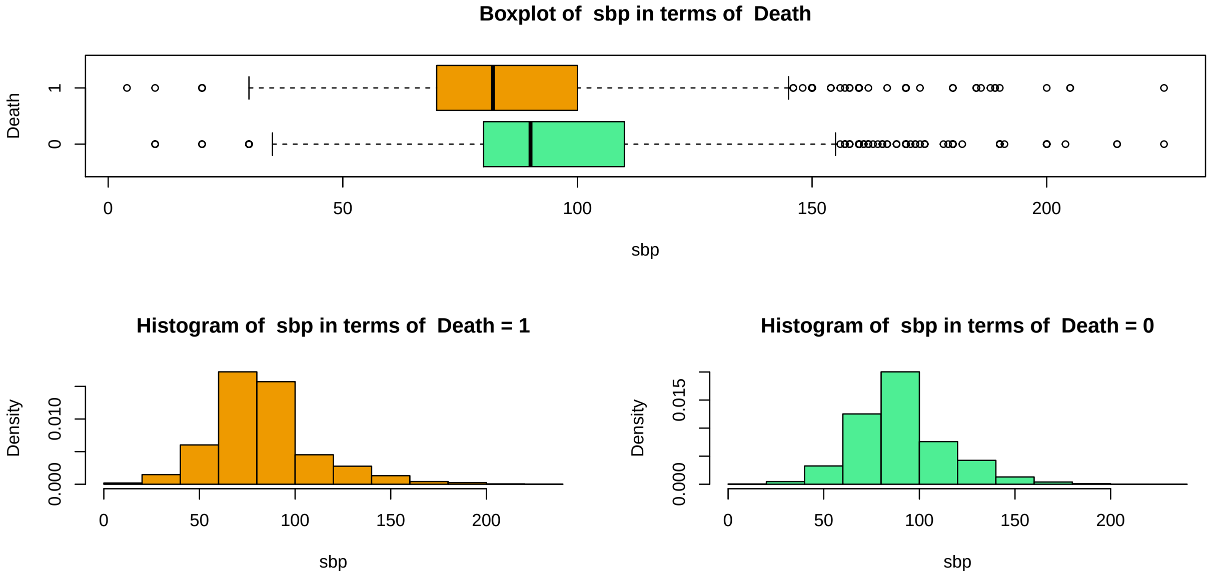


Figure 17: Distribution of sbp (Systolic Blood Pressure) in terms of death

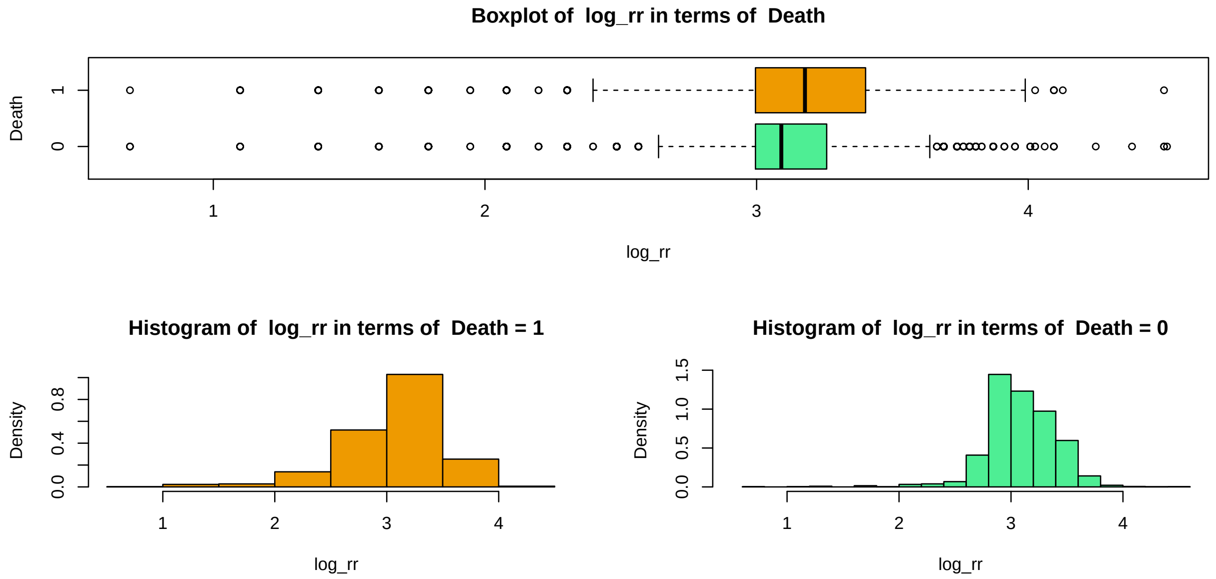


Figure 18: Distribution of log transformation of rr (Respiratory Rate) in terms of death

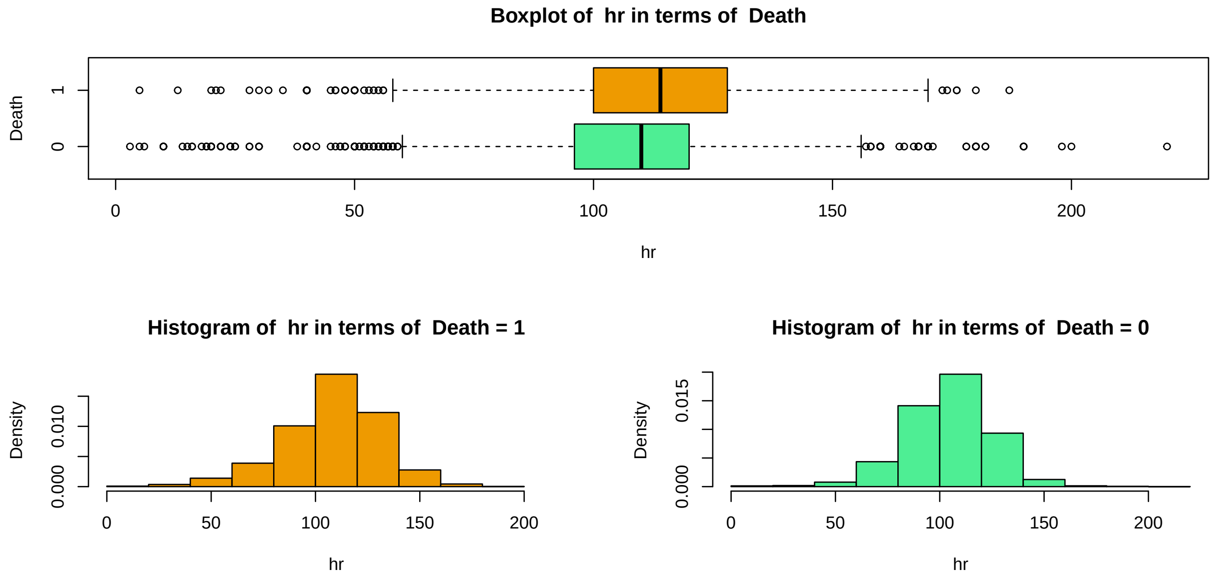


Figure 19: Distribution of hh (hearth Rate) in terms of death

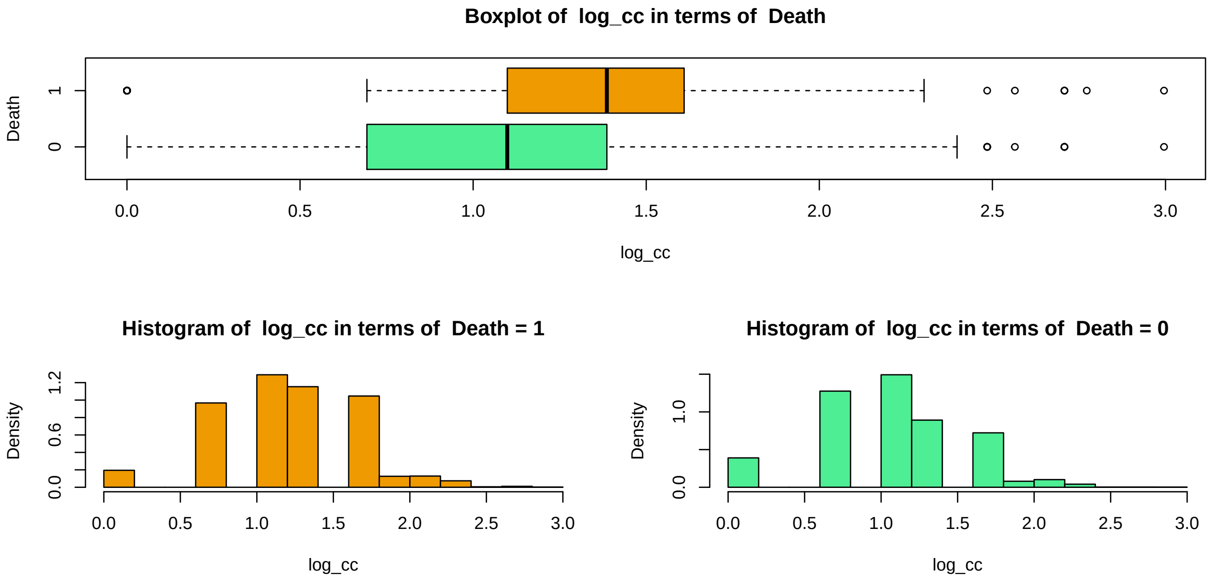


Figure 20: Distribution of log transformation of cc (Central capillary) in terms of death

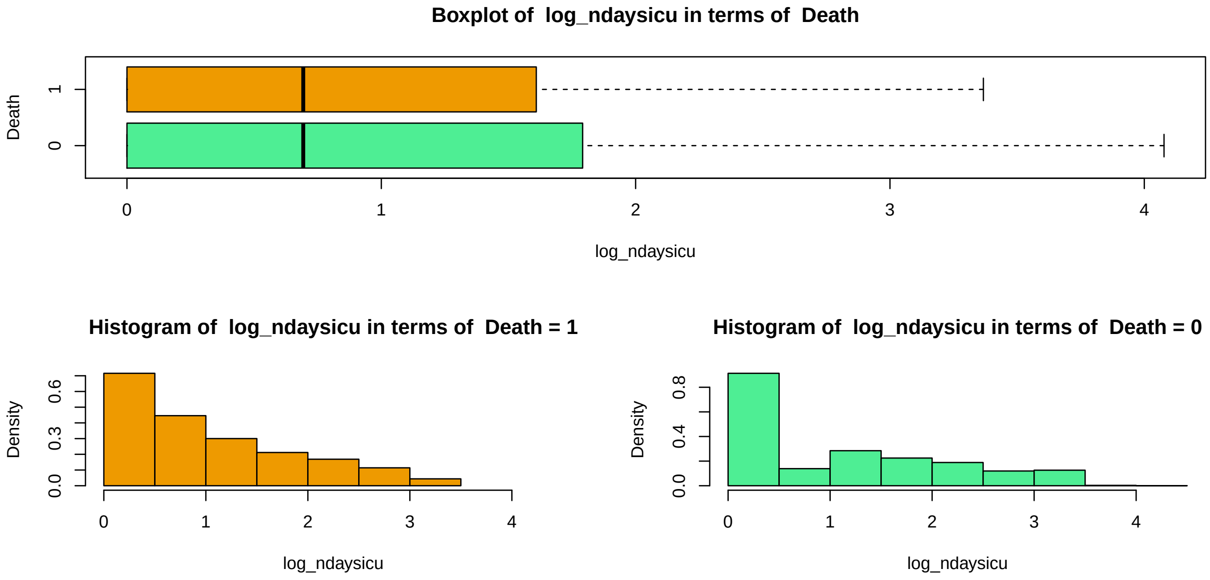


Figure 21: Distribution of log ndaysicu in terms of death

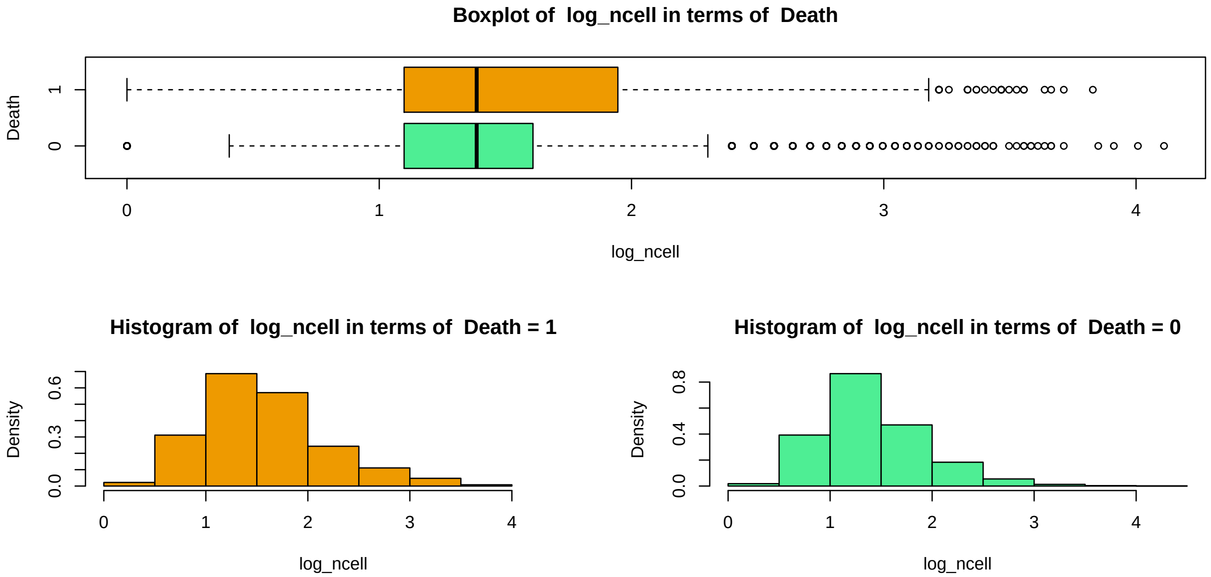


Figure 22: Distribution of log ncell in terms of death

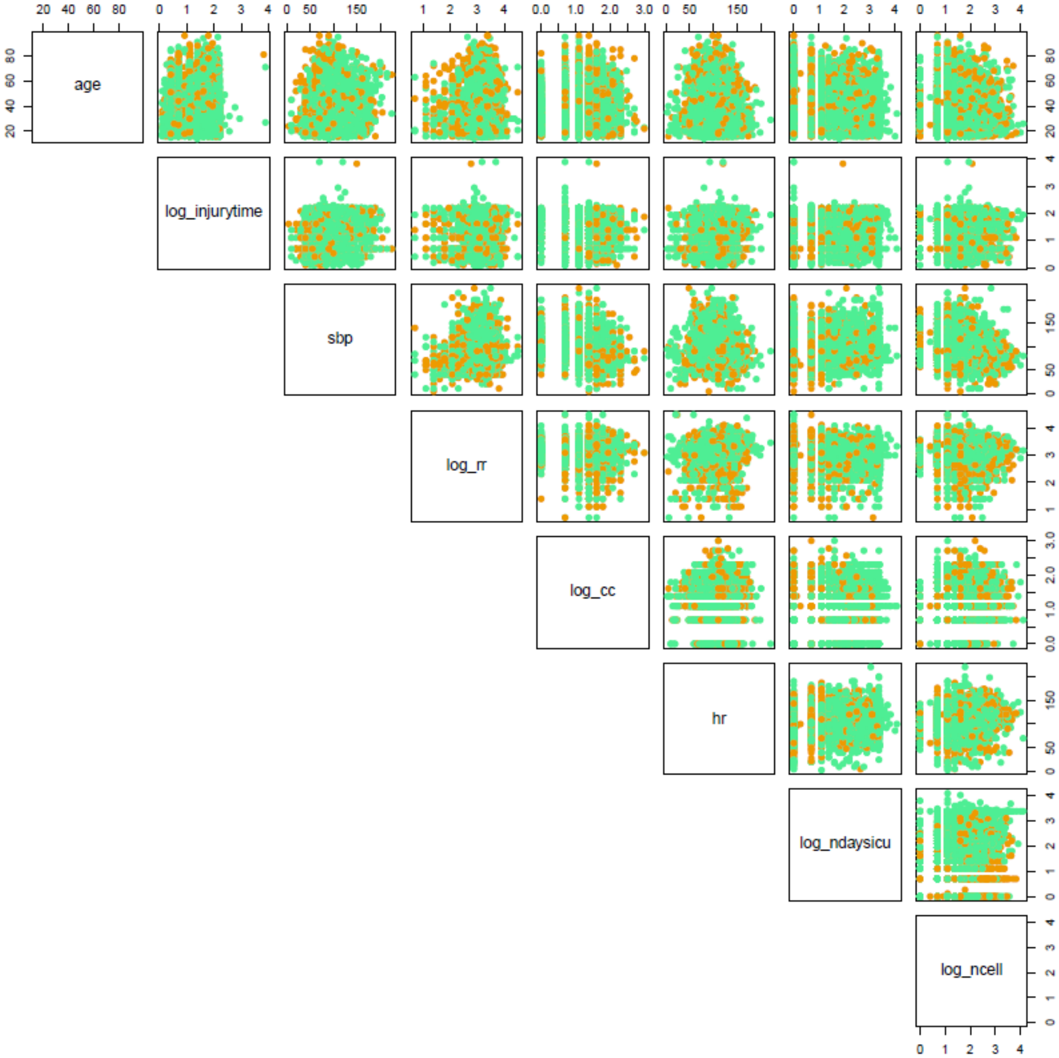


Figure 23: Scatter plot of all quantitiative variables

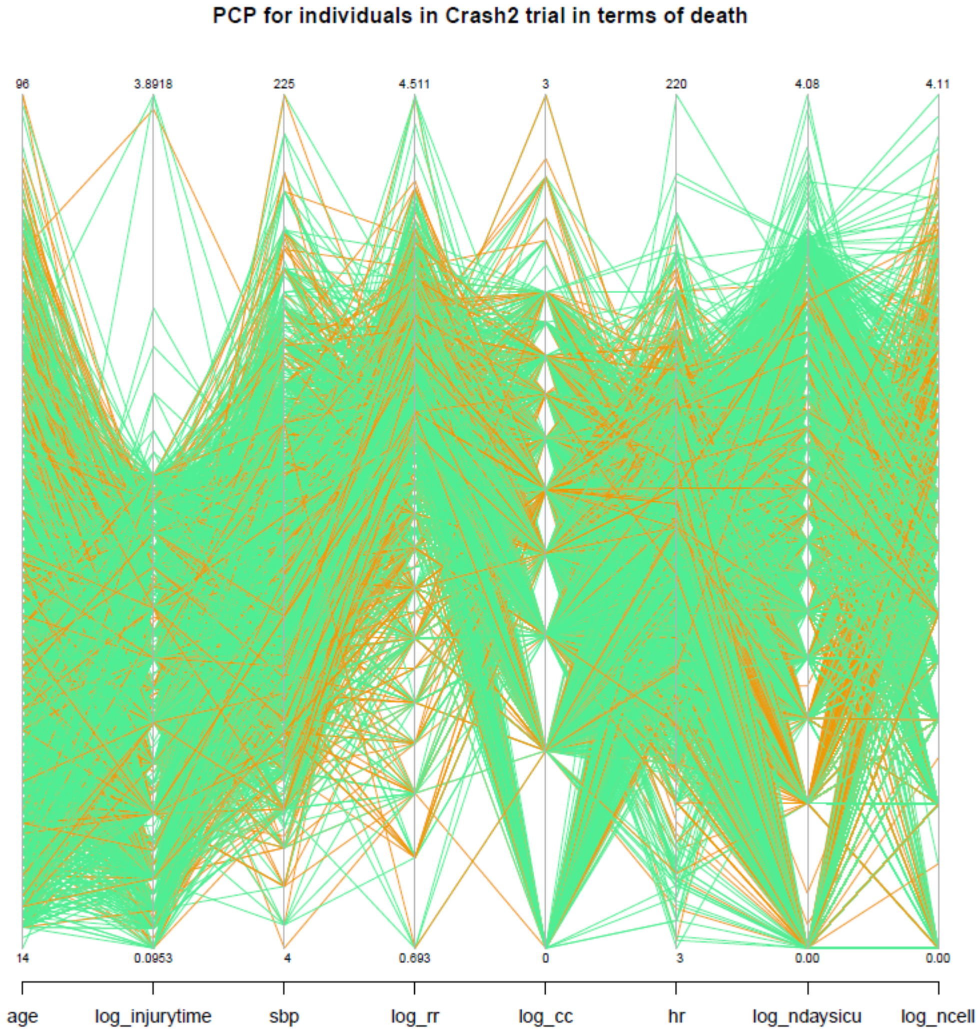


Figure 24: PCP plot of all quantitiative variables

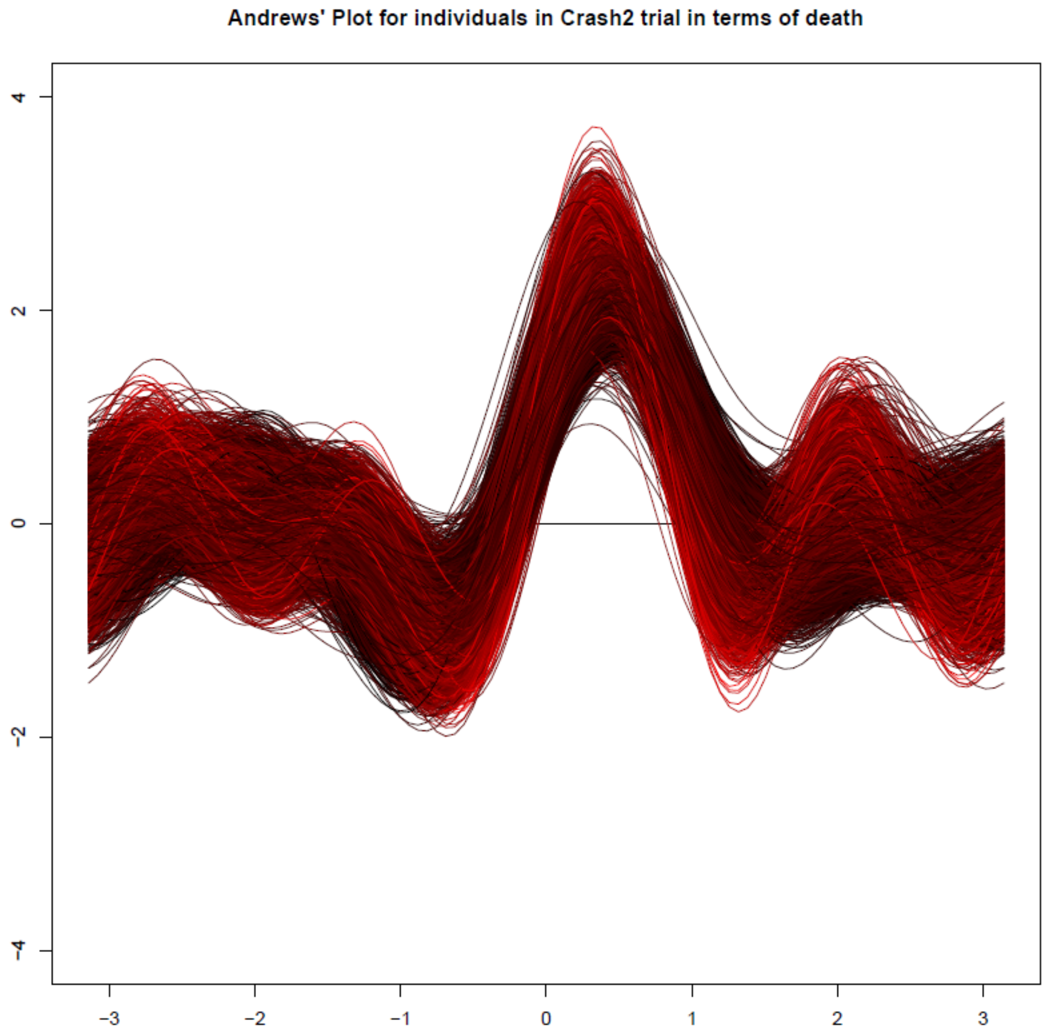


Figure 25: Andres' plot of all quantitiative variables