Team Project Multivariate Analysis

Adrian White, Cesar Conejo, Xavier Bryant

11/14/2020

Team members

- Adrian White: 1004391004
- Cesar Conejo: 100443596 (Representative member)
- Xavier Bryant: 100445659

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:

- entryid: (Numerical) Unique Numbers for Entry Forms
- sex: (Boolean) The sex of the patient (Male/Female)
- age : (Numerical) Age of the patient(Years)
- injurytime: (Numerical) Hours since injury (Hours)
- injurytype: (Categorical) Type of injury {Blunt, Penetrating, Blunt and Penetrating}

- sbp: (Numerical) Systolic Blood Pressure (mmHg)
- rr: (Numerical) Respiratory Rate (rate per minute)
- cc: (Numerical) Central Capillary Refille Time (seconds)
- hr: (Numerical) Heart Rate (rate per minute)
- ndaysicu: (Numerical) Number of days in ICU (days)
- btransf: (Boolean) Blood Products Transfusion
- ncell: (Numerical) Number of Units of Red Call Products Transfused
- nplasma: (Numerical) Number of Units of Fresh Frozen Plasma Transfused
- nplatelets: (Numerical) Number of Units of Platelets Transfused
- ncryo: (Numerical) Number of Units of Cryoprecipitate Transfused
- bvii: (Boolean) Recombinant Factor VIIa Given
- Death: (Boolean) Indicator if the patient survived after the procedure
- bloading: (Boolean) Complete Loading Dose of Trial Drug Given

A summary of the data type is the following:

variable	type_variable	sub_type_variable
entryid	Quantitative	Continuous
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
btransf	Quanlitative	Nominal
ncell	Quantitative	Discrete
nplasma	Quantitative	Discrete
nplatelets	Quantitative	Discrete
ncryo	Quantitative	Discrete
bvii	Qualitative	Nominal
death	Qualitative	Nominal
bloading	Qualitative	Nominal

Summary and Graphical display of individual variables

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

```
## 'data.frame':    9497 obs. of  18 variables:
##  $ entryid   : int  1 3 4 6 7 8 9 11 12 14 ...
##  $ sex       : Factor w/ 2 levels "male","female": 1 1 1 1 1 1 1 1 1 2 ...
##  $ age       : int  50 30 40 19 27 16 29 41 56 37 ...
##  $ injurytime: num  1 1 2 3 0.5 1 1 0.5 0.5 8 ...
##  $ injurytype: Factor w/ 3 levels "blunt","penetrating",..: 1 1 2 2 2 2 1 2 1 2 ...
##  $ sbp       : int  75 70 60 90 90 90 116 120 60 104 ...
##  $ rr        : int  28 26 20 30 26 28 15 15 9 23 ...
##  $ cc        : int  5 6 5 5 5 2 3 3 3 5 ...
##  $ hr        : int  120 130 120 90 96 118 118 70 100 92 ...
##  $ ndaysicu  : num  0 6 2 9 7 0 7 7 23 2 ...
##  $ btransf   : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ ncell     : num  1 2 4 2 1 1 16 8 4 4 ...
##  $ nplasma   : int  0 0 0 0 0 0 9 11 9 0 ...
##  $ nplatelets: int  0 0 0 0 0 0 22 10 0 0 ...
##  $ ncryo     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ bvii      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ death     : Factor w/ 2 levels "0","1": 2 1 2 2 1 1 1 1 1 1 ...
##  $ bloading  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

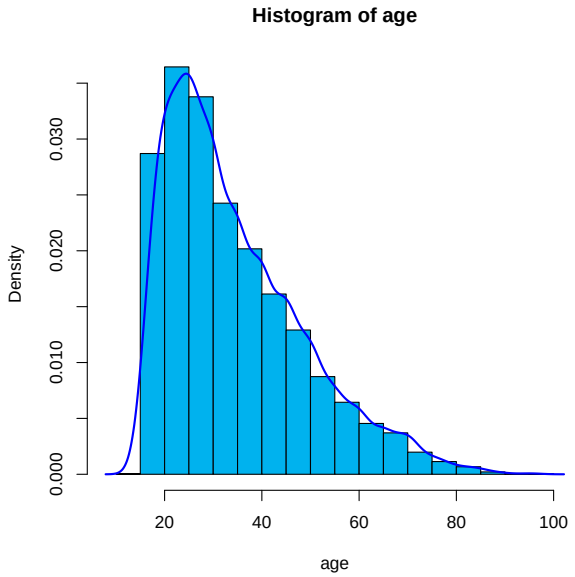
A summary of the values in the data set are:

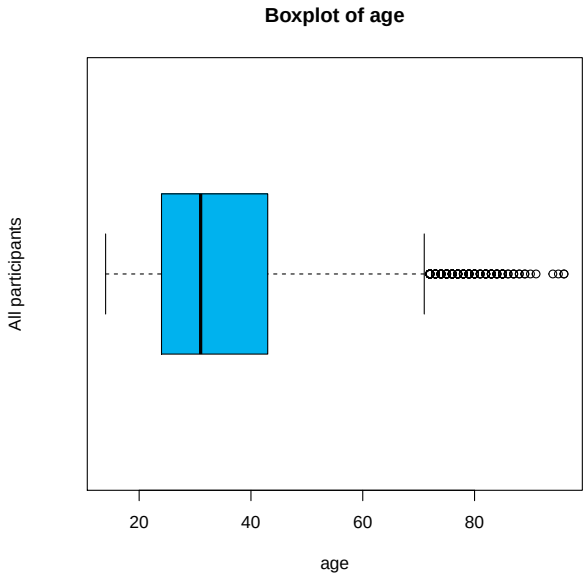
```
##     entryid          sex            age         injurytime
##  Min.   :    1   male  :7906   Min.   :14.0   Min.   : 0.10
##  1st Qu.: 4720   female:1591   1st Qu.:24.0   1st Qu.: 1.00
##  Median : 9333                 Median :31.0   Median : 3.00
##  Mean   : 9657                 Mean   :34.7   Mean   : 3.09
##  3rd Qu.:14598                 3rd Qu.:43.0   3rd Qu.: 4.50
##  Max.   :20270                 Max.   :96.0   Max.   :48.00
##                  injurytype        sbp              rr             cc
##  blunt                :5211   Min.   :  4.0   Min.   : 2.0   Min.   : 1.00
##  penetrating          :2937   1st Qu.: 80.0   1st Qu.:20.0   1st Qu.: 2.00
##  blunt and penetrating:1349   Median : 90.0   Median :22.0   Median : 3.00
##                               Mean   : 93.1   Mean   :23.5   Mean   : 3.44
##                               3rd Qu.:104.0   3rd Qu.:28.0   3rd Qu.: 4.00
##                               Max.   :225.0   Max.   :91.0   Max.   :20.00
##        hr          ndaysicu     btransf      ncell          nplasma
##  Min.   :  3   Min.   : 0.00   0:  12   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 96   1st Qu.: 0.00   1:9485   1st Qu.: 2.00   1st Qu.: 0.00
##  Median :110   Median : 1.00            Median : 3.00   Median : 0.00
##  Mean   :108   Mean   : 4.14            Mean   : 3.91   Mean   : 1.44
##  3rd Qu.:120   3rd Qu.: 5.00            3rd Qu.: 5.00   3rd Qu.: 1.00
##  Max.   :220   Max.   :58.00            Max.   :60.00   Max.   :60.00
##    nplatelets        ncryo       bvii     death    bloading
##  Min.   : 0.00   Min.   : 0.00   0:9456   0:7672   0:  39
##  1st Qu.: 0.00   1st Qu.: 0.00   1:  41   1:1825   1:9458
##  Median : 0.00   Median : 0.00
##  Mean   : 0.54   Mean   : 0.26
##  3rd Qu.: 0.00   3rd Qu.: 0.00
##  Max.   :87.00   Max.   :61.00
```

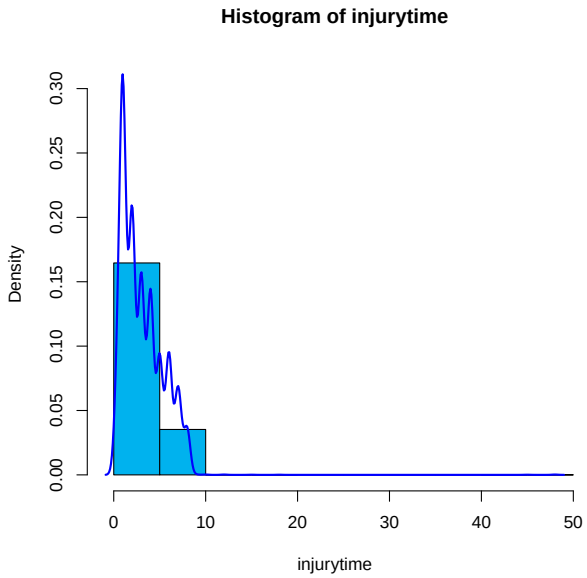
Finally, the list of different values by column is the following:

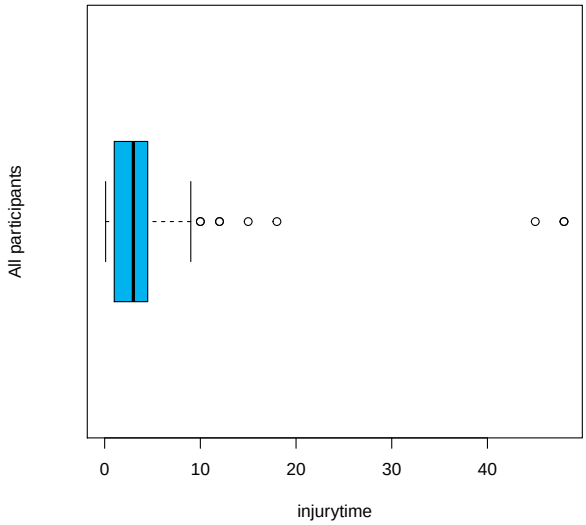
```
##    entryid        sex        age injurytime injurytype        sbp         rr
##       9497          2         81         78          3        153         58
##         cc         hr   ndaysicu    btransf      ncell    nplasma nplatelets
##         16        154         47          2         47         45         39
##      ncryo       bvii      death   bloading
##         28          2          2          2
```

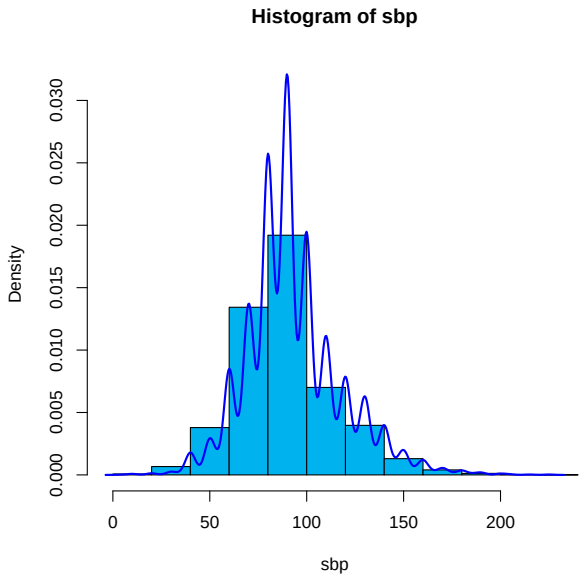
Some visualizations of the distributions of the quantitative variables are:

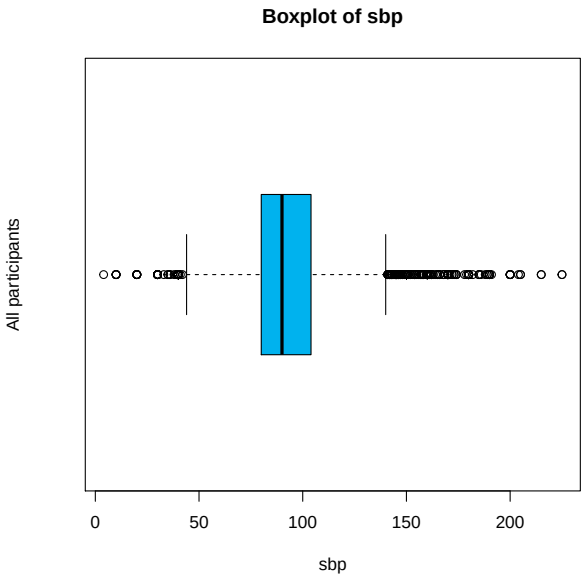


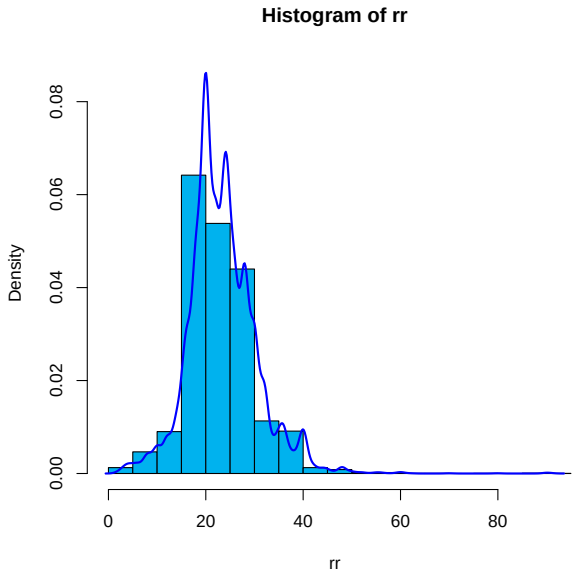


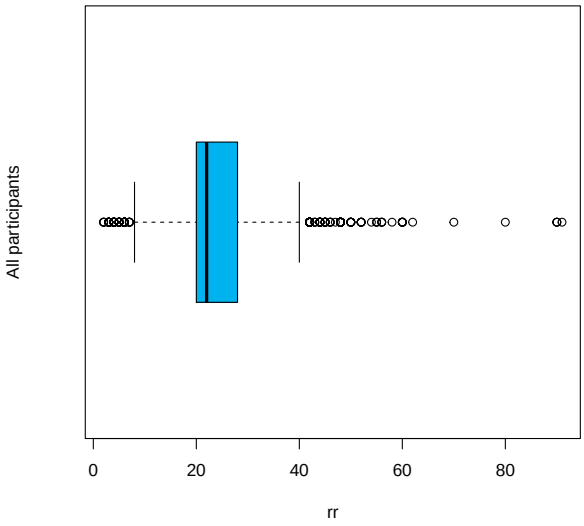


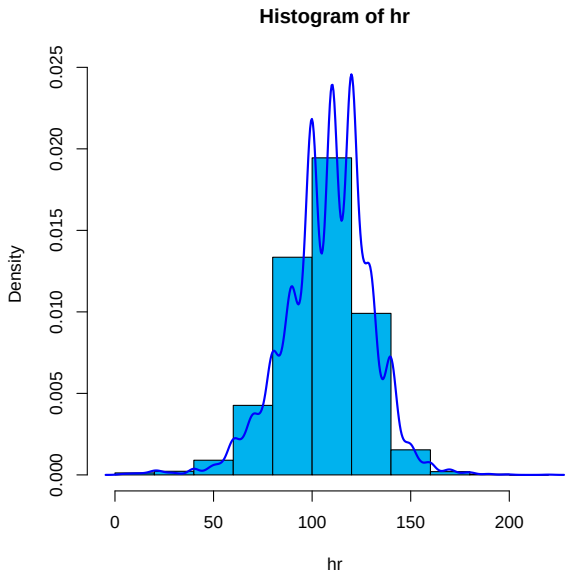


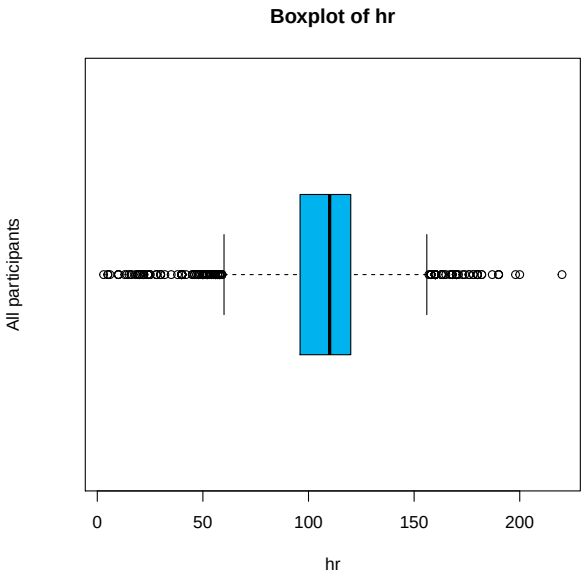

Histogram of sbp
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
50
100
150
200
sbp

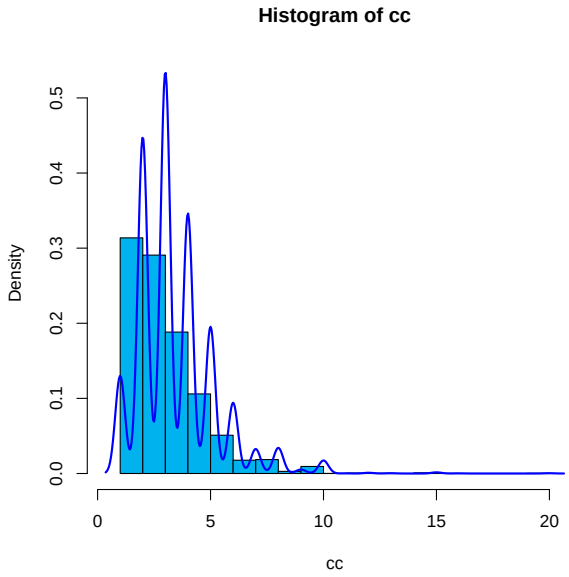

Boxplot of sbp
All participants
0
50
100
150
200
sbp

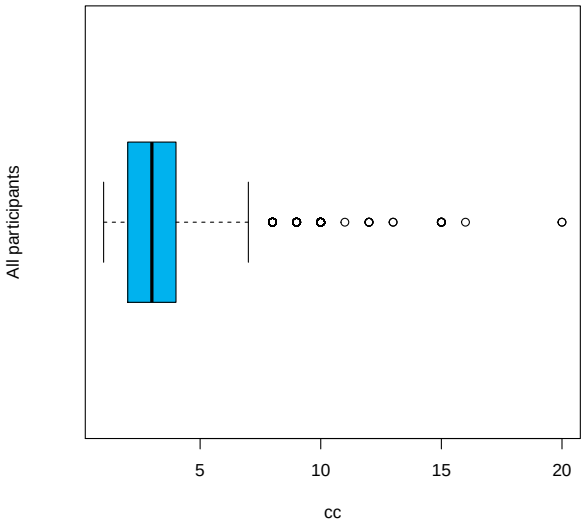

Histogram of rr
Density
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
rr

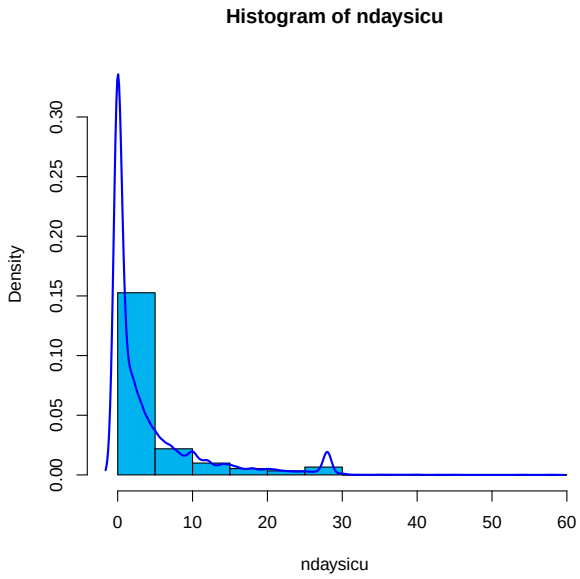

Boxplot of rr
All participants
0
20
40
60
80
rr

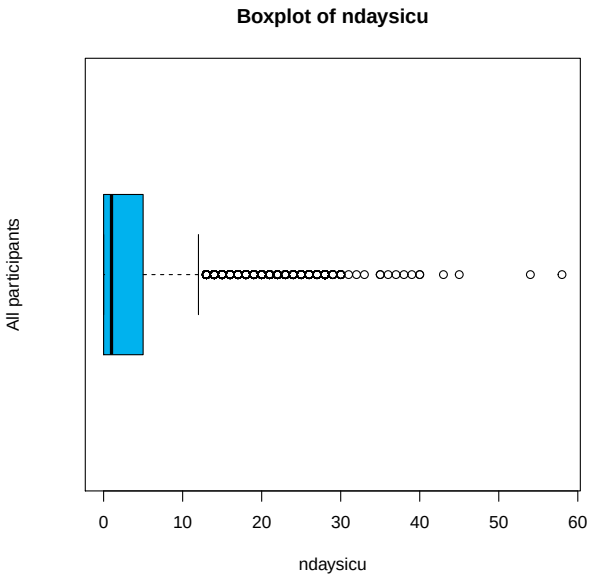

Histogram of hr
Density
0.000
0.005
0.010
0.015
0.020
0.025
0
50
100
150
200
hr

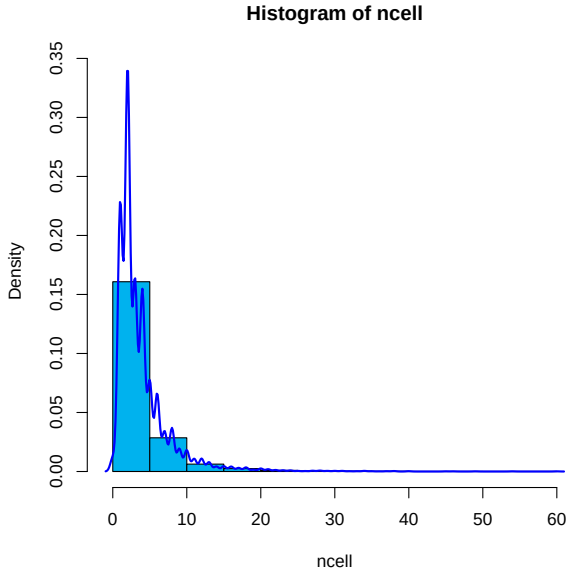

Boxplot of hr
All participants
0
50
100
150
200
hr

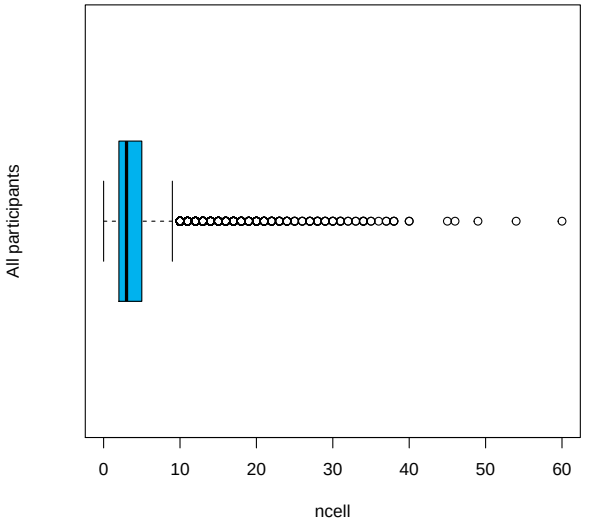

Histogram of cc
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
20
cc

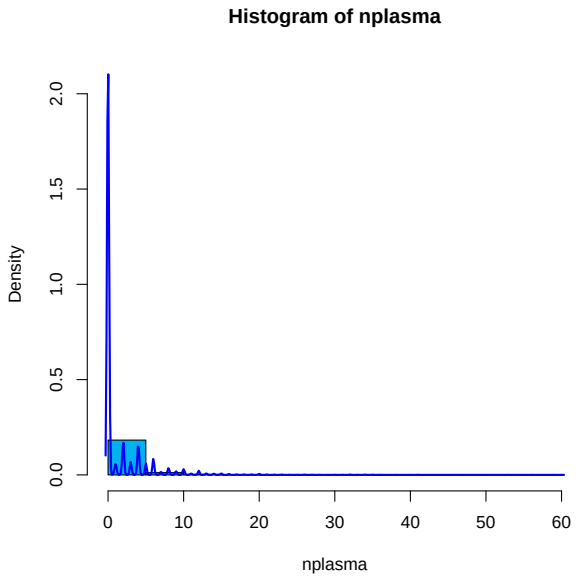

Boxplot of cc
All participants
5
10
15
20
cc

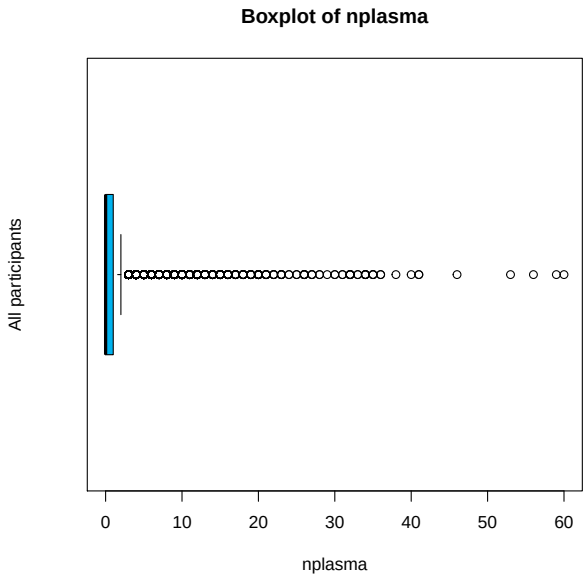

Histogram of ndaysicu
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
10
20
30
40
50
60
ndaysicu

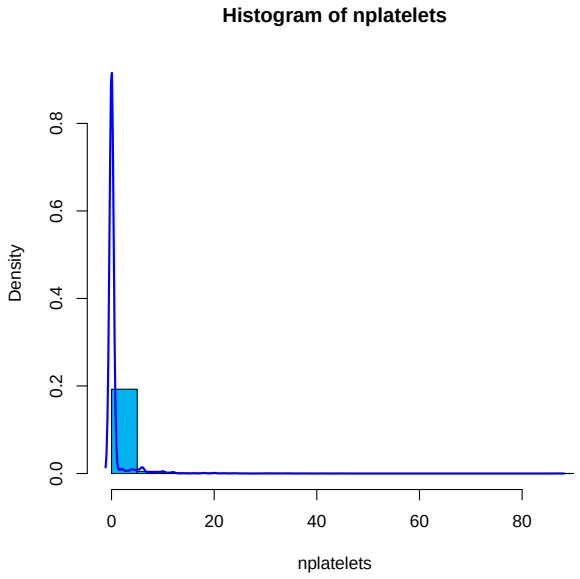

Boxplot of ndaysicu
All participants
0
10
20
30
40
50
60
ndaysicu

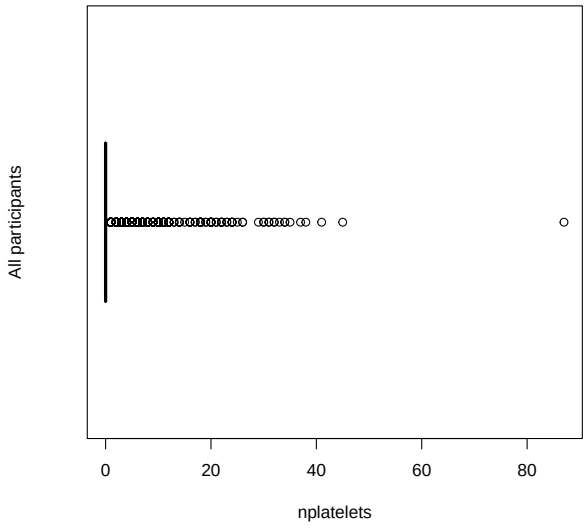

Histogram of ncell
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0
10
20
30
40
50
60
ncell

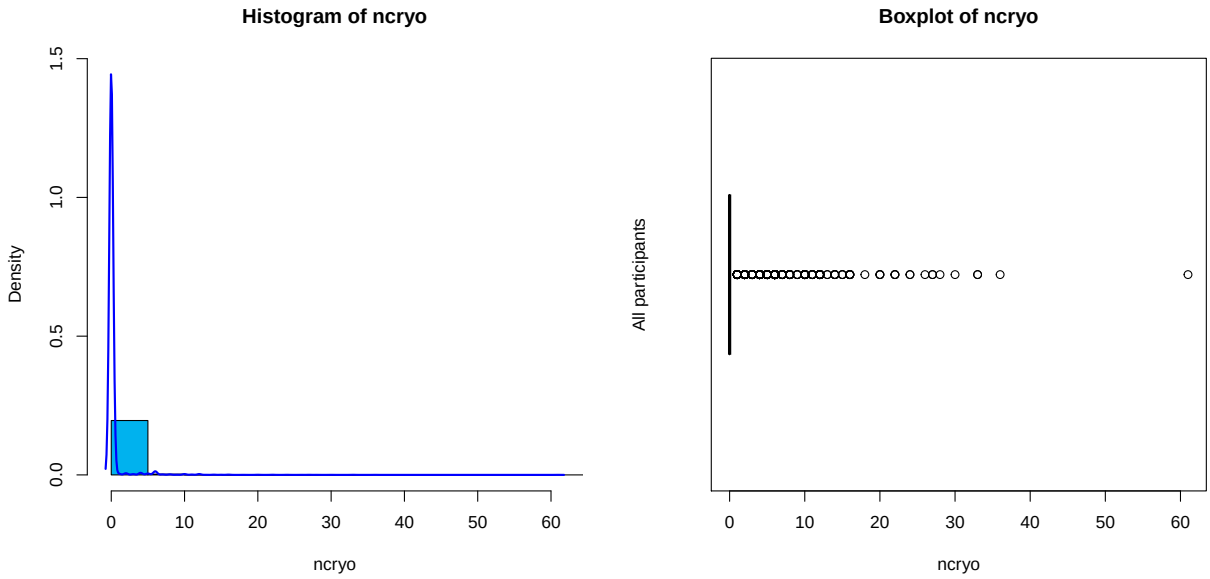

Boxplot of ncell
All participants
0
10
20
30
40
50
60
ncell


Histogram of nplasma
Density
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
50
60
nplasma


Boxplot of nplasma
All participants
0
10
20
30
40
50
60
nplasma


Histogram of nplatelets
Density
0.0
0.2
0.4
0.6
0.8
0
20
40
60
80
nplatelets


Boxplot of nplatelets
All participants
0
20
40
60
80
nplatelets



Discussion on the individual variable distributions

- **age:** Age 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.
- **sex:** Sex is heavily weighed to males, similar to age, possibly demonstrating that men take more risk and work at higher risk occupations. increasing the likelihood of experiencing trauma that would involve large amounts of bleeding.
- **injurytime:** Injury time is weighted to the right, 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.
- **injurytype:** Shows that the majority of patients have a blunt trauma, at 5211, then 2937 have a penetrating injury and 1349 with both.
- **sbp:** It seems that Sbp (Systolic Blood Pressure) 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.
- **rr:** Respiratory rate appears, similar to sbp, resembling a moderately balanced distribution around 22 respirations per minute, although is weighted more to the right.
- **hr:** Heart rate also seems fairly balanced at around 110, similar to the variables above, like rr and sbp.
- **cc:** Central capillary refill has 75% of the observations below of 4. However, the distribution is right-skewed.
- **ndaysicu:** 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.
- **ncell:** The distribution is weighted to the left with a median of 3, with many patients, only needing a small number of or zero units of red cell products transfused.

- **nplasma:** Similar to ncell, with, in fact, a median at 0, the distribution is weighted to the left and therefore many patients only need a small number of or zero units of plasma transfused. The kernel density is at zero and very tightly distributed on zero.
- **nplatelets:** Even more extreme than nplasama, the kernel density distribution is centered at 0 with only a number of values above zero for nplatelets. Clearly, most patients do not need platelets transfused.
- **ncryo:** A very similar kernel density distribution to nplatelets, centered around 0 with only a few patients requiring cryoprecipitate transfused.
- **bvii:** The value for bvii is 9456 for 0 and 41 for 1, showing that the vast majority of patients do not receive recombinant factor VIIa.
- **death:** Shows that most patients survive the trail with 7672 surviving in comparison to 1825 who did not.
- **bloading:** Shows the patients that receive the complete dose of the trial drug.

Graphical visulations of quantitative variable relationships

1. **Scatterplot Matrix of quantitative variables**
2. **Scatterplot Matrix of quantitative variables by sex**

We separated our quantitative variables into two sets of scatter-plot matrices as we have too many variables to do one set. It is hard to see any definite relationships between variables for any set of variables. We can see that some variables, as we discussed above, are very unbalanced, and therefore the observations are concentrated in specific regions of the scatter-plot like for the plots of *ncryo*, *nplatelets*, and more. Other distributions, almost appear as "shotgun" distributions with no distinct grouping such as those for age in relation to *sbp*, *rr*, *cc*, and hr as well as for hr and other variables, for example. These relationships are sensible for more balanced distributions as they are not weighted or skewed in a certain area of the plot. We do not see any remarkable difference in terms of sex.

3. **Andrews plot**

Andrew's plot agrees with our observation above. We do see a difference in the distributions of sex in the Andrews' plot, but not very substantially.