W271 Spring 18: Lab 2

Alyssa Eisenberg, Jeffrey Hsu, Gerard Kelly

Alcohol Consumption, Self-Esteem and Romantic Interactions

Introduction

The researchers stated the hypothesis as follow: *"We hypothesized that negative interactions with romantic partners would be associated with alcohol consumption (and an increased desire to drink). We predicted that people with low trait self-esteem would drink more on days they experienced more negative relationship interactions compared with days during which they experienced fewer negative relationship interactions. The relation between drinking and negative relationship interactions should not be evident for individuals with high trait self-esteem."*

EDA

```
library(car); require(dplyr); library(Hmisc); library(mcprofile); library(ggplot2); library(gridExtra);
dehart <- read.table(file="DeHartSimplified.csv", header=TRUE, sep=",")
#describe(dehart) #with a 10-page limit, should we include this type of output?
```

The dataset contains 623 observations of 13 variables representing entries in records kept by study participants. The variable **id** is a numeric identifier for each of the 89 study participants. Each participant recorded entries for seven consecutive days, indexed by the **studyday** variable, with the **dayweek** variable indicating which days of the week these correspond to (Monday = 1). The variable **gender** takes on one of two values according to whether the participant is male (1) or female (2); about 56% of the participants are female.

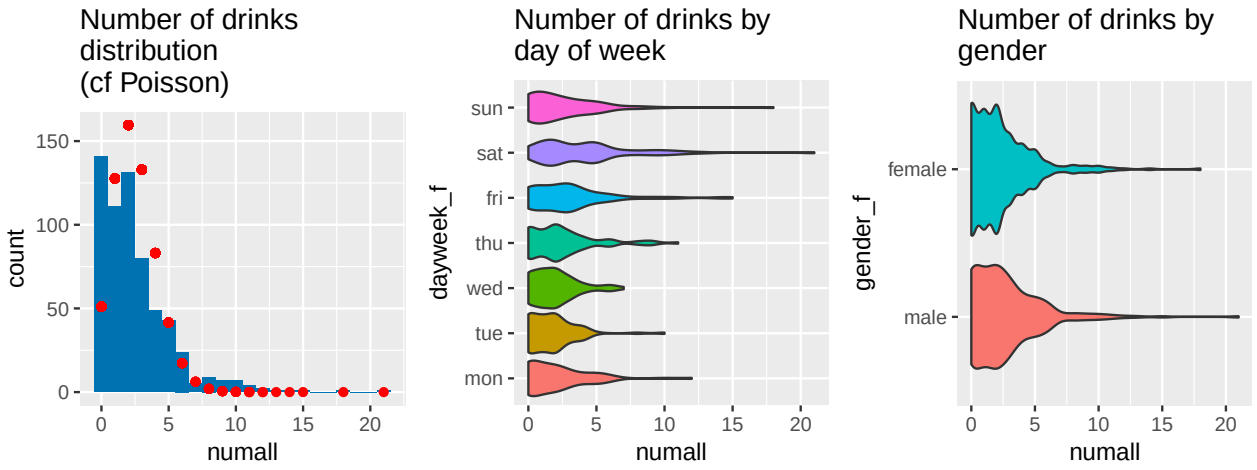
```
dehart$dayweek_f <- factor(dehart$dayweek); levels(dehart$dayweek_f) = c("mon","tue","wed","thu","fri",
dehart$gender_f <- factor(dehart$gender); levels(dehart$gender_f) = c("male","female")
```

For each of the seven days, participants record the number of drinks consumed with the integer **numall** count variable. There is one missing value. Values range from 0 to 15 with single outliers at 19 and 21. Observations are concentrated in the range 0 to 5. The sample mean and sample variance are 2.52 and 2.66 respectively. The median number of drinks is 3 for Friday, 4 for Saturday and 2 for all other days. There is a pronounced positive skew to the number of drinks for Fridays, Saturdays and Sundays.

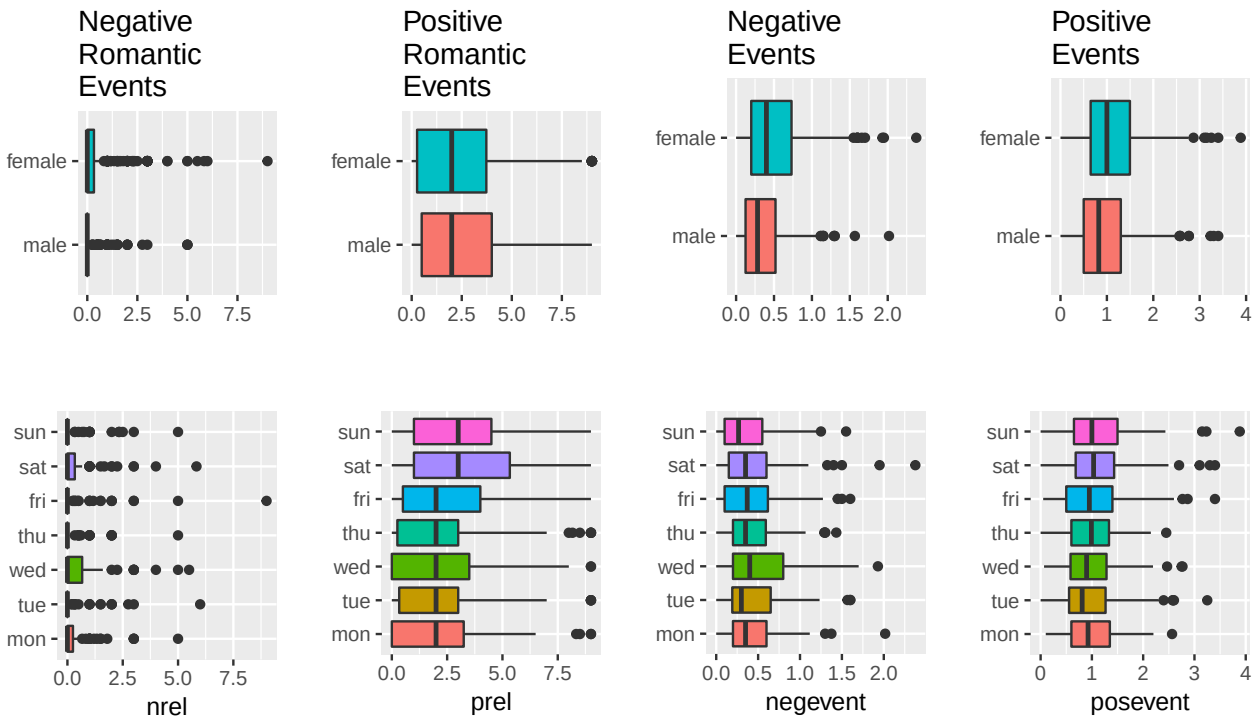
The variables **nrel** and **prel** are index measures for the negative and positive romantic-relationship events experienced by the subject each day (ranging from 0 to around 9), while **negevent** and **posevent** are index values combining the total number and intensity of negative and positive events experienced that day (ranging from 0 to around 4). There are no missing values for these variables. Distributions (and scales) for these measurements are very different, although tend to be similar for males and females, and similar across days of the week, although **prel** has signifiantly higher average (mean and median) values on Saturdays and Sundays. Distributions for all these variables are strongly positively-skewed, but this is particularly extreme for **nrel**, where a small fraction of outlying observations accounts for almost all of the variation.

```
#mean(dehart$numall, na.rm=TRUE); sd(dehart$numall, na.rm=TRUE)
#aggregate(numall ~ dayweek_f, data = dehart, FUN = function(x) c(m = mean(x), n = median(x)))
#The distribution of 622 times a Poisson random variable with parameter lambda is overlaid on the histo
lambda = 2.5
p1 <- ggplot(na.omit(dehart), aes(x = numall)) + geom_histogram(aes(y = ..count..), binwidth = 1, fill=
  geom_point(aes(y = 622*dpois(x = numall,lambda)), color = "red")+ ggtitle("Number of drinks\ndistribu
```

```
p2<-ggplot(na.omit(dehart), aes(dayweek_f, numall)) + geom_violin(aes(fill = dayweek_f)) + ggtitle("Numb
p3<-ggplot(na.omit(dehart), aes(gender_f, numall)) + geom_violin(aes(fill = gender_f)) + ggtitle("Numbe
grid.arrange(p1, p2, p3, ncol = 3)
```




```
p1a<-ggplot(dehart, aes(gender_f, nrel)) + geom_boxplot(aes(fill = gender_f)) + labs(x = "", y = "") + g
p1b<-ggplot(dehart, aes(dayweek_f, nrel)) + geom_boxplot(aes(fill = dayweek_f)) + labs(x = "") + ggtitle
p2a<-ggplot(dehart, aes(gender_f, prel)) + geom_boxplot(aes(fill = gender_f)) + labs(x = "", y = "") + g
p2b<-ggplot(dehart, aes(dayweek_f, prel)) + geom_boxplot(aes(fill = dayweek_f)) + labs(x = "") + ggtitle
p3a<-ggplot(dehart, aes(gender_f, negevent)) + geom_boxplot(aes(fill = gender_f)) + labs(x = "", y = ""
p3b<-ggplot(dehart, aes(dayweek_f, negevent)) + geom_boxplot(aes(fill = dayweek_f)) + labs(x = "") + gg
p4a<-ggplot(dehart, aes(gender_f, posevent)) + geom_boxplot(aes(fill = gender_f)) + labs(x = "", y = ""
p4b<-ggplot(dehart, aes(dayweek_f, posevent)) + geom_boxplot(aes(fill = dayweek_f)) + labs(x = "") + gg
grid.arrange(p1a, p2a, p3a, p4a, p1b, p2b, p3b, p4b, ncol = 4)
```



```
aggregate(nrel ~ dayweek_f, data = dehart, FUN = mean)
```

```
##   dayweek_f      nrel
## 1       mon 0.3471910
## 2       tue 0.2882022
## 3       wed 0.5541466
## 4       thu 0.2295880
## 5       fri 0.4014981
## 6       sat 0.4033708
## 7       sun 0.2893258
```

The **rosn** variable measures trait (long-term) self-esteem, a single measurement for each participant taken at the beginning of the study that does not change over the course of the seven days. This measurement ranges between 2 and 4, with a mean value around 3.4. Distributions for males and females differ, with a larger proportion of males recording lower values and a larger proportion of females recording higher values. The **age** variable measures age in years, ranging between 24.4 and 42.3 with a mean value of 34.3 and similar distributions for males and females but with a higher proportion of females recording higher values. Neither **rosn** nor **age** have any missing observations.

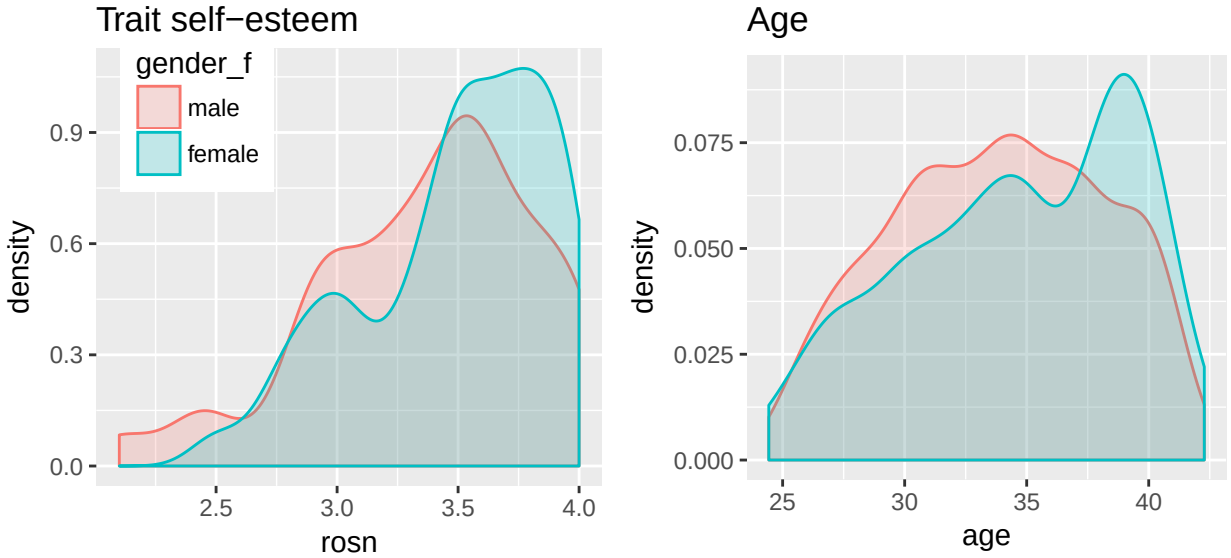
```
quantile(dehart$rosn[dehart$gender == 1])
```

```
##   0%  25%  50%  75% 100%
##  2.1  3.0  3.5  3.7  4.0
```

```
quantile(dehart$rosn[dehart$gender == 2])
```

```
##   0%  25%  50%  75% 100%
## 2.50 3.30 3.55 3.80 4.00
```

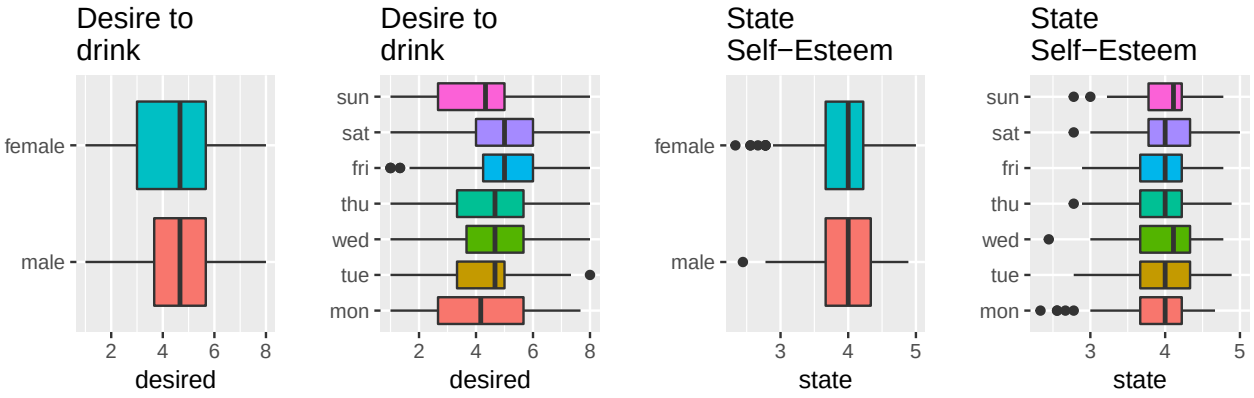
```
p1<-ggplot(dehart, aes(x = rosn, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2)+ ggtitl
p2<-ggplot(dehart, aes(x = age, fill = gender_f, colour = gender_f)) + geom_density(alpha=0.2)+
  ggtitle("Age")+theme(legend.position="none")
grid.arrange(p1, p2, ncol = 2)
```



The **desired** variable is a measure of the participant's recorded desire to drink, with values ranging between 1 and 8, a mean of 4.5 and a fairly symmetric distribution. with a significant share of responses at minimum and maximum values. Average values are slightly higher on average for males than for females and are highest on Friday and Saturday and lowest on Sunday and Monday. The **state** variable is a record of the participant's state (short-term) self-esteem as it varies each day. This ranges between 2 and 5 with a mean of

3.97 and a moderately negative-skewed distribution, little difference between males and females and little difference between days of the week (although with more negative outliers on Mondays). The variables **desired** and **state** both contain three missing values, two of these being from the same observation.

```
# aggregate(desired ~ gender_f, data = dehart, FUN = mean)
# aggregate(desired ~ dayweek_f, data = dehart, FUN = mean)
# aggregate(state ~ gender_f, data = dehart, FUN = mean)
# aggregate(state ~ dayweek_f, data = dehart, FUN = mean)
p1<-ggplot(na.omit(dehart), aes(gender_f, desired)) + geom_boxplot(aes(fill = gender_f)) + labs(x = "")
  ggtitle("Desire to\ndrink") + theme(legend.position="none") + coord_flip()
p2<-ggplot(na.omit(dehart), aes(dayweek_f, desired)) + geom_boxplot(aes(fill = dayweek_f)) + labs(x = "
  ggtitle("Desire to\ndrink") + theme(legend.position="none") + coord_flip()
p3<-ggplot(na.omit(dehart), aes(gender_f, state)) + geom_boxplot(aes(fill = gender_f)) + labs(x = "") +
  ggtitle("State\nSelf-Esteem") + theme(legend.position="none") + coord_flip()
p4<-ggplot(na.omit(dehart), aes(dayweek_f, state)) + geom_boxplot(aes(fill = dayweek_f)) + labs(x = "")
  ggtitle("State\nSelf-Esteem") + theme(legend.position="none") + coord_flip()
grid.arrange(p1, p2, p3, p4, ncol = 4)
```

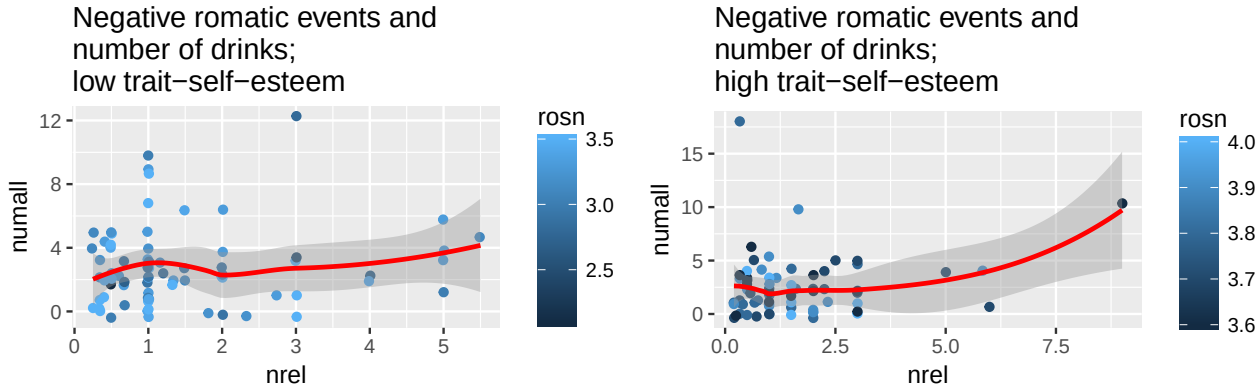


Bivarate relationships between these continuous variables are summarized in the below plot of correlation coefficients below. Most of these relationship are weak or insignificant. The strongest positive correlations are between **nrel** and **negevent** (negative events and negative romantic events), and between **prel** and **posevent** (positive events and positive romantic events). There is also a moderate positive correlation between the number of drinks and the desire to drink (**numall** and **desired**), and a mild positive correlation between trait and state self-esteem (**rosn** and **state**). State self-esteem has a weak positive correlation to positive event variables and a weak negative correlation to negative event variables, but no correlation with the number of drinks nor the desire to drink. Trait self-esteem has a weak negative correlation with the desire to drink but no correlation with the number of drinks. Age is uncorreated with any other variable.

```
data <- na.omit(dehart[,c(4,5,6,7,8,10,11,12,13)])
corr <- round(cor(data), 2)
ggcorrplot(corr, p.mat = cor_pmat(data),hc.order = TRUE, type = "lower", color = c("#FC4E07", "white",
```

The heavily skewed distribution for **nrel** indicates that the negative romatic relationship events relevant to the hypothesis are relatively infrequent. A subset of the dataset can be created for participant-days involving a non-zero rating on this variable. The relationship between **nrel** and numall can then be compared for participants with below-median trait self-esteem and above-median trait self-esteem. The relationship between negative romatic relationship events and the number of drinks appears to be stronger for individuals with higher trait-self esteem, however the range of this is driven by a relatively small number of individuals recording high **nrel** values that do not appear in the lower trait-self-esteem subset.

```
dehart_nrel = dehart[which(dehart$nrel != 0),]
p1 <- ggplot(na.omit(dehart_nrel[which(dehart_nrel$rosn<=3.5),]), aes(nrel, numall)) + geom_jitter(aes(
p2 <- ggplot(na.omit(dehart_nrel[which(dehart_nrel$rosn>3.5),]), aes(nrel, numall)) + geom_jitter(aes(co
grid.arrange(p1, p2,ncol = 2)
```



The same comparison can be made using state self-esteem rather than trait self-esteem, however since state-self-esteem is affected by negtive romantic relationship events, this comparson may be less pertinent to the hypothesis.

```
p1 <- ggplot(na.omit(dehart_nrel[which(dehart_nrel$state<=4),]), aes(nrel, numall)) + geom_jitter(aes(co
p2 <- ggplot(na.omit(dehart_nrel[which(dehart_nrel$state>4),]), aes(nrel, numall)) + geom_jitter(aes(col
grid.arrange(p1, p2,ncol = 2)
```