Simulate classification dataset

Contents

Purpose **1**

Simulate the data **1**

Boxplot and histogram **1**

References **3**

Purpose

This script shows how to simulate a dataset that can be used in classification problems.

In classification, the variable that we are measuring is continuous, while the variable that we are predicting is categorical.

Simulate the data

```
# From Harvard data science class (see references at the end of this notebook)

set.seed(11)

# Our measuring variable is continuous, numeric
x <- c(rnorm(30),rnorm(30,mean=2))

# Our outcome is categorical
y <- rep(c("A","B"),each=30)
```

Boxplot and histogram

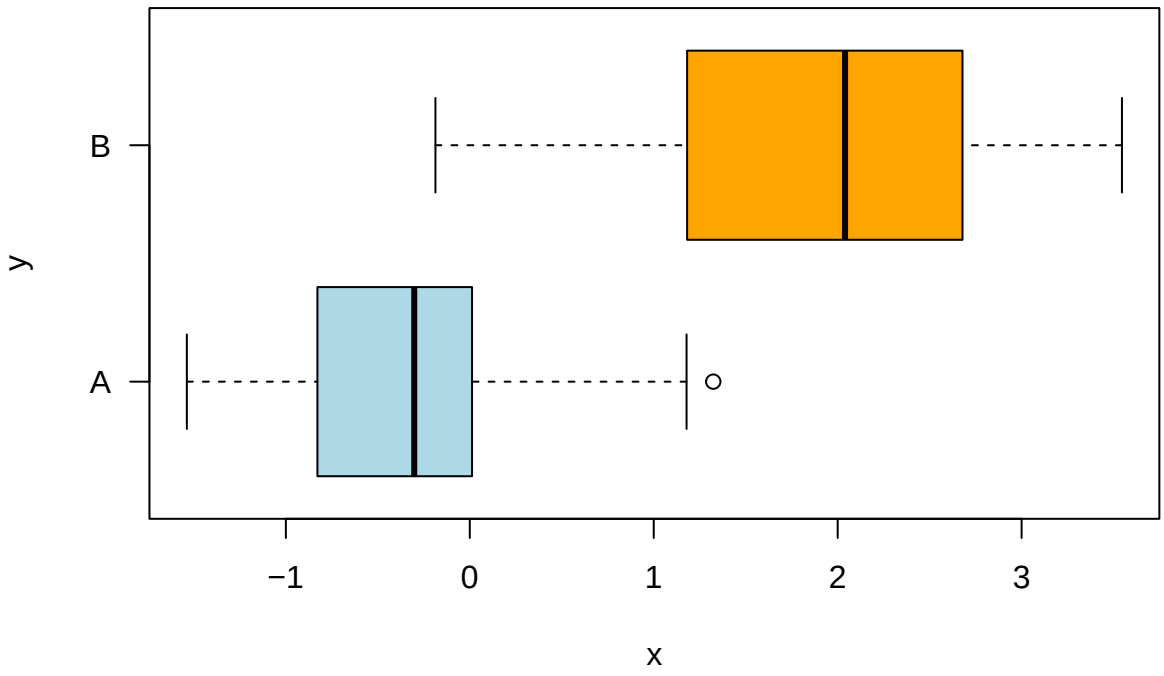
In a boxplot, we want to have the categorical variable in the horizontal axis.

That is why we see a formula x~y below.

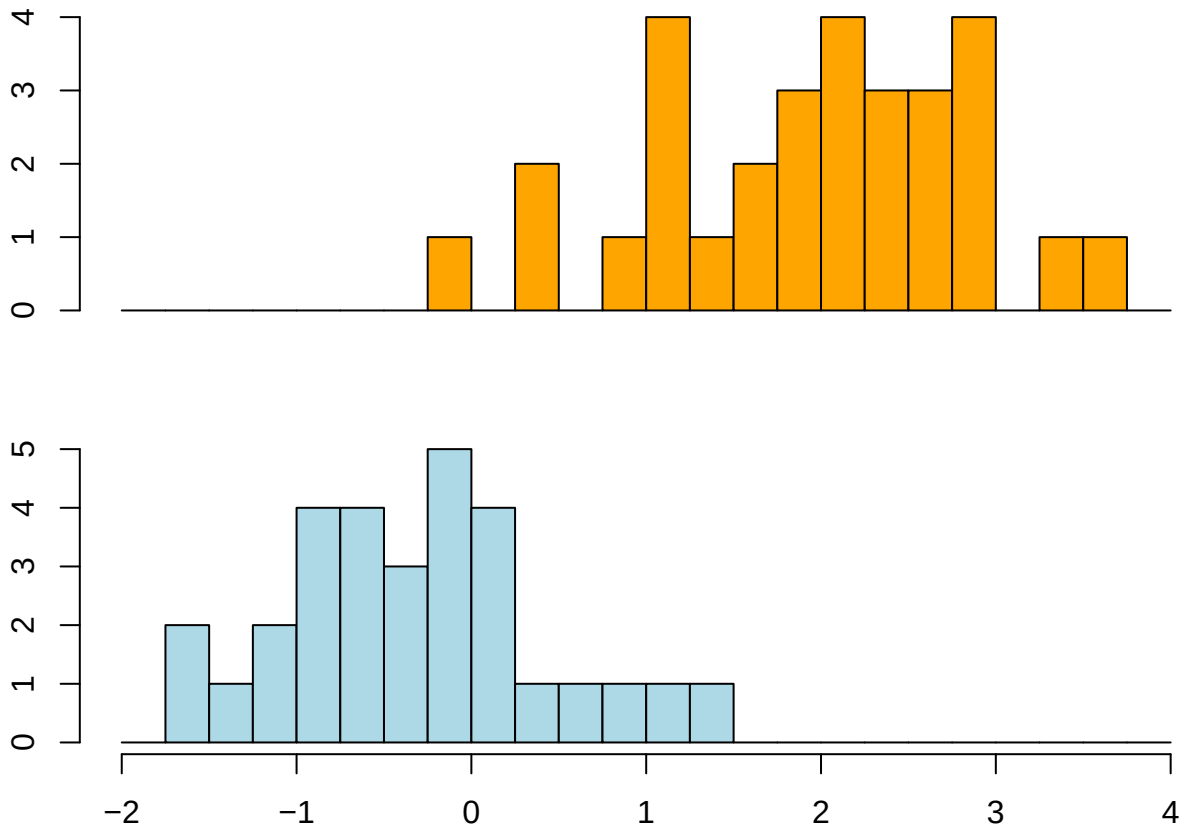
The chart gives us more information if we plot it horizontally in this case.

The histogram adds information to visualize the behavior relationship between the outcome, categorical valuable, and predictor, numeric variable.

```r
# From Harvard data science class (see references)
boxplot(x~y,col=c("lightblue","orange"), horizontal=T,las=1)
```




```r
oldpar <- par(mfrow=c(2,1),mar=c(2,2,1,1))
breaks <- seq(-2,4,by=0.25)
hist(x[y=="B"],breaks=breaks,col='orange',main="", xaxt='n')
hist(x[y=="A"],breaks=breaks,col='lightblue',main="")
```




```
par(oldpar)
```

References

- Harvard “Elements of Statistical Learning” (2021) taught by professors Dr. Sivachenko, Dr. Farutin