Simulate continuous variables dataset

Contents

Purpose

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

In regression, the variables that we are measuring are continuous, and the variable that we are predicting is continuous as well.

Dependent variables

Simulate the data

Create two Normally distributed datasets that have a relationship.

Play with the number of samples and we move the means around.

```
# From Harvard data science class (see references at the end of this notebook)
x <- rnorm(10000, mean=10, sd=sqrt(5))

# Initialize y with x...We would have a straight line if plotting y~x
y <- x

# Now inject variability to each, and we will not have a straight line exactly
x <- x + rnorm(10000, sd=2)
y <- y + rnorm(10000, sd=2)
```

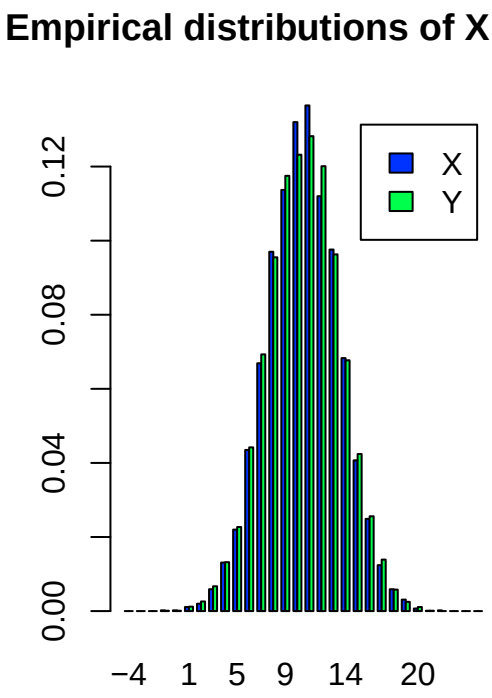
Plot histograms and scatterplot

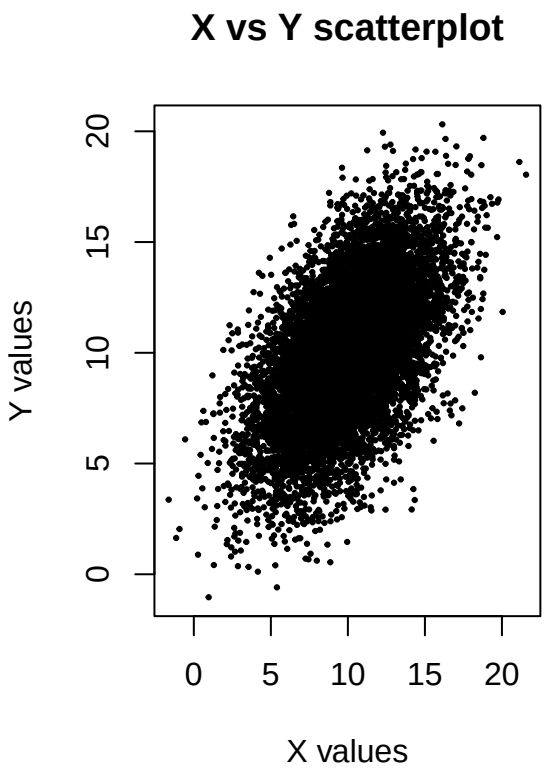
```
br<- -5:25 # set manually bins for histograms
# save histograms for X and Y , don't plot yet
hx <- hist(x, breaks=br, plot=F)
hy <- hist(y, breaks=br, plot=F)

# prepare 2 panels in one plot:
old.par <- par(mfrow=c(1,2))

# plot histograms side by side using rbind
barplot(rbind(hx$density,hy$density),
        beside=T,
        col=c(rgb(0,0.2,1), rgb(0,1,0.3)),
        legend=c('X','Y'),
        main='Empirical distributions of X and Y',
        names=br[-1])

# Scatter plot
plot(x,y,
     xlab='X values',
     ylab='Y values',
     main='X vs Y scatterplot',
     pch=19,
     cex=0.3)
```





```
# restore graphical attributes to previous values:
par(old.par)
```

Independent variables

Simulate the data

Create two independent Normally distributed datasets x and y.

Play with the number of samples and we move the means around.

```
# From Harvard data science class (see references at the end of this notebook)
# simulate sampling of 10000 values for X and for Y.
# We can play with the mean and sd. Shoud have same size to keep it balanced.
x <- rnorm(10000, mean=10, sd=3)
y <- rnorm(10000, mean=10, sd=3)
```

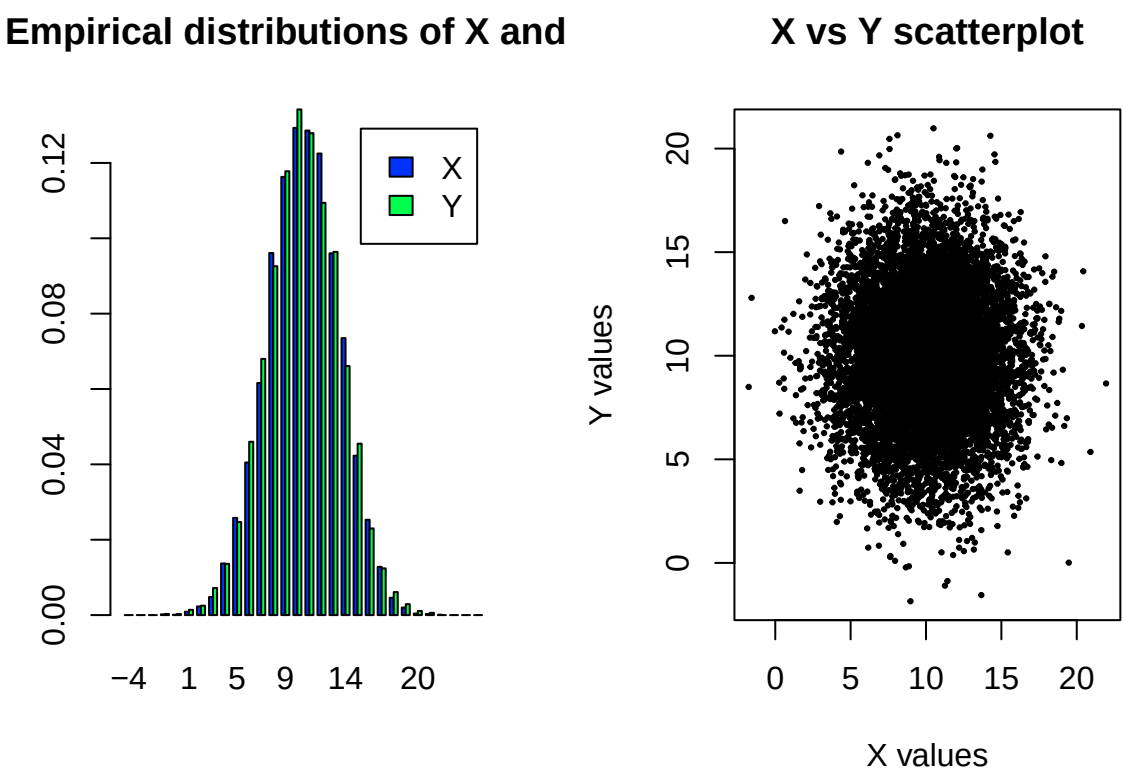
Plot histograms and scatterplot

```
br<- -5:25 # set manually bins for histograms
# save histograms for X and Y , don't plot yet:
hx <- hist(x, breaks=br, plot=F)
hy <- hist(y, breaks=br, plot=F)

# prepare 2 panels in one plot:
old.par <- par(mfrow=c(1,2))

# plot histograms side by side using rbind
barplot(rbind(hx$density,hy$density),
        beside=T,
        col=c(rgb(0,0.2,1), rgb(0,1,0.3)),
        legend=c('X','Y'),
        main='Empirical distributions of X and Y',
        names=br[-1])

# Scatter plot
plot(x,y,
     xlab='X values',
     ylab='Y values',
     main='X vs Y scatterplot',
     pch=19,
     cex=0.3)
```




```
# restore graphical attributes to previous values:
par(old.par)
```

References

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