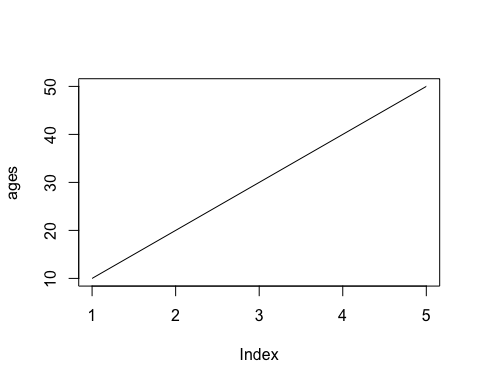
R Notebook

ages <- c(10, 20, 30, 40 , 50)

summary(ages)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10 20 30 30 40 50



# Work with the data set

## Import the data set

## Rows: 1000 Columns: 21  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (13): checking\_account, credit\_history, purpose, savings, present\_employ...  
## dbl (8): duration\_months, credit\_amount, installment\_rate, present\_residenc...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

summary(data$age\_years)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 19.00 27.00 33.00 35.55 42.00 75.00

table(data$purpose)

##   
## business car (new) car (used) domestic appliances   
## 97 234 103 12   
## education furniture/equipment others radio/television   
## 50 181 12 280   
## repairs retraining   
## 22 9

lm(credit\_amount ~ age\_years, data)

##   
## Call:  
## lm(formula = credit\_amount ~ age\_years, data = data)  
##   
## Coefficients:  
## (Intercept) age\_years   
## 2982.684 8.118

### Show measures of spread

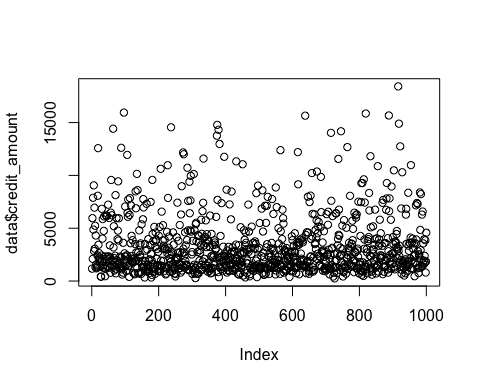
Provide a summary

summary(data$credit\_amount)

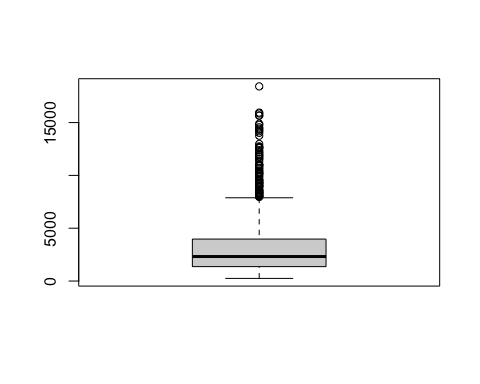
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 250 1366 2320 3271 3972 18424

Visualize the data to detect possible outliers

plot(data$credit\_amount)



boxplot(data$credit\_amount)



Basic formula of an outlier:

Mean - 2 \* sd

Mean + 2 \* sd

outlier\_min <- mean(data$credit\_amount) - (2\* sd(data$credit\_amount))  
outlier\_max <- mean(data$credit\_amount) + (2\* sd(data$credit\_amount))

print(outlier\_min)

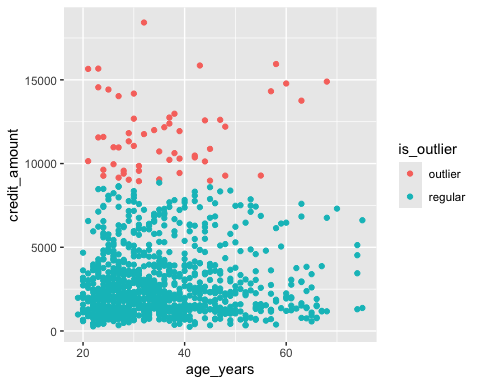
## [1] -2374.216

print(outlier\_max)

## [1] 8916.732

data$is\_outlier <- ifelse(data$credit\_amount < outlier\_min | data$credit\_amount > outlier\_max,  
 "outlier", "regular")

library(ggplot2)  
  
plot\_1 <- ggplot(aes(x=age\_years, y=credit\_amount), data = data)  
plot\_1 + geom\_point(aes(colour=is\_outlier))



mean(data$credit\_amount)

## [1] 3271.258

summary(data$credit\_amount)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 250 1366 2320 3271 3972 18424

### Percentiles

# use the 'quantile' function to calculate percentiles  
quantile(data$credit\_amount, 0.95)

## 95%   
## 9162.7

# Bi-variate statistics

### Using the Co-variance

cov(data$age\_years, data$credit\_amount)

## [1] 1050.523

cov(data$duration\_months, data$credit\_amount)

## [1] 21273.75

### Using the Correlation

cor(data$age\_years, data$credit\_amount)

## [1] 0.03271642

cor(data$duration\_months, data$credit\_amount)

## [1] 0.6249842

Causation research

model\_1 <- lm(credit\_amount ~ duration\_months, data=data)  
  
summary(model\_1)

##   
## Call:  
## lm(formula = credit\_amount ~ duration\_months, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5151.6 -1260.0 -432.9 653.2 13805.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 213.216 139.569 1.528 0.127   
## duration\_months 146.297 5.784 25.292 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2205 on 998 degrees of freedom  
## Multiple R-squared: 0.3906, Adjusted R-squared: 0.39   
## F-statistic: 639.7 on 1 and 998 DF, p-value: < 2.2e-16

# Binomial distribution

dbinom(6, 10, 0.5)

## [1] 0.2050781

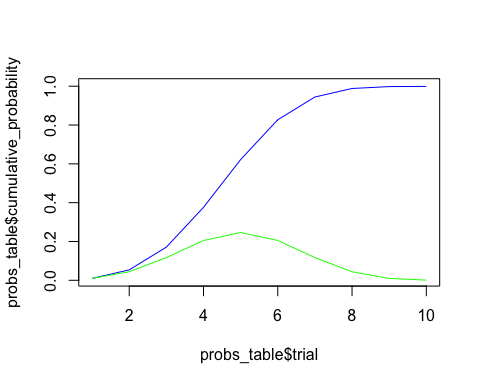
trials <- 1:10  
  
probabilities <- numeric(0)  
  
  
for (x in trials){  
 prob <- dbinom(x, 10, 0.5)  
 probabilities <- append(probabilities, prob)  
}  
  
probabilities

## [1] 0.0097656250 0.0439453125 0.1171875000 0.2050781250 0.2460937500  
## [6] 0.2050781250 0.1171875000 0.0439453125 0.0097656250 0.0009765625

Create a table to show the cumulative distribution

probs\_table <- data.frame(trial = trials,  
 probability = probabilities,  
 cumulative\_probability = cumsum(probabilities))

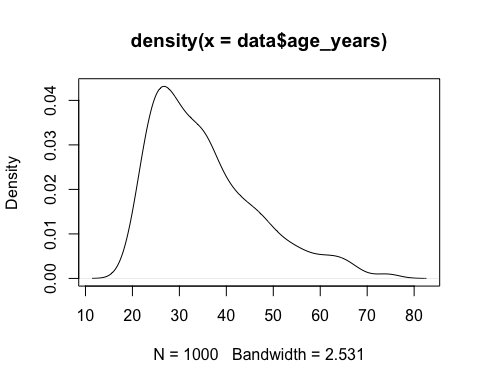
Plot the single probability



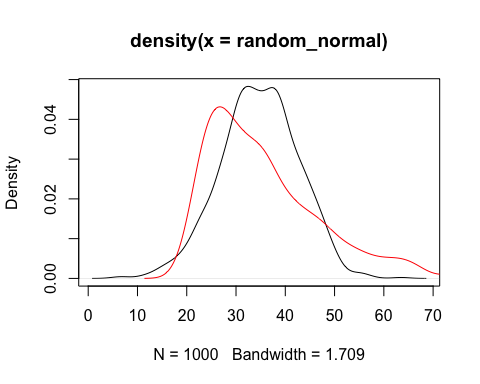
# Normal distribution

## Display the distribution in R

plot(density(data$age\_years))



random\_normal <- rnorm(1000, 35, 8)  
plot(density(random\_normal))  
lines(density(data$age\_years), col="red")



### Manually calculate a z-score

mean(data$age\_years)

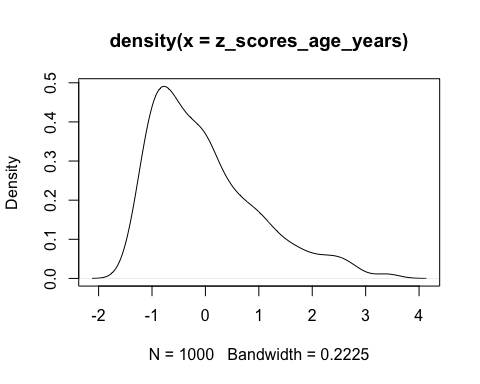
## [1] 35.546

(50 - mean(data$age\_years)) / sd(data$age\_years)

## [1] 1.270629

(35.546 - mean(data$age\_years)) / sd(data$age\_years)

## [1] -6.246272e-16



## Check if the data is normally distributed

**H0:** The data is normally distributed.

**H1:** The data is **not** normally distributed.

shapiro.test(data$age\_years)

##   
## Shapiro-Wilk normality test  
##   
## data: data$age\_years  
## W = 0.91747, p-value < 2.2e-16

p-value is significantly lower than 0.05, so this variable is **not** normally distributed.

This means, we cannot choose a parametric.

* (So not t-test, no anova, but the **non**-parametric equivalents for these tests).

## First statistical test (2 groups)

Test if the average age of people having a loan for new cars if (significantly) different then people having a loan for used cars.

### Hypotheses

**H0:** The average age of people with loans for used cars is the same.

**H1:** The average age of people with loans for used cars is the **not** same.

### Check for normality of the variable

# filter for only car-related purposes  
data\_cars <- data[data$purpose %in% c("car (new)", "car (used)"), ]  
table(data\_cars$purpose)

##   
## car (new) car (used)   
## 234 103

shapiro.test(data\_cars$age\_years)

##   
## Shapiro-Wilk normality test  
##   
## data: data\_cars$age\_years  
## W = 0.92305, p-value = 3.802e-12

The p-value, is lower than 0.05, so the data is not normally distributed. We cannot use a parametric test.

### Decide the confidence level (p-value) for our test

For this research we choose a confidence level of 95%, which means we have a critical boundary for the p-value of 0.05

### Decide/execute the test that will be used

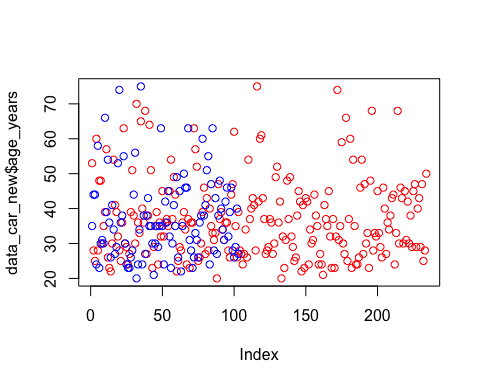
Based on the fact that the data is not normally distributed, we use for the alternative for a t-test, which is the Mann-Whitney U test, which is in R the wilcox.test

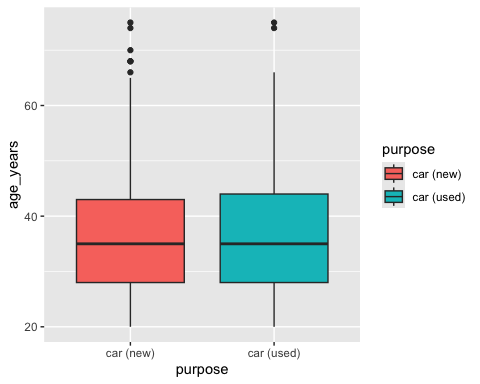
wilcox\_result <- wilcox.test(age\_years ~ purpose, data = data\_cars)  
wilcox\_result

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: age\_years by purpose  
## W = 12076, p-value = 0.9758  
## alternative hypothesis: true location shift is not equal to 0

wilcox\_result$p.value

## [1] 0.9757799





### Analyse results of the test

The p-value is 0.9758, which is larger than 0.05, so there is not enough evidence to reject the **H0** hypothesis.

### Draw conclusions (about our hypotheses) for our test

The average age between people for loans with used or new cars, is not significantly different.

### What would the t-test

t.test(data\_car\_used$age\_years,   
 data\_car\_new$age\_years)

##   
## Welch Two Sample t-test  
##   
## data: data\_car\_used$age\_years and data\_car\_new$age\_years  
## t = 0.11602, df = 187.84, p-value = 0.9078  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -2.603437 2.928804  
## sample estimates:  
## mean of x mean of y   
## 37.25243 37.08974