

Exercise set 1 - Week 35

Bjørn Christian Weinbach

20.08.2020

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

This notebook shows my solution to the exercise sheet for week 35 in STA510 course at the University of Stavanger.

Import libraries

```
library(ggplot2)    # Library for plotting
```

Problem 1

Consider the experiment of flipping an unbiased coin 5 independent times. Let H denote the outcome “head” and T to denote “tails”.

1. Compute the probability of HHTHT and THHHT.

We have two events, HHTHT and THHHT which are two events in the sample space of all possible outcomes of 5 coin flips, the probability $P(\text{HHTHT})/n$ where n is all possible outcomes.

n can be calculated using the **fundamental counting principle** which gives us:

$$n = 2 * 2 * 2 * 2 * 2 = 2^5 = 32 \quad (1)$$

Therefore, for each of the two events above, the probability of one of them occurring is $1/32$.

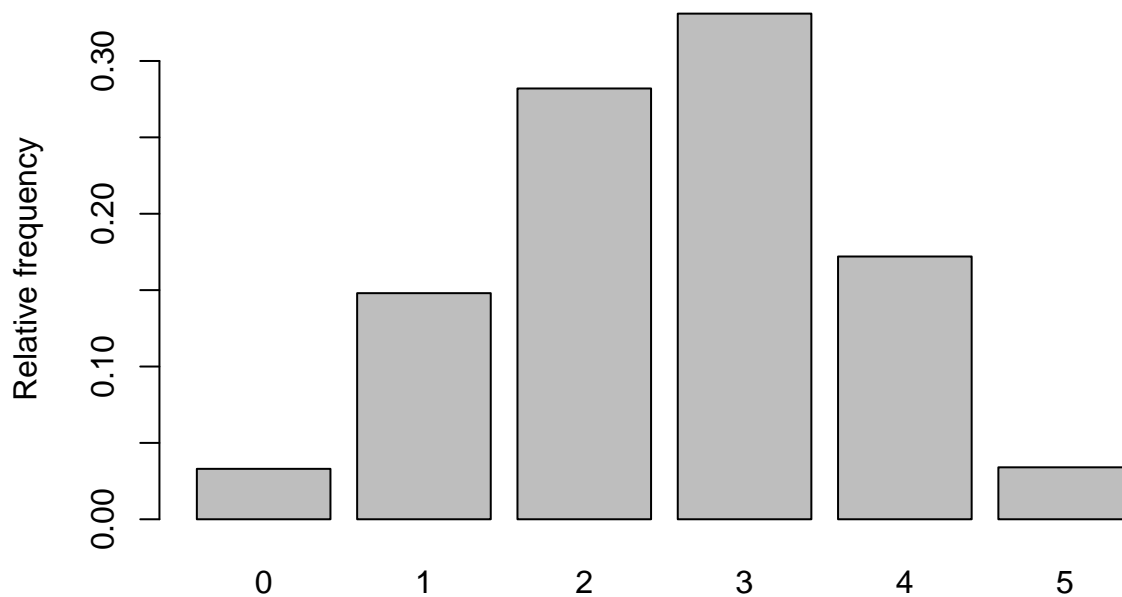
2. Modify the R-code from the number of heads example in the lectures to simulate the probability of three heads occurring in the 5 trials.

Function for simulationg number of heads

```
# Simulate the number of heads in k throws of a dice repeated Nsim times
# We let 0 correspond to tail and 1 correspond to head,
# The sample()-command below randomly draw a 0 or 1 with
# equal probability k times, and then sum() will summarize the number of heads
simheads <- function(Nsim,k){
  nheads <- vector(length=Nsim) # Define a vector to store the number of heads
  for(i in 1:Nsim) # Generate the number of heads Nsim times
    nheads[i] <- sum(sample(0:1,size=k,replace=TRUE)) # See explanation above
  return(nheads) # Return the resulting vector
}
```

Simulate and plot barplot

```
# Summarize the results by histograms and numbers
# Repeat the lines below many times for different numbers of Nsim and k
# What happens when Nsim gets large?
k <- 5
Nsim <- 1000
nheadsim <- simheads(Nsim=Nsim,k=k)
relfreq <- table(nheadsim)/Nsim # Calculate relative frequency for each outcome
barplot(relfreq,ylab="Relative frequency")
```



Exact probability We introduce the random variable X = number of heads. The exact probability $P(X = 3)$ is calculated using the binomial distribution with parameters $n = 5$ and $p = 0.5$.

In R, this is done so:

```
x <- dbinom(3, size=5, prob=0.5)
print(x)
```

```
## [1] 0.3125
```

Which tells us that the probability of getting 3 heads out of 5 coin tosses is 31.25 %.

3. Redo point 2 for a biased coin for which the probability of heads is 0.6.

Biased function

```
# Same function as above but with a biased coin. b is bias variable
simbias <- function(Nsim,k,b){
```

```

nheads <- vector(length=Nsim) # Define a vector to store the number of heads
for(i in 1:Nsim)               # Generate the number of heads Nsim times
  nheads[i] <- sum(sample(0:1,size=k,replace=TRUE, prob=c(1-b, b)))
return(nheads)                 # Return the resulting vector
}

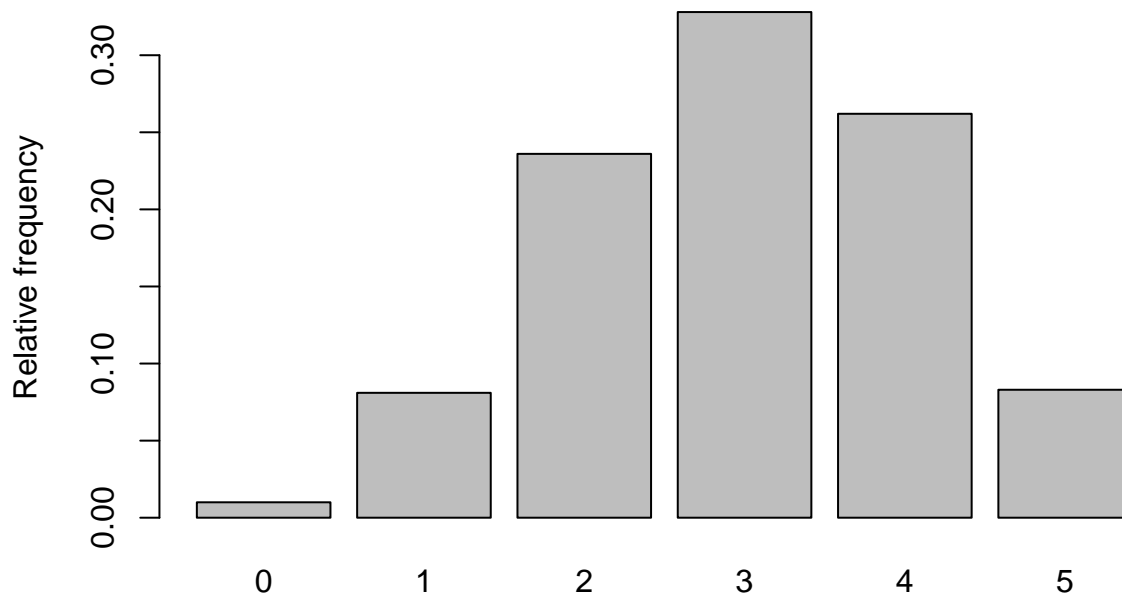
```

Simulate and summarise

```

# Summarize the results by histograms and numbers
# Repeat the lines below many times for different numbers of Nsim and k
# What happens when Nsim gets large?
k <- 5
Nsim <- 1000
b <- 0.6
nheadsim <- simbias(Nsim=Nsim,k=k,b=b)
relfreq <- table(nheadsim)/Nsim # Calculate relative frequency for each outcome
barplot(relfreq,ylab="Relative frequency")

```



Problem 2

For each of the following, determine the constant c such that $f(x)$ satisfies the condition of being the probability mass function (pmf) for a discrete random variable X , and then depict each $f(x)$ as a bar graph:

a)

Given the function

$$f(x) = x/c, \quad x = 1, 2, 3, 4 \quad (2)$$

for the function qualify as a pmf the constant c should have a value such that

$$\sum_{i=1}^4 f(x_i) = 1 \quad (3)$$

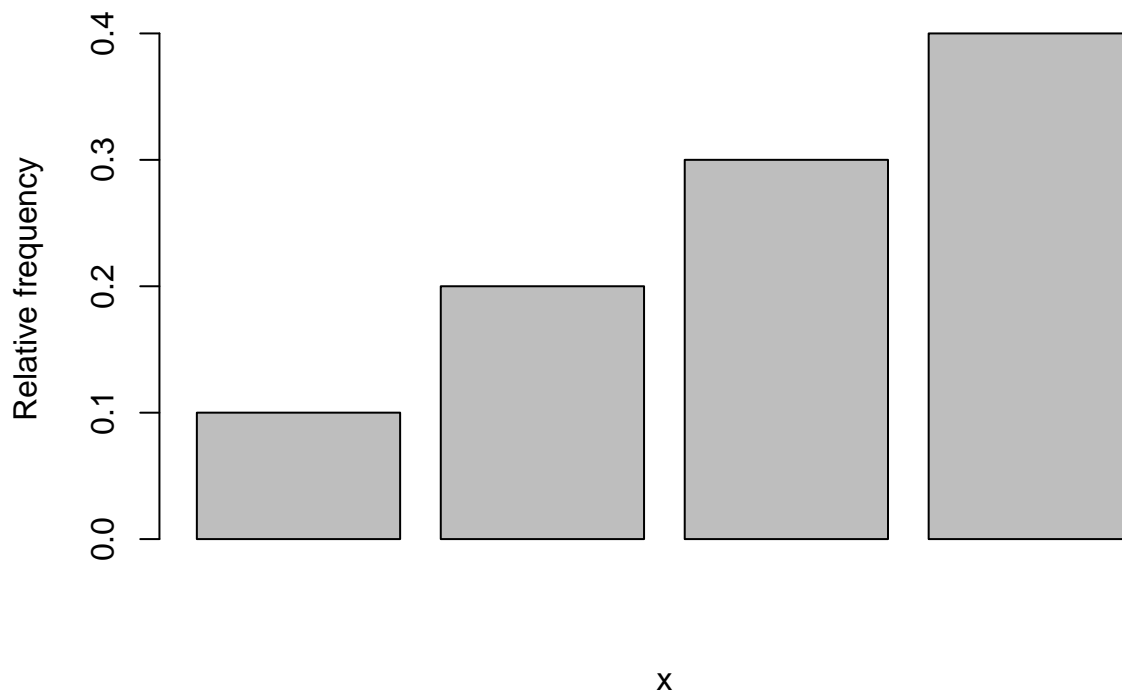
Therefore, c can be calculated as follows

$$1/c + 2/c + 3/c + 4/c = 1 \implies c = 10 \quad (4)$$

The function i R

```
# Define f1(x)
# The function takes in a vector and checks if it is valid
# function calculates pmf
f1 <- function(x) {
  ifelse(x %in% c(1, 2, 3, 4),
        x/10,
        0)
}

barplot(f1(1:4),ylab="Relative frequency", xlab="x")
```



we see that the function is a valid pmf since:

```
sum(f1(-100:100))
```

```
## [1] 1
```

b)

Given the function

$$f(x) = c(x+1)^2, \quad x = 0, 1, 2, 3 \quad (5)$$

for the function qualify as a pmf the constant c should have a value such that

$$\sum_{i=0}^3 f(x_i) = 1 \quad (6)$$

Therefore, c can be calculated as follows

$$c + 4c + 9c = 1 \implies c = 1/14 \quad (7)$$

The function in R

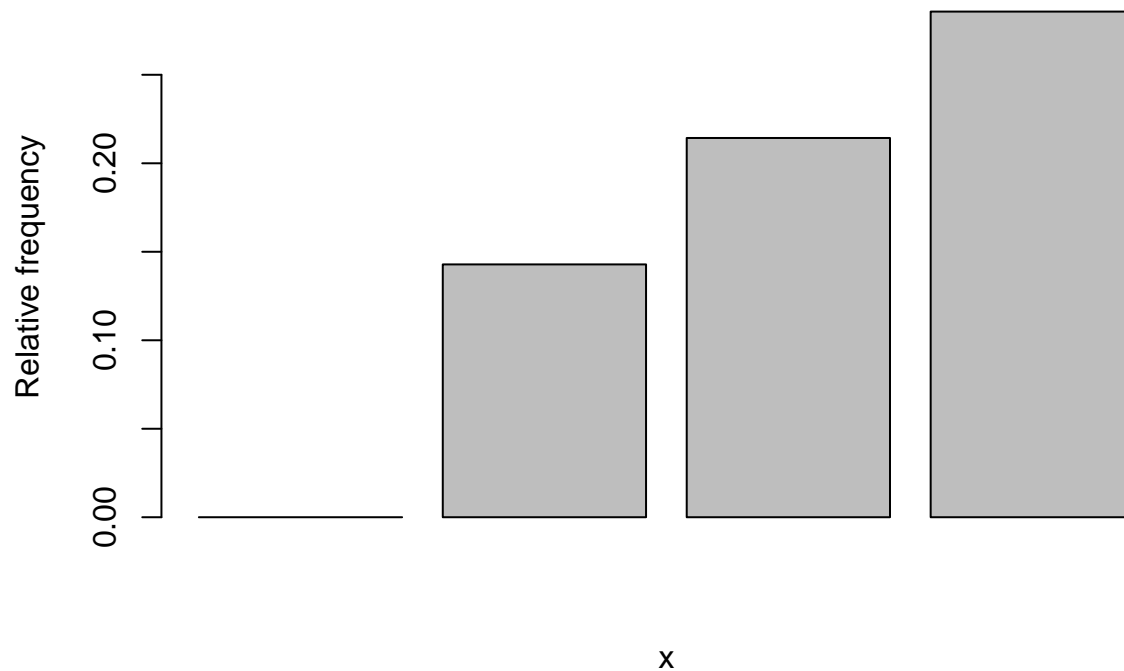
```
# Define f2(x)  
# The function takes in a vector and checks if it is valid
```

```

# function calculates pmf
f2 <- function(x) {
  ifelse(x %in% c(1, 2, 3, 4),
        (1/14)*(x + 1),
        0)
}

barplot(f2(0:3),ylab="Relative frequency", xlab="x")

```



we see that the function is a valid pmf since:

```
sum(f2(-100:100))
```

```
## [1] 1
```

Problem 3

For the probability distributions in problem 2 find $E(X)$ and $E(g(X))$ with $g(x) = x^3$.

Function for expected value

```

expected <- function(func, vec)
  return(sum(vec * func(vec)))

```

Expected value of f1 and f2

```
expected(f1, c(1,2,3,4))
```

```
## [1] 3
```

```
expected(f2, c(1,2,3,4))
```

```
## [1] 2.857143
```

Function for expected value with $g(x)$

```
gexpected <- function(func, vec, gvec)
  return(sum(gvec * func(vec)))
```

Expected value $E(g(x))$

```
gexpected(f1, c(1,2,3,4), c(1,2,3,4)^3)
```

```
## [1] 35.4
```

```
gexpected(f2, c(1,2,3,4), c(1,2,3,4)^3)
```

```
## [1] 32.42857
```

Problem 4

a)

Calculate $P(X < 0.5)$

$P(X < 0.5)$ is calculated using the integral

$$\int_{-\infty}^{0.5} f(x) dx \quad (8)$$

where $f(x)$ is the probability density function for X

Calculate $P(X < 0.5)$ in R

```
f3 <- function(x) {
  if (x >= 0 && x <= 1) {
    return(4*x*(1 - x^2))
  } else {
    return(0)
  }
}
```

```
integrate(Vectorize(f3),
  lower=-Inf,
  upper=0.5)$value
```

```
## [1] 0.4375
```

Calculate $E(X)$

```
integrate(Vectorize(function(x) x * f3(x)),
  lower = -Inf,
  upper = Inf)$value
```

```
## [1] 0.5333333
```

Calculate $\text{Var}(X)$

```
integrate(Vectorize(function(x) x^2 * f3(x)),
  lower = -Inf,
  upper = Inf)$value
```

```
## [1] 0.3333333
```

Calculate $\text{SD}(X)$

```
sqrt(integrate(Vectorize(function(x) x^2 * f3(x)),
  lower = -Inf,
  upper = Inf)$value)
```

```
## [1] 0.5773503
```

b) Calculate $P(X < 0.7)$ and $P(X > 7)$

$P(X < 0.7)$

```
integrate(Vectorize(f3),
  lower=-Inf,
  upper=0.7)$value
```

```
## [1] 0.7399155
```

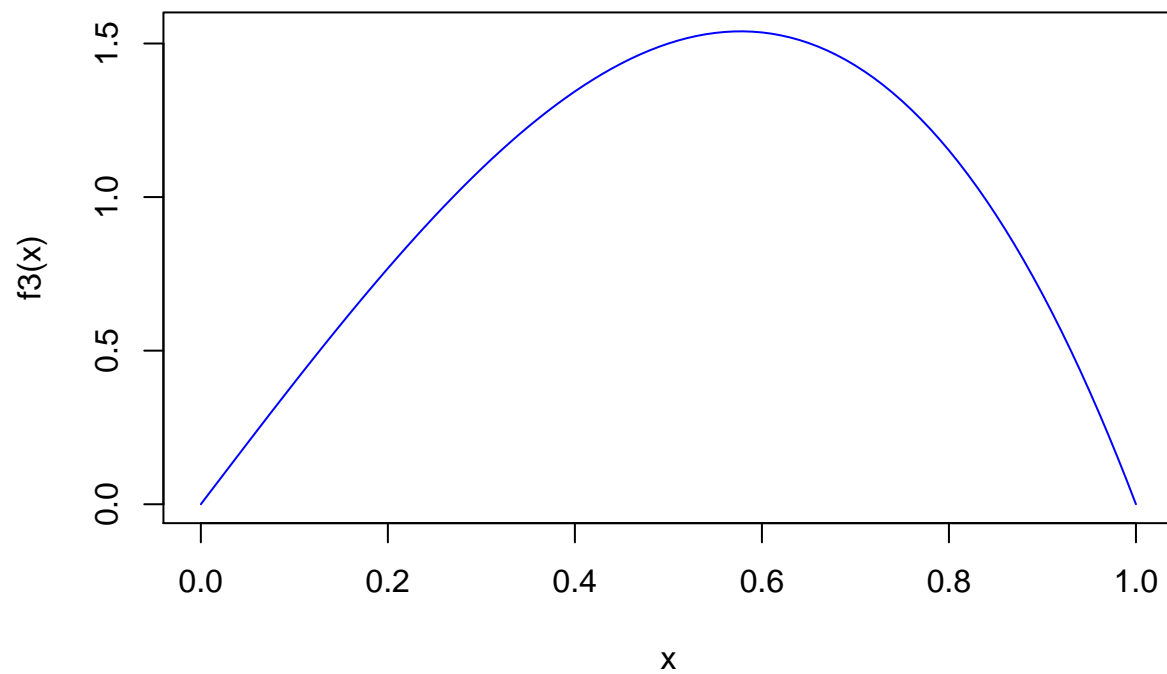
$P(X < 0.7)$

```
1 - integrate(Vectorize(f3),
  lower=-Inf,
  upper=0.7)$value
```

```
## [1] 0.2600845
```

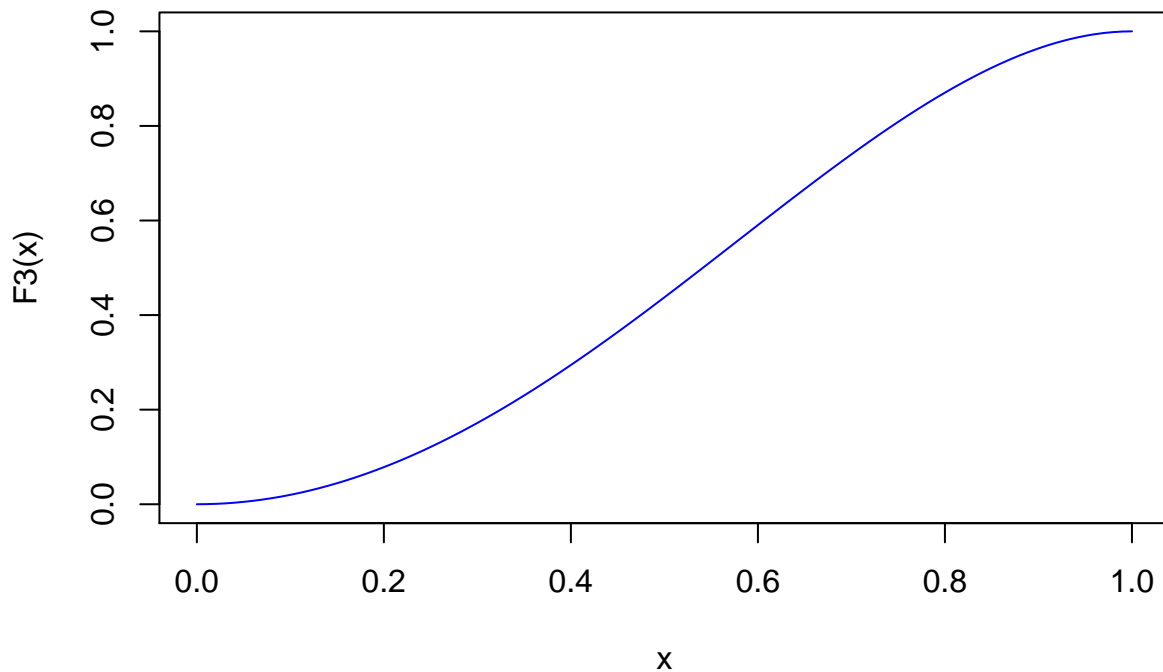
c) Plot $f(x)$

```
x <- seq(0, 1, 0.01)
plot(x, f3(x), type="l", col= "blue")
```

c) Plot $F(x)$

```
F3 <- function(x) {  
  return(-4*((x^4 / 4)-(x^2 / 2)))  
}  
  
x <- seq(0, 1, 0.01)  
plot(x, F3(x), type="l", col= "blue")
```



Problem 5

Let X be a continuous random variable with pdf

$$f(x) = \begin{cases} k(1 - x^2) & , -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where k is a constant.

a. Determine k and make a sketch of $f(x)$ in \mathbb{R} .

to determine k , we use the fact that $f(x)$ is a pdf which implies that the following integral

$$\int_{-\infty}^{\infty} f(x) dx = 1 \quad (10)$$

since the pdf is 0 for all x besides $x \in [-1, 1]$ we calculate

$$k \int_{-1}^1 (1 - x^2) dx = k \left[1 - \frac{1}{3} + 1 - \frac{1}{3} \right] = \frac{4}{3}k = 1 \implies k = \frac{3}{4} \quad (11)$$

This can also be validated in R

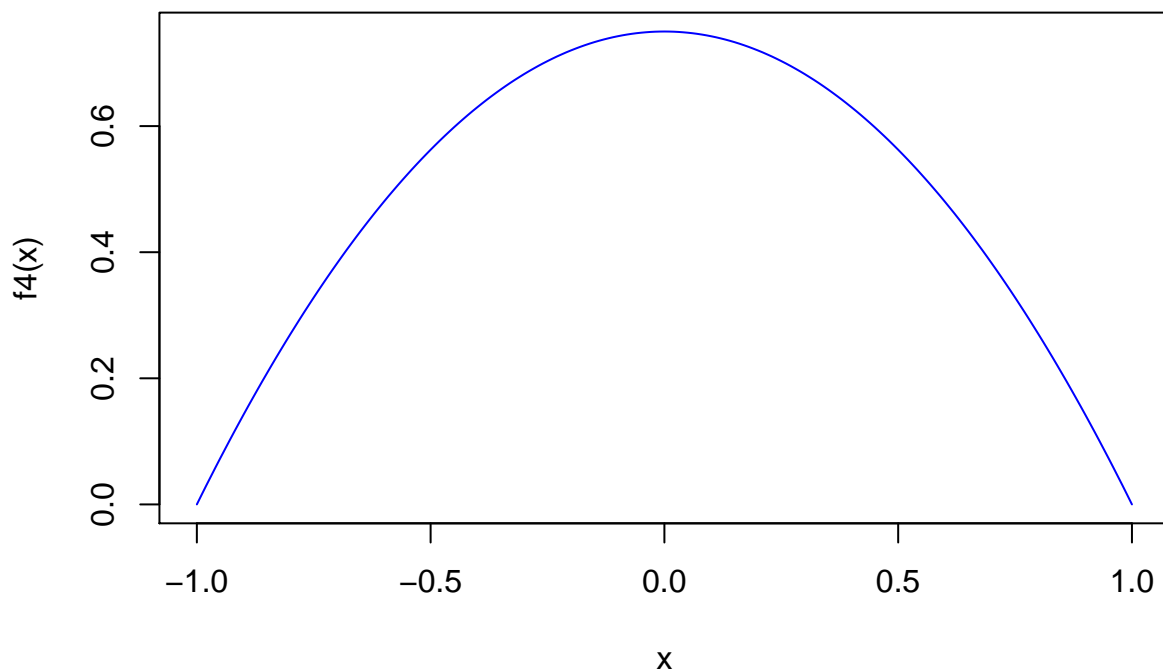
```
f4 <- function(x) {  
  return(3/4 * (1 - x^2))  
}
```

```
integrate(f4,
  lower = -1,
  upper = 1)$value
```

```
## [1] 1
```

Plot $f(x)$

```
x <- seq(-1, 1, 0.01)
plot(x, f4(x), type="l", col= "blue")
```



b. Calculate $P(X < 0.5)$

```
integrate(f4,
  lower = -1,
  upper = 0.5)$value
```

```
## [1] 0.84375
```

c. Let $Y = 1 + x^2$ Calculate $E(Y)$

We know from Theorem 4.5 in Walpole, Ronald E., et al.

$$E(aX + b) = aE(X) + b \quad (12)$$

This gives us that

$$E(Y) = E(aX^2 + b) = aE(X^2) + b \quad (13)$$

```
# Function for linear combination of random variable
expectedLinearComb <- function(g, f, lower, upper, a, b) {
  return (a*integrate(Vectorize(function(x) g(x) * f(x)),
    lower = lower,
    upper = upper)$value + b)
}

g <- function(x) x^2
expectedLinearComb(g, f4, -1, 1, 1, 1)

## [1] 1.2
```

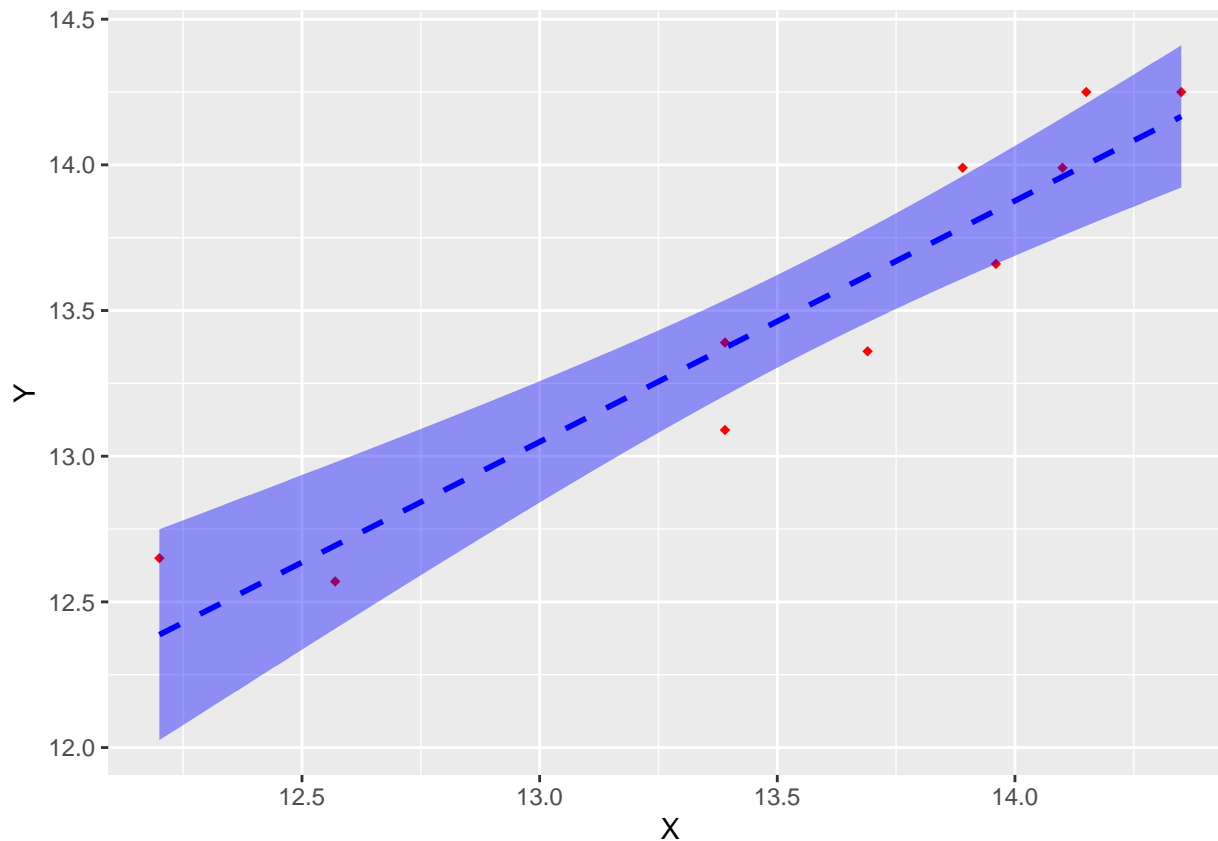
Problem 6

```
xi = c(13.89, 13.39, 12.20, 14.35, 14.10, 13.39, 13.96, 14.15, 13.69, 12.57)
yi = c(13.99, 13.39, 12.65, 14.25, 13.99, 13.09, 13.66, 14.25, 13.36, 12.57)

df <- data.frame("X" = xi, "Y" = yi)

ggplot(df, aes(x=X, y=Y)) +
  geom_point(shape=18, color="red")+
  geom_smooth(method=lm, linetype="dashed",
    color="blue", fill="blue")

## `geom_smooth()` using formula 'y ~ x'
```



Calculate correlation

```
cor(xi, yi)
```

```
## [1] 0.9431302
```

Problem 7

10 people are trying to identify one pepsi out of 5 total sodas. Given the null-hypothesis that pepsi and cola taste the same, there is a $\frac{1}{5}$ probability of choosing the correct one.

The 10 experiments across 10 people are independent and can be seen as repeated trials. The variable X is therefore a binomial distribution with parameters $p = \frac{1}{5}$ and $n = 10$

We do the calculations in R

R code calculating $P(X \leq 3)$

```
# Calculate P(X >= 3)
px = pbinom(3, size=10, prob=0.2, lower.tail = FALSE)
print(px)
```

```
## [1] 0.1208739
```

R code calculating $E(X)$

```
# Calculate E(X)
f5 = function(x) dbinom(x, size=10, prob=0.2)
mu = sum(0:10*f5(0:10))
mu

## [1] 2
```

R code calculating $Var(X)$

```
# Calculate Var(X)
sum((0:10)**2 * f5(0:10)) - mu**2

## [1] 1.6
```

- b. Let us now imagine that the test goes on and on (with more than 10 persons if necessary) and let N be the number of persons needed for the first correct identification. What kind of distribution does N have and what is $E(N)$?

This example follows the properties for a geometric distribution.

Geometric distribution has a infinite number of elements, since there is always a possibility, though an unlikely one, that we need N trials until success where N approaches infinity.

Problem 8

The height of Norwegian men in their twenties follows a normal distribution with mean 180 cm and standard deviation 6.5 cm.

Use the R function `pnorm` to calculate the following probabilities:

- a. The probability that a man has a height of less than 175.

```
# Calculate P(X < 175cm)
pnorm(175, mean = 180, sd = 6.5, lower.tail = TRUE, log.p = FALSE)

## [1] 0.2208782
```

- b. The probability that a man has a height of more than 190.

```
# Calculate P(X > 190cm)
pnorm(190, mean = 180, sd = 6.5, lower.tail = FALSE, log.p = FALSE)

## [1] 0.0619679
```

- c. The probability that a man has a height between 170 to 180.

```
# Calculate P(170cm < X < 180cm)
pnorm(180, mean = 180, sd = 6.5, lower.tail = TRUE, log.p = FALSE) -
pnorm(170, mean = 180, sd = 6.5, lower.tail = TRUE, log.p = FALSE)

## [1] 0.4380321
```

- d. The height which is such that only 1% of the men have a higher height. Problem

```
qnorm(0.99, mean = 180, sd = 6.5, lower.tail = TRUE, log.p = FALSE)

## [1] 195.1213
```

Problem 9

```
# All random variables has same expected value
```

```
ex = 0.4  
ey = 0.4  
ez = 0.4  
vx = 0.1  
vy = 0.1  
vz = 0.1  
corxy = 0  
corxz = 0.5  
coryz = -0.5
```

```
# 2X
```

```
e2x = ex+ex  
e2x
```

```
## [1] 0.8
```

```
v2x = vx + vx  
v2x
```

```
## [1] 0.2
```

```
# X + Y
```

```
exy = 2*ex  
exy
```

```
## [1] 0.8
```

```
vxy = vx + vy + 2*(corxy/(sqrt(vx)*sqrt(vy)))  
vxy
```

```
## [1] 0.2
```

```
# X + Z
```

```
exz = ex + ez  
exz
```

```
## [1] 0.8
```

```
vxz = vx + vx + 2*(corxz/(sqrt(vx)*sqrt(vz)))  
vxz
```

```
## [1] 10.2
```

```
# Y + Z
```

```
eyz = ey + ez  
eyz
```

```
## [1] 0.8
```

```
vyz = vy + vx + 2*(coryz/(sqrt(vy)*sqrt(vz)))  
vyz
```

```
## [1] -9.8
```