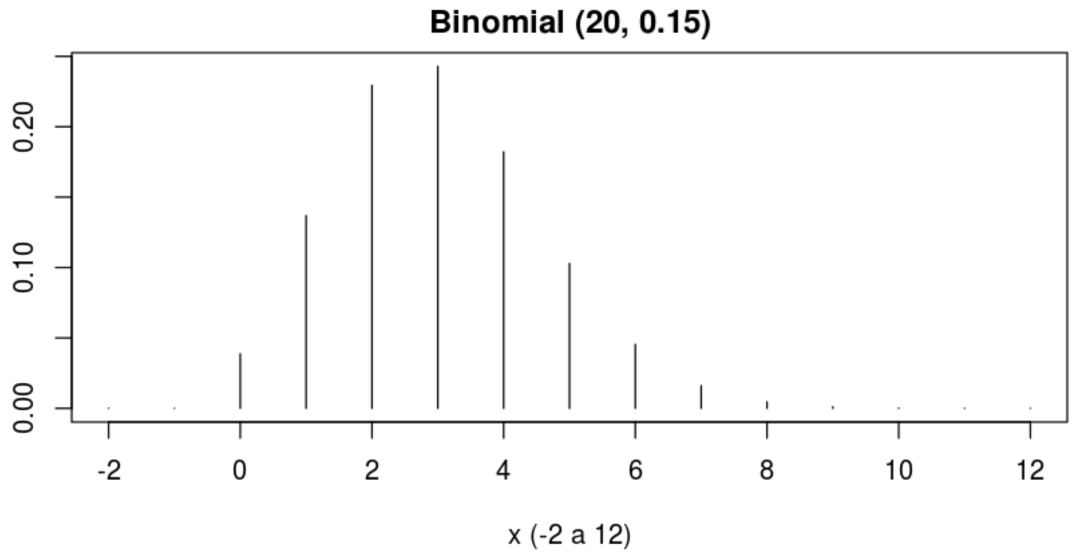
Fundamentos Est. para Ciencia dos Dados - Lista 4

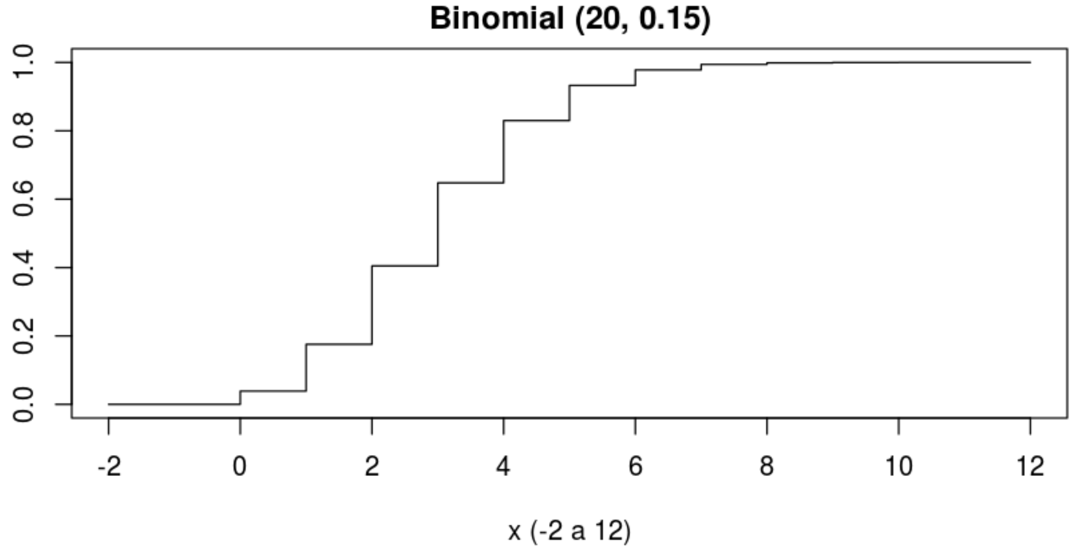
March 18, 2018

Arthur Iperoyg - 2018662796

1 Questão 1

```
In [189]: par(mar=c(4,2,2,2),mfrow=c(2,1))
          x <- seq(-2,12)
          pX <- dbinom(x, 20, 0.15)
          plot(x,pX, type = "h", main="Binomial (20, 0.15)",
               ylab="Distribuição", xlab="x (-2 a 12)")
          FX <- pbinom(x, 20, 0.15)
          plot(x,FX, type = "s", main="Binomial (20, 0.15)",
               ylab="Acumulada", xlab="x (-2 a 12)")
```



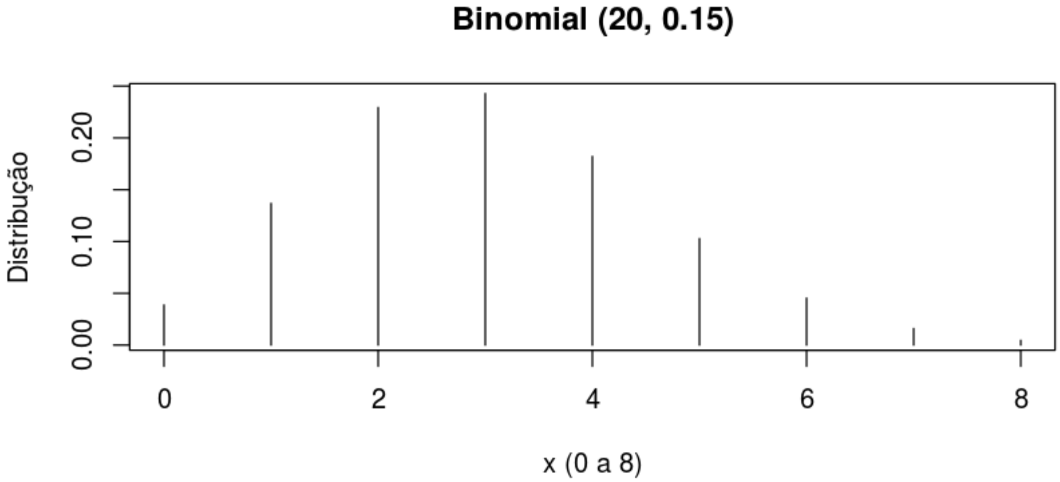


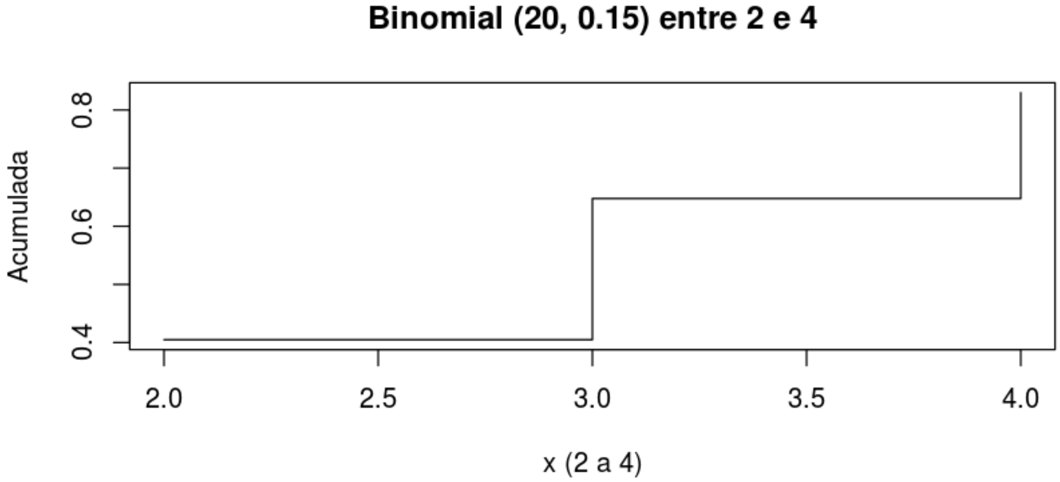
```
In [190]: k = x[match(max(pX),pX)] #P(X = k) máximo é quando k = 3
          maxP = max(pX) # a probabilidade máxima é 24.3% aproximadamente
          paste("Valor máximo quando k = ", k, " e a probabilidade é ", maxP)
```

'Valor máximo quando k = 3 e a probabilidade é 0.242828896149268'

```
In [191]: x2 <- c(0:8)
          pX2 <- dbinom(x2, 20, 0.15)
          par(mfrow=(c(2,1)))
          plot(x2,pX2, type = "h", main="Binomial (20, 0.15)",
               ylab="Distribução", xlab="x (0 a 8)")
          eX = 20 * 0.15
          li = eX-1
```

```
ls = eX+1
FX <- pbinom(li:ls, 20, 0.15)
plot(li:ls,FX, type = "s", main="Binomial (20, 0.15) entre 2 e 4"
     , ylab="Acumulada", xlab="x (2 a 4)")
#O valor acumulado no entorno de E(X) é bem alto (perto de 65%)
```





```
In [192]: eX = 20 * 0.15
          li = eX-1
          ls = eX+1
          calcSum <- sum(dbinom(li:ls, 20, 0.15))
          calcSub <- pbinom(ls,20, 0.15) - pbinom(li-0.01, 20, 0.15)
          # Eh necessario subtrair 0.01 porque senao o valor do limite inferior
          # será subtraido do resultado, o que nao eh o que queremos.
```

```
        paste("Somando as distribuições: ", calcSum,
              ". Subtraindo os acumulados dos extremos: ", calcSub)
```

```
'Somando as distribuições: 0.654288970179971 . Subtraindo os acumulados dos extremos: 0.654288970179971'
```

```
In [193]: paste("Quartil ~ 0.95 = ", qbinom(0.95, 20, 0.15))
```

```
'Quartil ~ 0.95 = 6'
```

```
In [194]: paste("Probabilidade do quartil ~ 0.95: ", pbinom(6,20,0.15))
```

```
'Probabilidade do quartil ~ 0.95: 0.978064899143155'
```

```
In [195]: size <- 1000
          v <- rbinom(size, 20, 0.15)
          eX = 20 * 0.15
          li = eX-1
          ls = eX+1
          # Sim, 63% dos valores cairam na faixa, o que era esperado.
          paste("% de valores que cairam na faixa delimitada: ",
                length(v[v >= li & v <= ls])/size * 100)
```

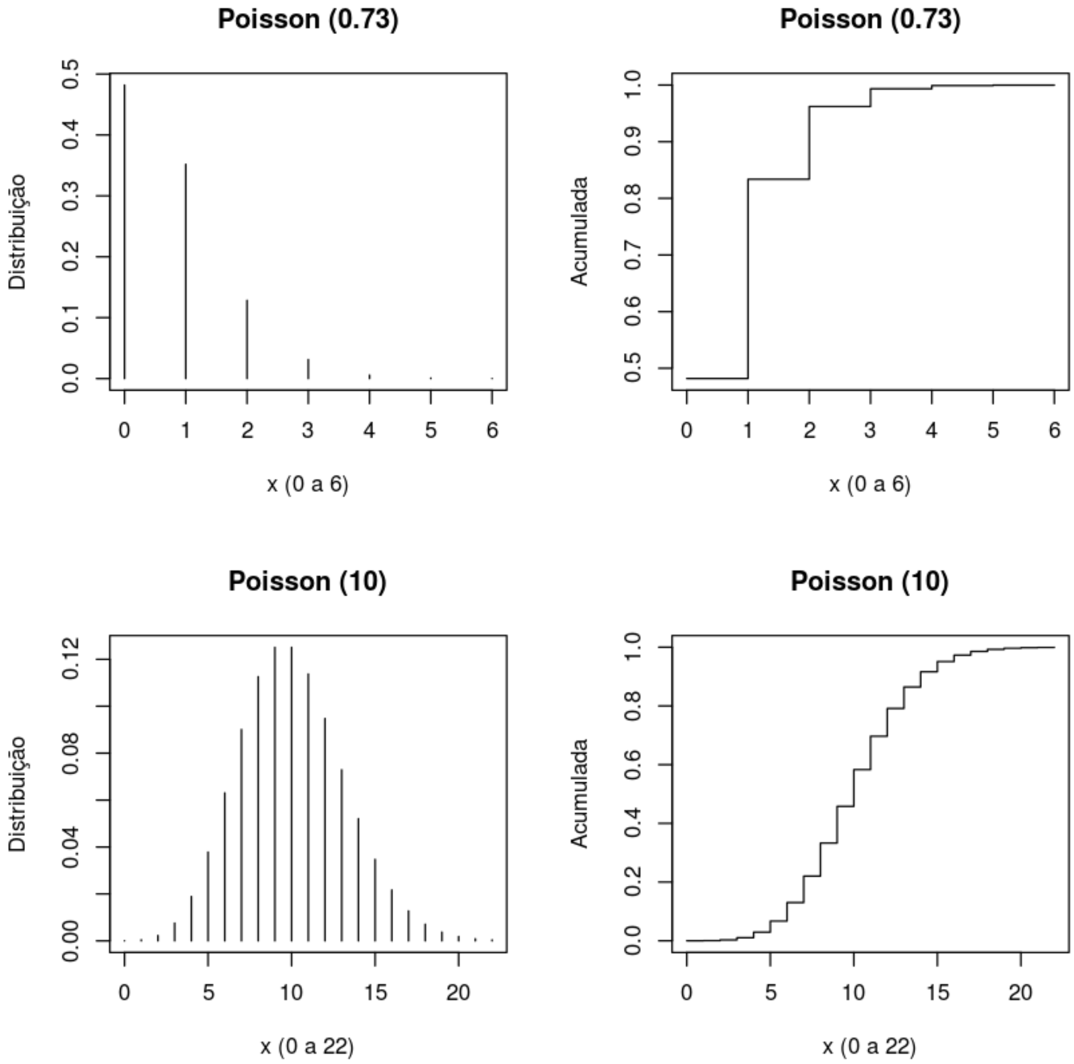
```
'% de valores que cairam na faixa delimitada: 65.1'
```

```
In [196]: size = 300
          v <- rbinom(size, 20, 0.15)
          paste("Bin(20, 0.15), X = 0, Distribuição: ",
                dbinom(0,20,0.15), "; Frequencia: ", length(v[v == 0])/size)
          paste("Bin(20, 0.15), X = 1, Distribuição: ",
                dbinom(1,20,0.15), "; Frequencia: ", length(v[v == 1])/size)
          paste("Bin(20, 0.15), X = 2, Distribuição: ",
                dbinom(2,20,0.15), "; Frequencia: ", length(v[v == 2])/size)
          paste("Bin(20, 0.15), X = 3, Distribuição: ",
                dbinom(3,20,0.15), "; Frequencia: ", length(v[v == 3])/size)
          paste("Bin(20, 0.15), X = 4, Distribuição: ",
                dbinom(4,20,0.15), "; Frequencia: ", length(v[v == 4])/size)
          paste("Bin(20, 0.15), X = 5, Distribuição: ",
                dbinom(5,20,0.15), "; Frequencia: ", length(v[v == 5])/size)
          paste("Bin(20, 0.15), X = 6, Distribuição: ",
                dbinom(6,20,0.15), "; Frequencia: ", length(v[v == 6])/size)
          # Sim, os valores sao parecidos.
```

```
'Bin(20, 0.15), X = 0, Distribuição: 0.0387595310845143 ; Frequencia: 0.0233333333333333'
'Bin(20, 0.15), X = 1, Distribuição: 0.136798345004168 ; Frequencia: 0.133333333333333'
'Bin(20, 0.15), X = 2, Distribuição: 0.229338401918753 ; Frequencia: 0.226666666666667'
'Bin(20, 0.15), X = 3, Distribuição: 0.242828896149268 ; Frequencia: 0.24'
'Bin(20, 0.15), X = 4, Distribuição: 0.182121672111951 ; Frequencia: 0.196666666666667'
'Bin(20, 0.15), X = 5, Distribuição: 0.102845179545572 ; Frequencia: 0.0966666666666667'
'Bin(20, 0.15), X = 6, Distribuição: 0.0453728733289289 ; Frequencia: 0.0533333333333333'
```

2 Questao 2

```
In [197]: x1 <- c(0:6)
          x2 <- c(0:22)
          eX1 = 0.73
          eX2 = 10
          pX1 <- dpois(x1, eX1)
          FX1 <- ppois(x1, eX1)
          pX2 <- dpois(x2, eX2)
          FX2 <- ppois(x2, eX2)
          par(mfrow=(c(2,2)))
          plot(x1,pX1, type = "h", main="Poisson (0.73)"
               , ylab="Distribuição", xlab="x (0 a 6)")
          plot(x1,FX1, type = "s", main="Poisson (0.73)"
               , ylab="Acumulada", xlab="x (0 a 6)")
          plot(x2,pX2, type = "h", main="Poisson (10)"
               , ylab="Distribuição", xlab="x (0 a 22)")
          plot(x2,FX2, type = "s", main="Poisson (10)"
               , ylab="Acumulada", xlab="x (0 a 22)")
```



```
In [198]: x1 <- c(-1:6)
          x2 <- c(0:21)
          eX1 = 0.73
          eX2 = 10
          pX1 <- dpois(x1, eX1)
          FX1 <- ppois(x1, eX1)
          pX2 <- dpois(x2, eX2)
          FX2 <- ppois(x2, eX2)
          paste("Poisson (0.73) E(X) = 0.73, k obtido: ",
                x1[match(max(pX1),pX1)])
          paste("Poisson (10) E(X) = 10, k obtido: ",
                x2[match(max(pX2),pX2)])
          #Sim, o valor k obtido é próximo ao valor esperado
```

```
        #(0 quando E(x) = 0.73 e 9 quando E(x) = 10)
```

```
'Poisson (0.73) E(X) = 0.73, k obtido: 0'
'Poisson (10) E(X) = 10, k obtido: 9'
```

```
In [199]: x1 <- c(0:3)
          x2 <- c(3:20)
          eX1 = 0.73
          eX2 = 10
          paste("Poisson (0.73) - 0 a 3 - soma: ",
                sum(dpois(x1, eX1)))
          paste("Poisson (10) - 3 a 20 - soma: ",
                sum(dpois(x2, eX2)))
```

```
'Poisson (0.73) - 0 a 3 - soma: 0.993352334865238'
'Poisson (10) - 3 a 20 - soma: 0.99564234362263'
```

```
In [200]: x1 <- c(0:3)
          x2 <- c(3:20)
          eX1 = 0.73
          eX2 = 10
          paste("Poisson (0.73) - 0 a 3 - soma pela dif: ",
                ppois(3,eX1)-ppois(0-0.01,eX1))
          paste("Poisson (10) - 3 a 20 - soma pela dif: ",
                ppois(20,eX2)-ppois(3-0.01,eX2))
```

```
'Poisson (0.73) - 0 a 3 - soma pela dif: 0.993352334865238'
'Poisson (10) - 3 a 20 - soma pela dif: 0.99564234362263'
```

```
In [201]: size = 1000
          eX1 = 0.73
          eX2 = 10
          x1_200poison = rpois(size, eX1)
          x2_200poison = rpois(size, eX2)
          #
          X_val = c(0,1,2,3,4,5,6)
          poi_0_73_d <- c(dpois(0,eX1))
          poi_0_73_f <- c(length(x1_200poison[x1_200poison == 0])/size)
          poi_10_d <- c(dpois(0,eX2))
          poi_10_f <- c(length(x2_200poison[x2_200poison == 0])/size)

          poi_0_73_d <- c(poi_0_73_d, dpois(1,eX1))
          poi_0_73_f <- c(poi_0_73_f, length(x1_200poison[x1_200poison == 1])/size)
          poi_10_d <- c(poi_10_d, dpois(1,eX2))
          poi_10_f <- c(poi_10_f, length(x2_200poison[x2_200poison == 1])/size)

          poi_0_73_d <- c(poi_0_73_d, dpois(2,eX1))
          poi_0_73_f <- c(poi_0_73_f, length(x1_200poison[x1_200poison == 2])/size)
          poi_10_d <- c(poi_10_d, dpois(2,eX2))
```

```
poi_10_f <- c(poi_10_f, length(x2_200poison[x2_200poison == 2])/size)

poi_0_73_d <- c(poi_0_73_d, dpois(3,eX1))
poi_0_73_f <- c(poi_0_73_f, length(x1_200poison[x1_200poison == 3])/size)
poi_10_d <- c(poi_10_d, dpois(3,eX2))
poi_10_f <- c(poi_10_f, length(x2_200poison[x2_200poison == 3])/size)

poi_0_73_d <- c(poi_0_73_d, dpois(4,eX1))
poi_0_73_f <- c(poi_0_73_f, length(x1_200poison[x1_200poison == 4])/size)
poi_10_d <- c(poi_10_d, dpois(4,eX2))
poi_10_f <- c(poi_10_f, length(x2_200poison[x2_200poison == 4])/size)

poi_0_73_d <- c(poi_0_73_d, dpois(5,eX1))
poi_0_73_f <- c(poi_0_73_f, length(x1_200poison[x1_200poison == 5])/size)
poi_10_d <- c(poi_10_d, dpois(5,eX2))
poi_10_f <- c(poi_10_f, length(x2_200poison[x2_200poison == 5])/size)

poi_0_73_d <- c(poi_0_73_d, dpois(6,eX1))
poi_0_73_f <- c(poi_0_73_f, length(x1_200poison[x1_200poison == 6])/size)
poi_10_d <- c(poi_10_d, dpois(6,eX2))
poi_10_f <- c(poi_10_f, length(x2_200poison[x2_200poison == 6])/size)


data.frame(X_val, poi_0_73_d, poi_0_73_f, poi_10_d, poi_10_f)
#Sim, os valores simulados sao proximos!
```

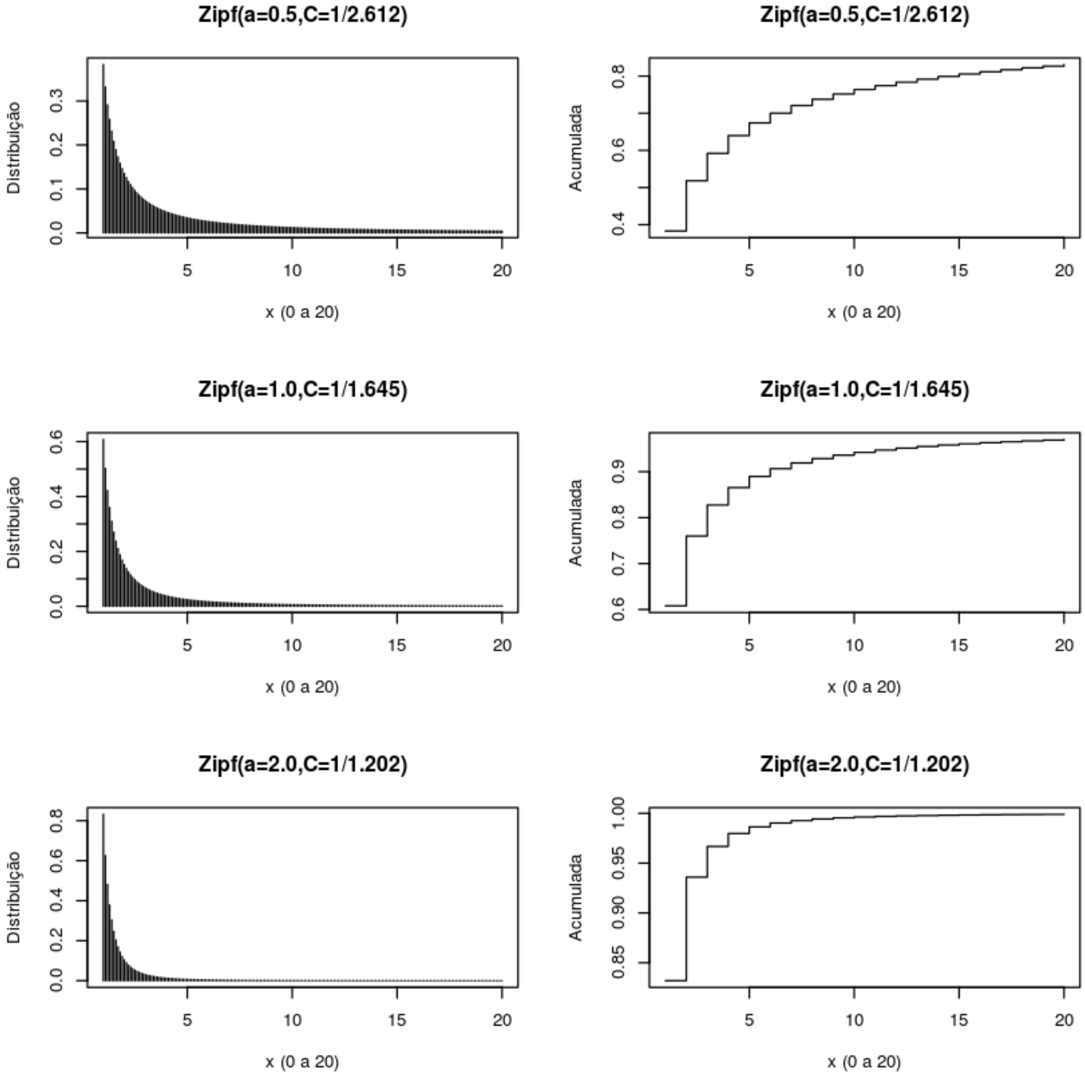
X_val	poi_0_73_d	poi_0_73_f	poi_10_d	poi_10_f
0	0.4819089901	0.504	4.539993e-05	0.000
1	0.3517935628	0.329	4.539993e-04	0.001
2	0.1284046504	0.131	2.269996e-03	0.003
3	0.0312451316	0.028	7.566655e-03	0.007
4	0.0057022365	0.007	1.891664e-02	0.017
5	0.0008325265	0.001	3.783327e-02	0.034
6	0.0001012907	0.000	6.305546e-02	0.064

3 Questao 3

```
In [202]: dzipf <- function(k, a, C){
               return(C / (k ^(a+1)))
          }

          pzipf <- function(k, a, C){
             return(cumsum(dzipf(k,a,C)))
          }
```

```
In [203]: a1 <- 1/2
          C1 <- 1/2.612
          a2 <- 1
```

```
C2 <- 1/1.645
a3 <- 2
C3 <- 1/1.202

range1 = seq(1,20,0.1)
range2 = seq(1,20,1)

par(mfrow=c(3,2))
plot(range1,dzipf(range1, a1, C1), type = "h",
    main="Zipf(a=0.5,C=1/2.612)" , ylab="Distribuição", xlab="x (0 a 20)")
plot(range2,pzipf(range2, a1, C1), type = "s",
    main="Zipf(a=0.5,C=1/2.612)" , ylab="Acumulada", xlab="x (0 a 20)")
plot(range1,dzipf(range1, a2, C2), type = "h",
    main="Zipf(a=1.0,C=1/1.645)" , ylab="Distribuição", xlab="x (0 a 20)")
plot(range2,pzipf(range2, a2, C2), type = "s",
    main="Zipf(a=1.0,C=1/1.645)" , ylab="Acumulada", xlab="x (0 a 20)")
plot(range1,dzipf(range1, a3, C3), type = "h",
    main="Zipf(a=2.0,C=1/1.202)" , ylab="Distribuição", xlab="x (0 a 20)")
plot(range2,pzipf(range2, a3, C3), type = "s",
    main="Zipf(a=2.0,C=1/1.202)" , ylab="Acumulada", xlab="x (0 a 20)")
```



```
In [204]: range1 = seq(0,10)
          range2 = range1+1

          ratio_zipf <- function(k, a){
              return ((k  / (k + 1 )) ^ (1+a))
          }

          paste("a1 = 1/2 - C1 = 1/2.612")
          a1_calc_ratio_10 <- dzipf(range2, a1,C1) / dzipf(range1, a1,C1)
          a1_form_ratio_10 <- ratio_zipf(range2,a1)
          a1_calc_ratio_500 <- dzipf(range2+500, a1,C1) / dzipf(range1+500, a1,C1)
          a1_form_ratio_500 <- ratio_zipf(range2+500,a1)
          data.frame(range1, a1_calc_ratio_10, a1_form_ratio_10,
           range1+500, a1_calc_ratio_500, a1_form_ratio_500)

paste("a1 = 1 - C1 = 1/1.645")
a2_calc_ratio_10 <- dzipf(range2, a2,C2) / dzipf(range1, a2,C2)
a2_form_ratio_10 <- ratio_zipf(range2,a2)
a2_calc_ratio_500 <- dzipf(range2+500, a2,C2) / dzipf(range1+500, a2,C2)
a2_form_ratio_500 <- ratio_zipf(range2+500,a2)
data.frame(range1, a2_calc_ratio_10, a2_form_ratio_10,
           range1+500, a2_calc_ratio_500, a2_form_ratio_500)

paste("a1 = 2 - C1 = 1/1.202")
a3_calc_ratio_10 <- dzipf(range2, a3,C3) / dzipf(range1, a3,C3)
a3_form_ratio_10 <- ratio_zipf(range2,a3)
a3_calc_ratio_500 <- dzipf(range2+500, a3,C3) / dzipf(range1+500, a3,C3)
a3_form_ratio_500 <- ratio_zipf(range2+500,a3)
data.frame(range1, a3_calc_ratio_10, a3_form_ratio_10,
           range1+500, a3_calc_ratio_500, a3_form_ratio_500)
```

'a1 = 1/2 - C1 = 1/2.612'

range1	a1_calc_ratio_10	a1_form_ratio_10	range1...500	a1_calc_ratio_500	a1_form_ratio_500
0	0.0000000	0.3535534	500	0.9970075	0.9970134
1	0.3535534	0.5443311	501	0.9970134	0.9970194
2	0.5443311	0.6495191	502	0.9970194	0.9970253
3	0.6495191	0.7155418	503	0.9970253	0.9970312
4	0.7155418	0.7607258	504	0.9970312	0.9970370
5	0.7607258	0.7935601	505	0.9970370	0.9970429
6	0.7935601	0.8184876	506	0.9970429	0.9970487
7	0.8184876	0.8380525	507	0.9970487	0.9970545
8	0.8380525	0.8538150	508	0.9970545	0.9970603
9	0.8538150	0.8667842	509	0.9970603	0.9970660
10	0.8667842	0.8776415	510	0.9970660	0.9970717

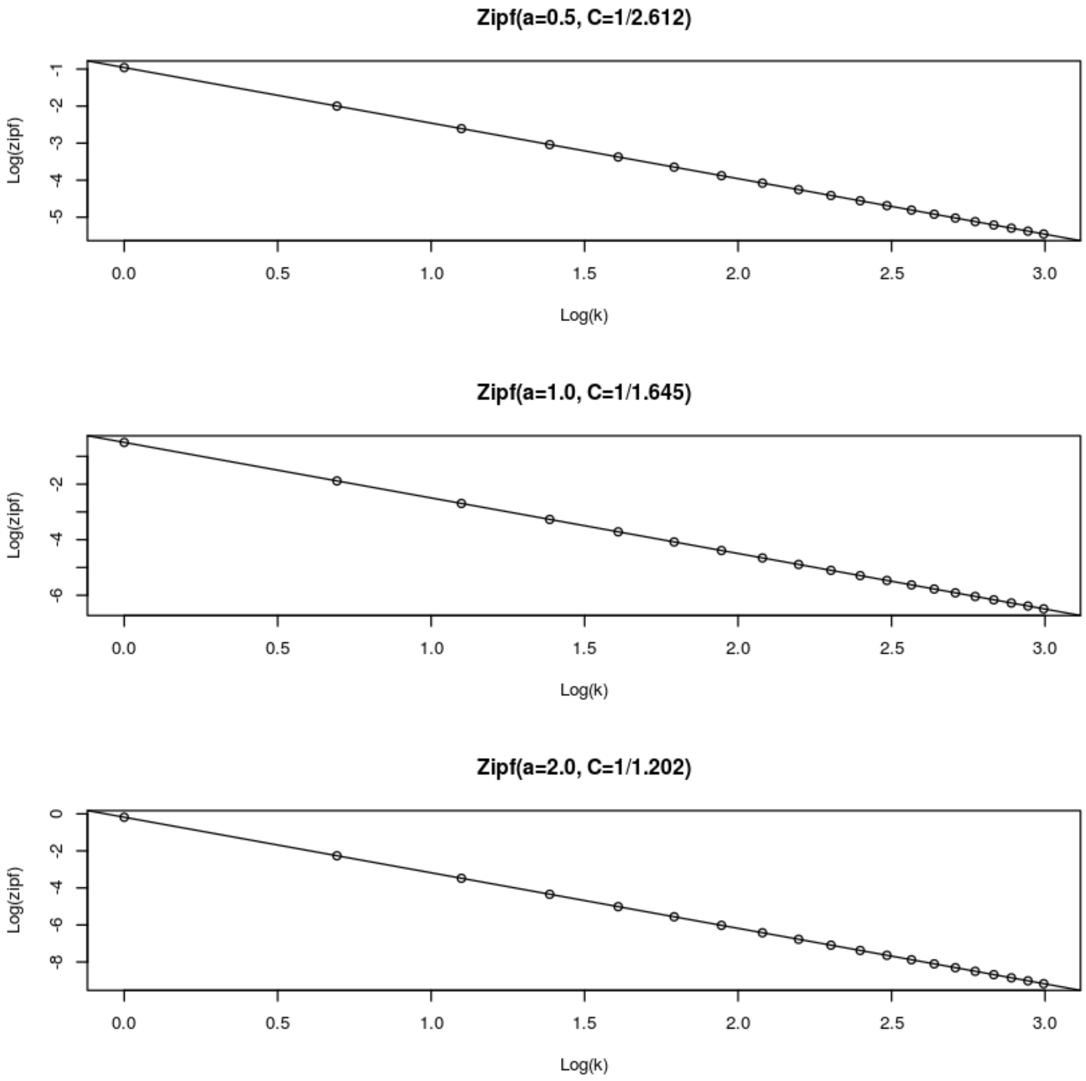
'a1 = 1 - C1 = 1/1.645'

range1	a2_calc_ratio_10	a2_form_ratio_10	range1...500	a2_calc_ratio_500	a2_form_ratio_500
0	0.0000000	0.2500000	500	0.9960120	0.9960199
1	0.2500000	0.4444444	501	0.9960199	0.9960278
2	0.4444444	0.5625000	502	0.9960278	0.9960357
3	0.5625000	0.6400000	503	0.9960357	0.9960435
4	0.6400000	0.6944444	504	0.9960435	0.9960513
5	0.6944444	0.7346939	505	0.9960513	0.9960591
6	0.7346939	0.7656250	506	0.9960591	0.9960669
7	0.7656250	0.7901235	507	0.9960669	0.9960746
8	0.7901235	0.8100000	508	0.9960746	0.9960823
9	0.8100000	0.8264463	509	0.9960823	0.9960899
10	0.8264463	0.8402778	510	0.9960899	0.9960976

'a1 = 2 - C1 = 1/1.202'

range1	a3_calc_ratio_10	a3_form_ratio_10	range1...500	a3_calc_ratio_500	a3_form_ratio_500
0	0.0000000	0.1250000	500	0.9940239	0.9940358
1	0.1250000	0.2962963	501	0.9940358	0.9940476
2	0.2962963	0.4218750	502	0.9940476	0.9940594
3	0.4218750	0.5120000	503	0.9940594	0.9940712
4	0.5120000	0.5787037	504	0.9940712	0.9940829
5	0.5787037	0.6297376	505	0.9940829	0.9940945
6	0.6297376	0.6699219	506	0.9940945	0.9941061
7	0.6699219	0.7023320	507	0.9941061	0.9941177
8	0.7023320	0.7290000	508	0.9941177	0.9941292
9	0.7290000	0.7513148	509	0.9941292	0.9941406
10	0.7513148	0.7702546	510	0.9941406	0.9941521

```
In [209]: range1 = seq(1,20)

          par(mfrow=c(3,1))
          plot(log(range1), log(dzipf(range1,a1,C1)),
              main="Zipf(a=0.5, C=1/2.612)", ylab="Log(zipf)", xlab="Log(k)")
          lines(abline(log(C1), -(1 + a1)))

          plot(log(range1), log(dzipf(range1,a2,C2)),
               main="Zipf(a=1.0, C=1/1.645)", ylab="Log(zipf)", xlab="Log(k)")
          lines(abline(log(C2), -(1 + a2)))

          plot(log(range1), log(dzipf(range1,a3,C3)),
               main="Zipf(a=2.0, C=1/1.202)", ylab="Log(zipf)", xlab="Log(k)")
          lines(abline(log(C3), -(1 + a3)))
```



```
In [206]: rzipf <- function(nsim = 1, alpha = 1, Cte = 1/1.645)
          {
              res = numeric(nsim)
              for(i in 1:nsim) {
                  x = -1
                  k = 1
                  F = p =  Cte
                  U = runif(1)
                  while( x == -1) {
                      if(U < F) {
                          x = k
                      }
                      else {
```

```
                p = p * (k/(k+1))^(1+alpha)
                F = F + p
                k = k+1
            }
        }
        res[i] = x
    }
    return (res)
}
```

In [212]:

```
a1 <- 1/2
C1 <- 1/2.612
a2 <- 1
C2 <- 1/1.645
a3 <- 2
C3 <- 1/1.202

#rzipf(400, a1, C1)
#rzipf(400, a2, C2)
#rzipf(400, a3, C3)

print(rzipf(400, a2, C2))
```

```
  [1]   1   3   1   1   1   1   1   1   2   4   1   1   5   1   2   1   3   3
 [19]   1   2   4   2   1   4   3   2   1  10   2   1   1   2   1   3  11   1
 [37]   2   1   1   1   1   1   1   1  17   2   2   4   1   2   2   2   1   1
 [55]   4   1   1   2   1   1   2   1   1   1   1   1   2   1   3   1   1   2
 [73]   1   6   1   1   3   2   1   7   1   2   1   1   1  14   1   1  16   1
 [91]  32   2   1   1   1   1  11   1   1   1   1   1   2   6   4   1   1   1
[109]   1   3   1   1   1   1   1   1   1   1   2   1   8   2   2   2  10   1
[127]   9   1   1   1   4   2   3   1   1   4   2 113   4   4   1   1   7   1
[145]   4   1  14   2   1   1   1   5   7   3   6   1   1   1   1   1   1   1
[163]   4   1   1   3   1  45   2   2  31   2   3   1   2   1  10  17   1   2
[181]   1   1   1   1   1   1   1   2   4   3   2   1   1   6   1   1   2   1
[199]   1   1   1   1   9   2   1   4   3   1   1   1   2   1   1   1   3   1
[217]   1   1   2   2   1   1   1   1   1   1   4   1   1   1   1   1   1   1
[235]   1   1   2   1   6   3   1   1   1   2   1   1   1   1   1   4   2   1
[253]   1   1   1   1   1   1   1   1   1   1   2   2   1   5   1   1   2   1
[271]   1   1   1   1   1   1   1   2   2   1   1   1   2   2   1   1   1   3
[289]   1   1   1   1   3   5   2   1   2   1   1   1   1   1   1   1   1   5
[307] 458   1   2   2   1   1   8   9 142   6   1   1   3   1   1   1   1 102
[325]   1   1   2   3   3   2   1   1   1   1   1   4   1   2   1   1   2   1
[343]   1   7   1   1   1   1   1   1   1   1   1   1  25   1   1   1   2   1
[361]   3   1   1   1   1   1   1   3   6   1   1   2   1   2   2   1   3   1
[379]   1   1   3   1   1   4   1   3  35   1   1   2   2   5   1   1   3   1
[397]   1   1   1   5
```