Exercise 8. Sampling distributions

Michal Béreš

Other selected continuous distributions

χ^2 - Chi-square distribution(Pearns distribution)

- Use: when estimating the standard deviation(using sampling)

- Has a single parameter - the number of degrees of freedom

- $\frac{S^2}{\sigma^2}(n-1) \sim \chi^2_{n-1}$
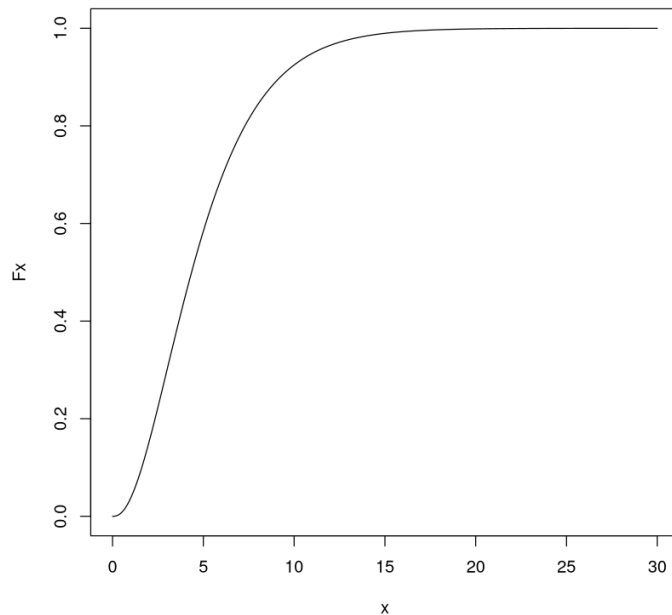
- S is the sample standard deviation

In [1]:
```
n =   5 # number of degrees of freedom
x = seq(from = 0, to = 30, by = 0.01) # x-axis
fx = dchisq(x = x, df = n) # chi-sq probability density. distribution

plot(x, fx, type='l')
```
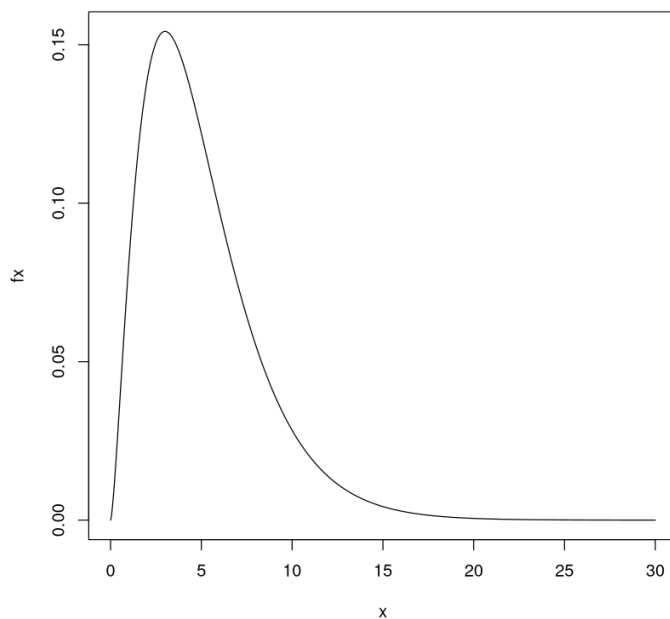

In [2]:
```
Fx = pchisq(q = x, df = n) # distrib. f. of chi-sq. distribution

plot(x, Fx, type='l')
```
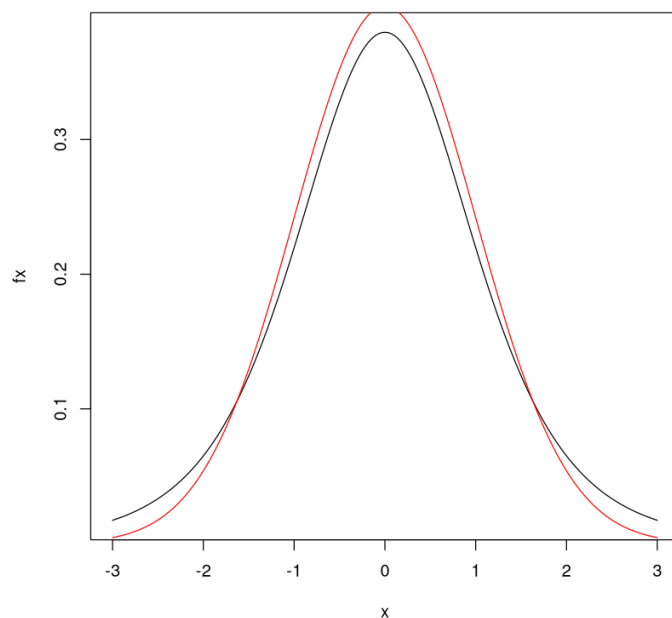
t - Student's distribution

- Application: when estimating the mean without exact knowledge of variance(sampling only)

- $\frac{\bar{X}-\mu}{S}\sqrt{n} \sim t_{n-1}$

- $\bar{X}$ is the sample average

- S is the sample standard deviation

- converges to a normalized normal distribution with increasing number of degrees of freedom

In [3]:
```
n =   5 # number of degrees of freedom
x = seq(from = -3, to = 3, by = 0.01) # x-axis
fx = dt(x = x, df = n) # probability density of the student's distribution

plot(x, fx, type='l')

fnorm = dnorm(x, mean=0, sd=1)   # norm values. normal dist
lines(x, fnorm, col="red")       # to the last graph
```
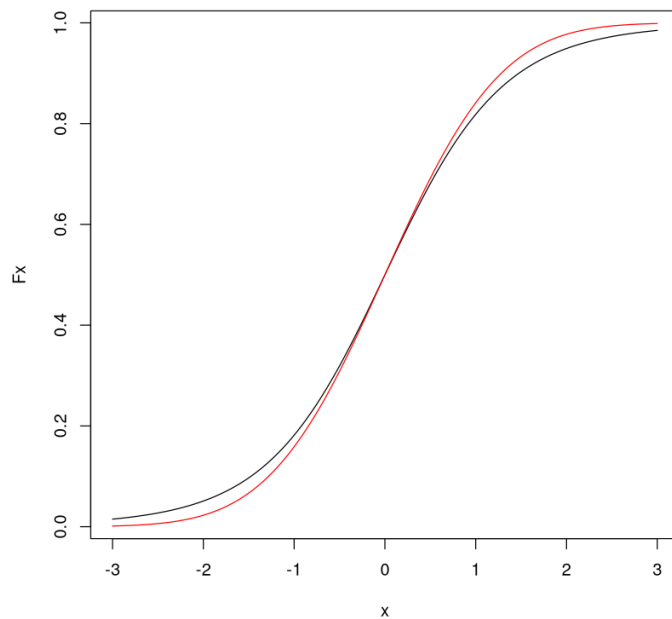
```
Fx = pt(q = x, df = n) # probability density of the student's distribution

plot(x, Fx, type='l')

Fnorm = pnorm(x, mean=0, sd=1)  # norm values. normal birth
lines(x, Fnorm, col="red")      # to the last graph
```


F - Fisher-Snedecor distribution
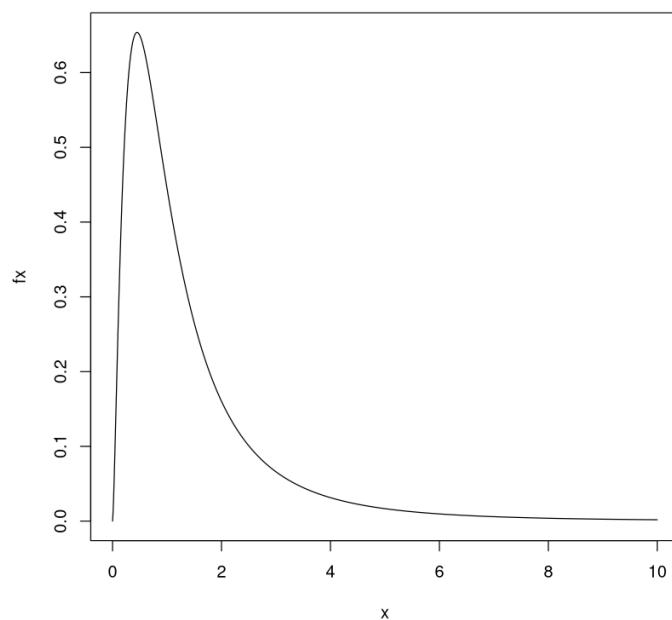
- Used to test ratio of variances

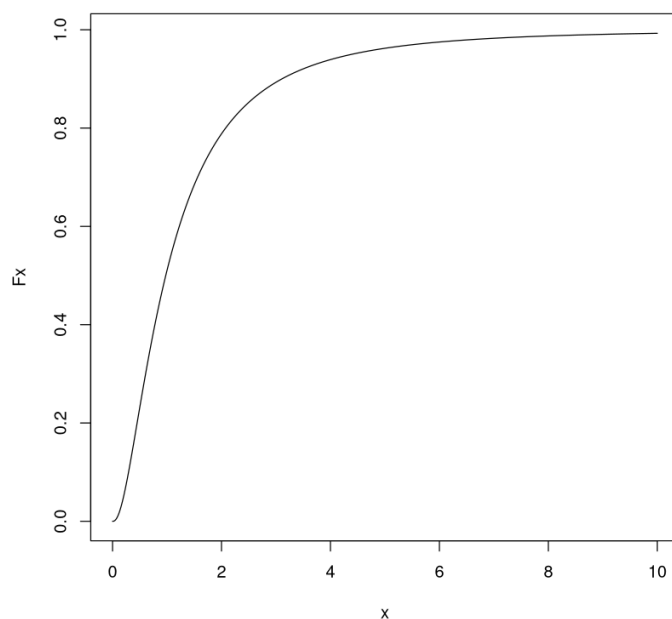- $\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$

```
n = 5   # number of degrees of freedom selection. 1
m = 6   # number of degrees of freedom selection. 2
x = seq(from = 0, to = 10, by = 0.01) # x-axis
fx = df(x = x, df1 = n, df2 = m) # chi-quad probability density. distribution

plot(x, fx, type='l')
```

```
Fx = pf(q = x, df1 = n, df2 = m) # chi-quad probability density. distribution

plot(x, Fx, type='l')
```


How does the average of the values from the normal distribution behave?

The function **rnorm(n, mean, sd)** generates **n** values from the normal distribution with themeanvalue **mean** and the standard deviation **sd**.

```
selection_size = 30
mu = 10
sigma = 3
random_selection = rnorm(n=selection_size, mean=mu, sd=sigma)
random_selection
```

11.3062265595301 · 10.809312184237 · 9.86643308059752 · 10.4910559756103 · 11.4521940495524 · 7.96251025569272 · 9.89800848482684 · 9.634293579656 · 16.2777724427846 · 10.1544097353766 · 12.0630410997955 · 8.77137878629639 · 5.71897635738176 · 10.1722782393151 · 6.49694491806961 · 4.7969033856175 · 9.68800049133389 · 13.2742851857312 · 4.80976309827595 · 7.70862309012849 · 14.4320500880304 · 12.132430525525 · 8.3266291701652 · 10.4439951878145 · 9.21068596255054 · 8.16114645393442 · 13.6877930358881 · 15.3427274521976 · 8.75682729929821 · 15.3915120630611

```
mean(random_selection)
sd(random_selection)
```

10.2412736079425

2.9777118337053

Random variable: average of values

```
n_selections = 1000
means = numeric(n_selections) # numeric produces vector 0
st_deviations = numeric(n_selections)
for(i in 1:n_selections){
  random_selection = rnorm(n=selection_size,mean=mu,sd=sigma)
  means[i] = mean(random_selection)
  st_deviations[i] = sd(random_selection)
}
```

In [10]:
```
hist(means)
```

Histogram of means
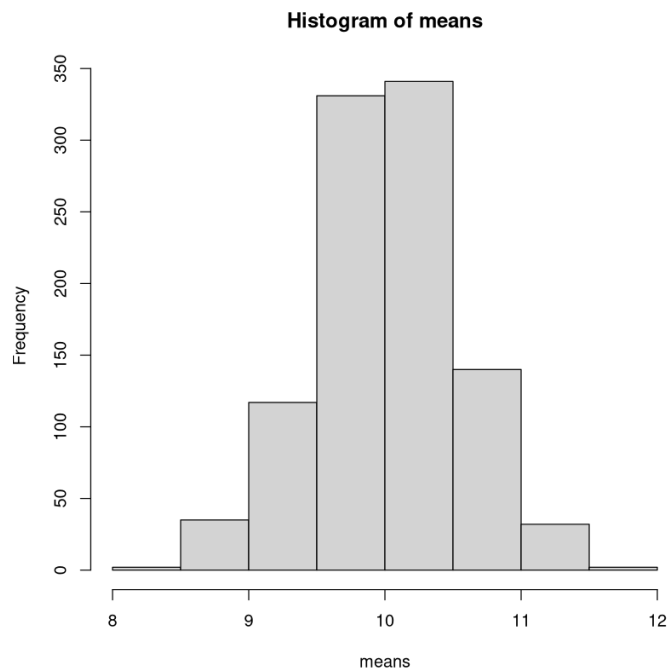
In [11]:
```
qqnorm(means)
qqline(means)
```

Normal Q-Q Plot
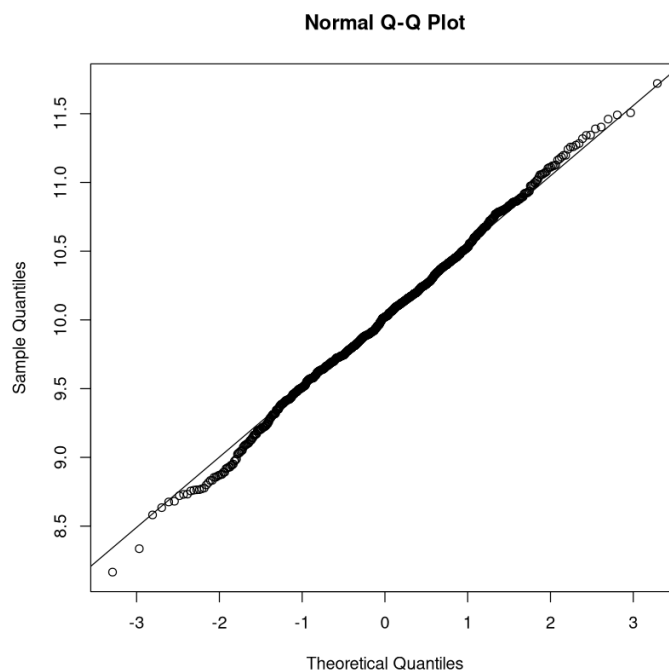
In [12]:
```
mean(means)
sd(means)
sigma/sqrt(selection_size)
```

10.0160254413041

0.53528150232229

0.547722557505166

# How does the average of values from the uniform distribution behave?

The function **runif(n, min, max)** generates **n** values from the uniform distribution U(**min, max**).

```
In [13]:  selection_size = 30
          a = 1
          b = 7
          random_selection=runif(n=selection_size, min=a, max=b)
          # random sample
```

```
In [14]:  hist(random_selection)
```

Histogram of random_selection
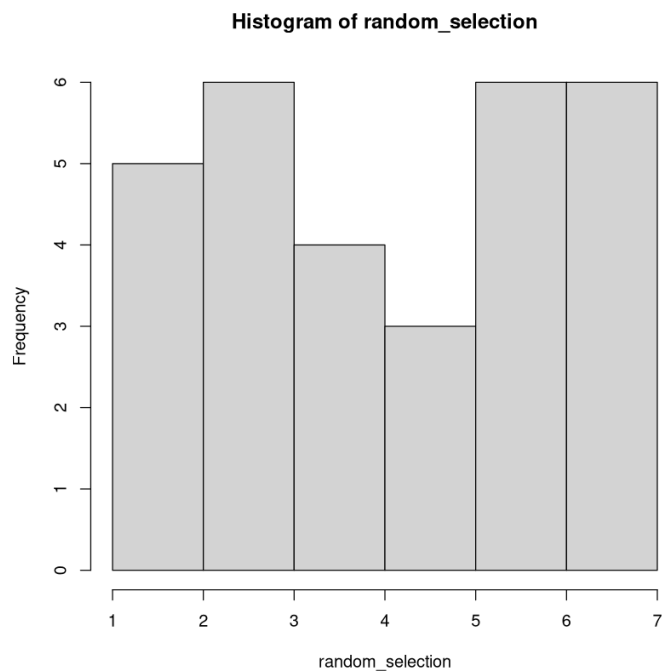

```
In [15]:  mu = (a+b)/2
          mu
          mean(random_selection)
          sigma = sqrt((b-a)^2/12)
          sigma
          sd(random_selection)
```

4

4.08716902770102

1.73205080756888

1.77611333114412

Random variable: average of values

```
In [16]:  n_selections = 1000
          means = numeric(n_selections)
          st_deviations = numeric(n_selections)
          for(i in 1:n_selections){
            random_selection = runif(n=selection_size, min=a, max=b)
            means[i] = mean(random_selection)
            st_deviations[i] = sd(random_selection)
          }
```
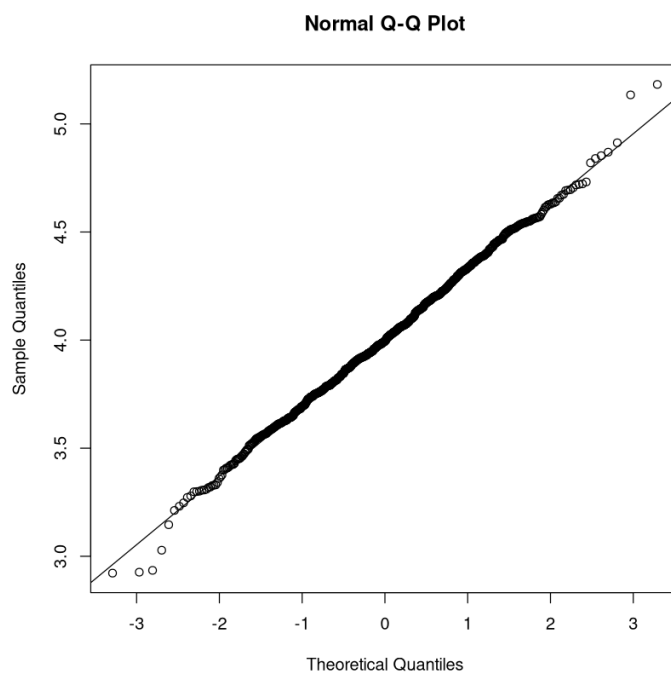
```
In [17]:  hist(means)
```

Histogram of means

```
qqnorm(means)
qqline(means)
```

Normal Q-Q Plot
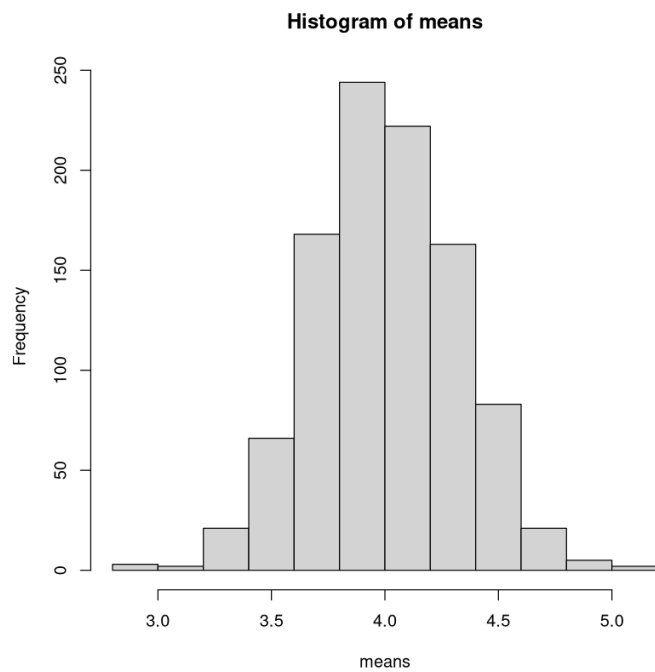
```
mean(means)
sd(means)
sigma/sqrt(selection_size)
```

4.00852905525048

0.318828420525004

0.316227766016838

Examples

Example 1.

The load on an aircraft with 64 seats shall not exceed 6,000 kg. What is the probability that this value will be exceeded at full occupancy if the passenger mass has a mean value of 90 kg and a standard deviation of 10 kg?

```
In [20]:   # X... weight 64 passengers
           # X~N(64 * 90; 64 * 100)
           # P(X>6000)=1 - F(6000)

           1 - pnorm(q=6000, mean=64*90, sd=sqrt(64*100))
```

0.0013498980316301

Example 2.

The consignment contains 300 products of a certain type. It is known that the probability of making a defective product of this type is 0.04.

a)

Estimate the probability that the proportion of defective products in the consignment (300 products) will differ from the probability of producing a defective product by less than 1% (i.e. $p - \pi \in< -0.01, 0.01 >$).

```
In [21]:   # p-pi from -0.01 to 0.01
           # for pi = 0.04, p is from 0.03 to 0.05
           # having 300 products, we should observe between 9 and 15 defective ones

           # X...number of defective products out of 300
           # X~Binom(n=300, pi=0.04)
           # P(9 <= X <= 15)=P(X<=15)-P(X<9)
           pbinom(15,300,0.04)-pbinom(9-1,300,0.04)
```

0.699166351176511

```
In [22]:   # X=(p - π)/sqrt(π*(1 - π)) * sqrt(n) ~ N(0, 1)
           # P(-0.01/sqrt(π*(1-π))*sqrt(n) < X < 0.01/sqrt(π*(1-π))*sqrt(n))
           pi = 0.04
           n = 300
           bound = 0.01/sqrt(pi*(1-pi))*sqrt(n)
           pnorm(q=bound, mean=0, sd=1) - pnorm(q=-bound, mean=0, sd=1)
```

0.623240882188418

b)

How will the result change if the shipment contains 3,000 products?

```
In [23]:   pbinom(150,3000,0.04)-pbinom(90-1,3000,0.04)
```

0.995471791017906

```
In [24]:   n = 3000
           bound = 0.01/sqrt(pi*(1-pi))*sqrt(n)
           pnorm(q=bound, mean=0, sd=1) - pnorm(q=-bound, mean=0, sd=1)
```

0.994811392447684

Example 3.

Passenger regularly travel to and from work using public transport. It is known that the waiting time for the arrival of public transport ranges from 0 to 3 minutes. What is the probability that the total waiting time for a passenger during in 23 working days will be less than 80 minutes?

```
In [25]:   # Y... time of the i-th waiting for public transport
           # y~R(0; 3)
           # X... total waiting time in 23 days(round trip ⇒ 46 waitings)
           # X~N(46 * EY; 46 * DY)
           # P(X<80)

           a = 0
           b = 3
           n = 46
```

```
EY = (a+b)/2
DY = (b-a)^2/12

pnorm(q=80, mean=n*EY, sd=sqrt(n*DY))
```

0.969449534780224

Example 4.

Assume that the average electricity consumption of households in a given city in January is 120 kWh and the standard deviation of consumption is 100 kWh. Determine the probability that the average consumption of 100 randomly selected households will be greater than 140 kWh.

In [26]:
```
# Xi... consumption of the i-th household
# X... average consumption of 100 households
# X~N(EXi; Dxi/n)
# P(X>140)

EXi = 120
DXi = 100^2
n = 100
1 - pnorm(q=140, mean=EXi, sd=sqrt(DXi/n))
```

0.0227501319481792

Example 5.

The Acme Battery Company has developed a new type of mobile phone battery. On average, batteries last 60 minutes on a single charge. The standard deviation of this time is 4 minutes. Assume that the production department runs a quality control test after 6 months. They performed two random selections with a range of 10 batteries and in both found a standard deviation of battery life greater than 6 minutes. How likely were they to expect such a result?

In [27]:
```
# X=(n - 1) * S^2/σ^2
# X ~ χ_n-1
# P(S>6)=P(X>...)

n = 10
S_obs = 6
sigma = 4
X_obs = (n - 1)*S_obs^2/sigma^2

P_one_observation = 1 - pchisq(q=X_obs, df=n-1)
P_one_observation
P_one_observation^2
```

0.0164307101306824
0.00026996823539851

Example 6.

The mortality tables show a probability of 0.99 that a 35-year-old man will live another year. The annual premium for this age group is CZK 2,000, in the event of death the insurance company will pay CZK 100,000. What is the probability that the profit of the company insuring 500 men aged 35 will be at least CZK 500,000?(Solve in two ways - using the binomial distribution and using the binomial distribution approximation by the normal distribution.)

In [28]:
```
# X... number of men out of 500 who won't live to see another year
# X~Bi(500; 0.01)
# Z=500 · 2,000 - X · 100,000
# P(Z ≥ 500,000)=P(X ≤ 5)

pbinom(5, size=500, prob=0.01)
```

0.615962131821453

In [29]:
```
# X~Bi(500; 0.01)~N(500 * 0.01; 500 * 0.01*(1-0.01))
# P(X ≤ 5)~P(X<5.5)(continuity correction)
```

```
pnorm(5.5, mean=500*0.01, sd=sqrt(500*0.01*(1-0.01)))
```

0.588906616380991

Example 7.

Assume that approximately 60% of young men in the population have higher than recommended serum cholesterol levels. In a random selection of 200 young men, how likely will more than 120 of them have higher than recommended serum cholesterol levels?

In [30]:
```
# X... number of young men out of 200 with higher than recommended serum cholesterol levels
# X ~ Bi(200; 0.6)
# P(X>120)=1 - P(X ≤ 120)

1 - pbinom(120, size=200, prob=0.6)
```

0.473155725672359

In [31]:
```
# X~N(200 * 0.6; 200 * 0.6(1-0.6)), ie X ≈ N(120; 48)
# 1 - P(X ≤ 120)~1 - P(X<120.5)(continuity correction)

1 - pnorm(120.5, mean=200*0.6, sd=sqrt(200*0.6*(1-0.6)))
```

0.471233793707209