Moodle Tasks (Continuous Random Variables)

Задача 1

Генерирайте 100 случайни наблюдения над X. Постройте боксплот и хистограма, добавете емперичните и теоретичната плътност. Ако:

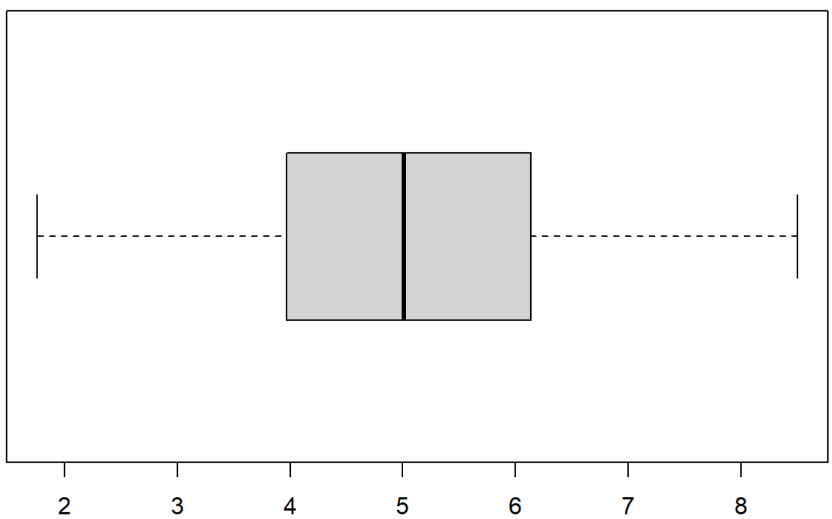
a) $X \in N(5,2)$
b) $X \in U(1,5)$
c) $X \in Exp(3)$
d) $X \in \Gamma(5,1)$
e) $X \in \mathscr{X}^2(3)$
f) $X \in t(5)$
g) X е съчетание от две разпределения $N(1,2)$ и $N(5,2)$ с вероятност за първото $p = 0.4$.

Определете вида на разпределението (симетрично или изместено, леки или тежки опашки, едномодални и т.н.)

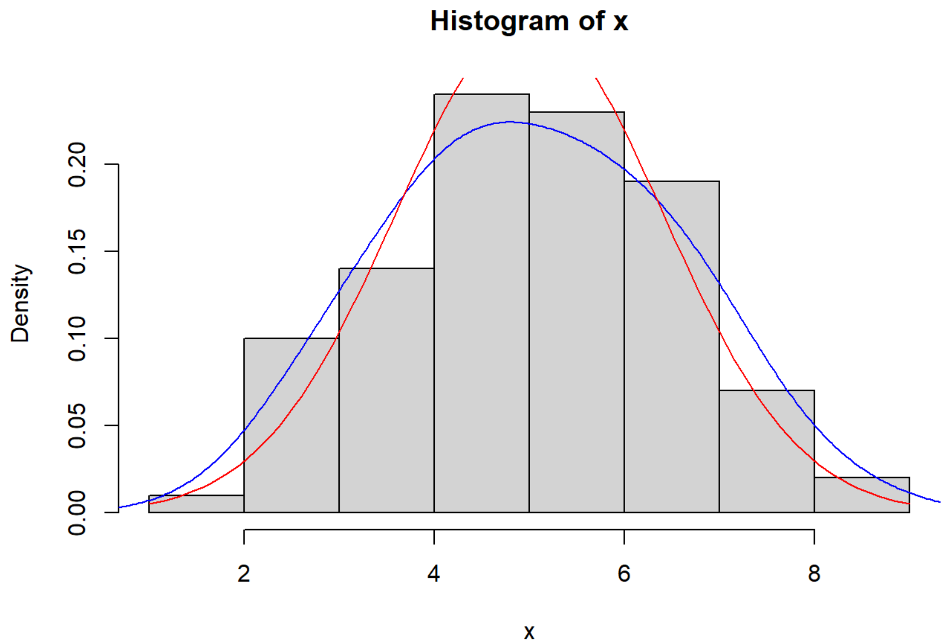
Решение:

a) $X \in N(5,2)$

```
> x <- rnorm(100, mean = 5, sd = sqrt(2))
> boxplot(x, horizontal = TRUE)
```




```
> hist(x, probability = TRUE)
> lines(density(x, bw = "SJ"), col = "blue")
> curve(dnorm(x, mean = 5, sd = sqrt(2)), add = TRUE, col = "red")
```




```
> library(EnvStats)
> skewness(x)
[1] 0.07139546
```

Коефициентът на асиметрия е приблизително 0, т.е. имаме симетрично разпределение. Какво значи приблизително в случая?
Знаем, че коефициентът на асиметрия на стандартното нормално разпеделение е 0, с дисперсия на съответния емпеeричен коефициент

$$D_s = \frac{6(n-2)}{(n+1)(n+3)} = \frac{6(100-2)}{(100+1)(100+3)} \approx 0.05$$

Този коефициент не зависи от параметрите на разпределението, т.е. дисперсията на коефициента на асиметрия, който наблюдаваме е също толкова.

Емпиричният коефициент на асиметрия $= skew_n \in N(0, 0.05)$

Тогава стандартното отклонение на този коефициент е

$$\sqrt{D_s} \approx \sqrt{0.05} \approx 0.23$$

```
> qnorm(0.995, 0, 1)
[1] 2.575829
```

Ето защо в 99% от извадките, които ще генерираме при нормално разпределение коефициентът на асиметрия, който ще получим е в интервала

$$(-2.576\sqrt{D_s};\ 2.576\sqrt{D_s})$$

Ако искаме да получим по-добра точност трябва да увеличим обема на извадката.

```
> kurtosis(x)
[1] -0.6267447
```

Коефициентът на изостреност /ексцес/ е приблизително 0, т.е. имаме mesokurtic разпределение (с нормален ексцес). Какво значи приблизително в случая?

Знаем, че коефициентът на ексцес на стандартното нормално разпеделение е 0, с дисперсия на съответния емперичен коефициент

$$D_k = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)} = \frac{24 \times 100(100-2)(100-3)}{(100+1)^2(100+3)(100+5)} \approx 0.21$$

Този коефициент не зависи от параметрите на разпределението, т.е. дисперсията на коефициента на ексцес, който наблюдаваме е също толкова.

Емпиричният коефициент на ексцес $= kurt_n \in N(0, 0.21)$

Тогава стандартното отклонение на този коефициент е

$$\sqrt{D_k} \approx \sqrt{0.21} \approx 0.45$$

```
> qnorm(0.995, 0, 1)
[1] 2.575829
```

Ето защо в 99% от извадките, които ще генерираме при нормално разпределение коефициентът на ексцес, който ще получим е в интервала

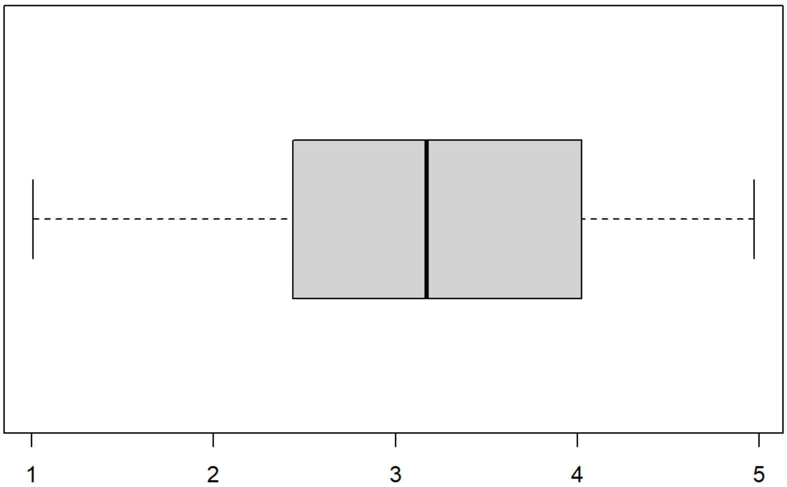
$$(-2.576\sqrt{D_k};\ 2.576\sqrt{D_k})$$
$$(-2.576 \times 0.45;\ 2.576 \times 0.45)$$
$$(-1.1592;\ 1.1592)$$

Ако искаме да получим по-добра точност трябва да увеличим обема на извадката.

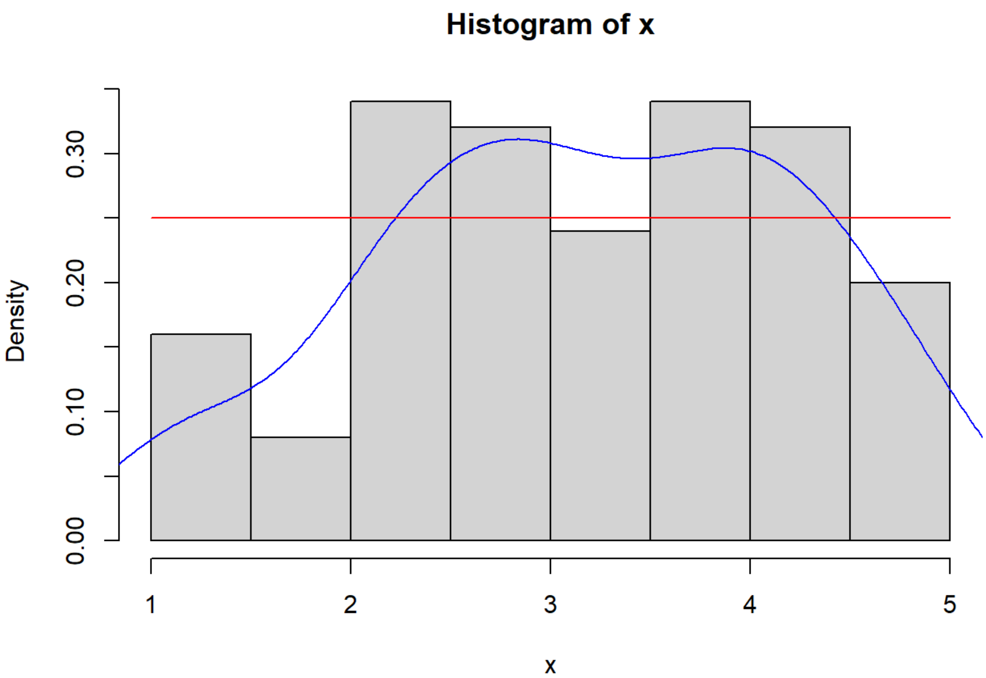
Разпределението ни е едномодално.

b) $X \in Unif(1,5)$

```
> x <- runif(100, min = 1, max = 5)
> boxplot(x, horizontal = TRUE)
```




```
> hist(x, probability = TRUE)
> lines(density(x, bw = "SJ"), col = "blue")
> curve(dunif(x, min = 1, max = 5), add = TRUE, col = "red")
```




```
> skewness(x)
[1] -0.2302741
```

Коефициентът на асиметрия на равномерното разпределение е 0. Коефициентът на асиметрия в извадката е приблизително 0, т.е. разпределението е симетрично. Т.к. точното разпределение на емпиричния коефициент на асметрия в случая не е известно, смисълът на думата приблизително можем да го анализираме само при големи извадки.

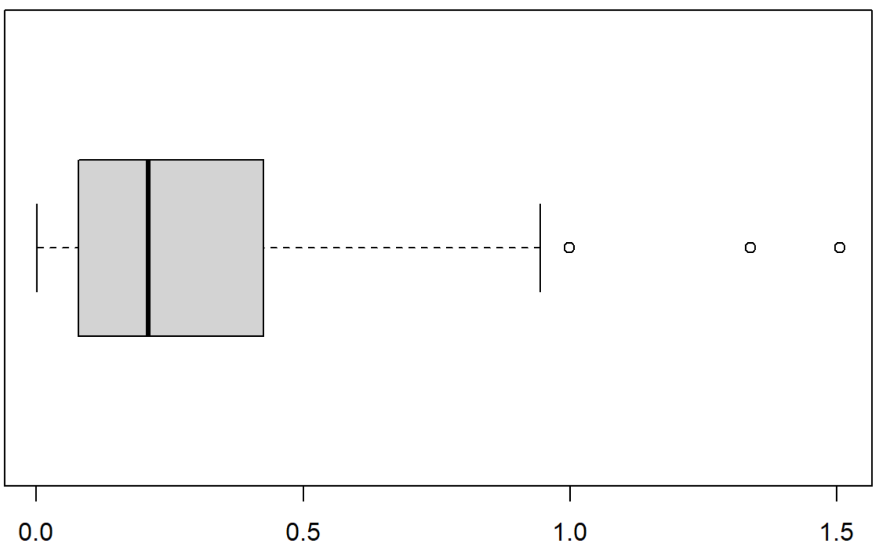
```
> kurtosis(x)
[1] -0.7617731
```

Коефициентът на изостреност /ексцес/ на равномерното разпределение е отрицателен $-\frac{6}{5}$. Коефициентът на изостреност /ексцес/ в извадката е също отрицателен -1.104895, т.е. имаме platykurtic разпределение (поднормален ексцес).

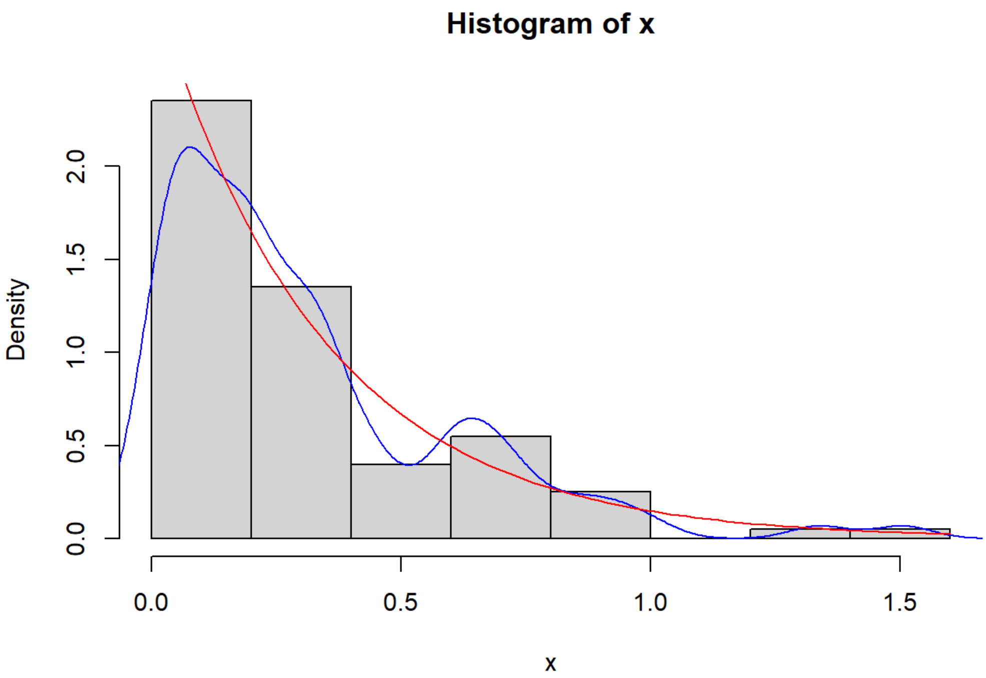
Разпределението ни е едномодално.

c) $X \in Exp(3)$

```
> x <- rexp(100, rate = 3)
> boxplot(x, horizontal = TRUE)
```




```
> hist(x, probability = TRUE)
> lines(density(x, bw = "SJ"), col = "blue")
> curve(dexp(x, rate = 3), add = TRUE, col = "red")
```




```
> skewness(x)
[1] 1.572087
```

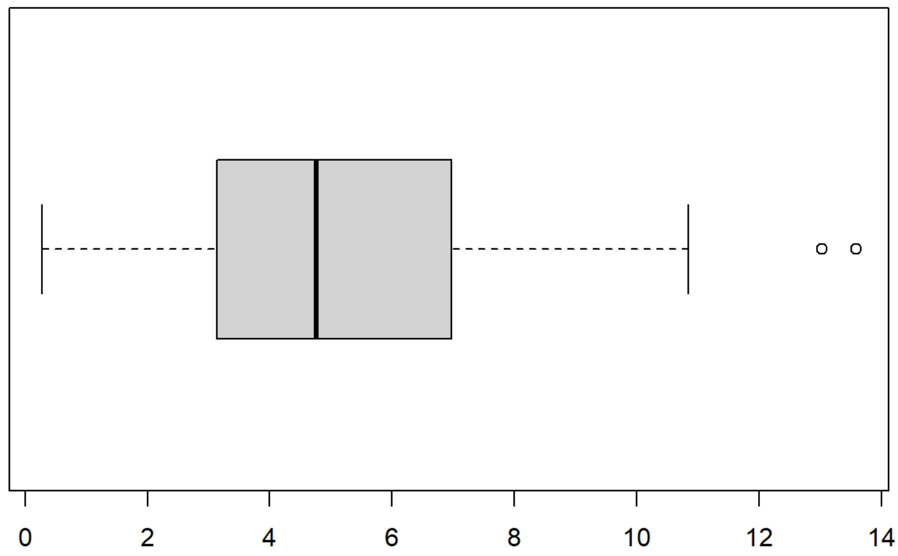
Коефициентът на асиметрия на експоненциалното разпределение е 2. За това получените коефициенти в случая ще бъдат положителни приблизително 2, т.е. имаме разпределение с дясна асиметрия. Т.к. точното разпределение на емперичния коефициент на асметрия в случая не е известно, смисълът на думата приблизително можем да го анализираме само при големи извадки.

```
> kurtosis(x)
[1] 2.904433
```

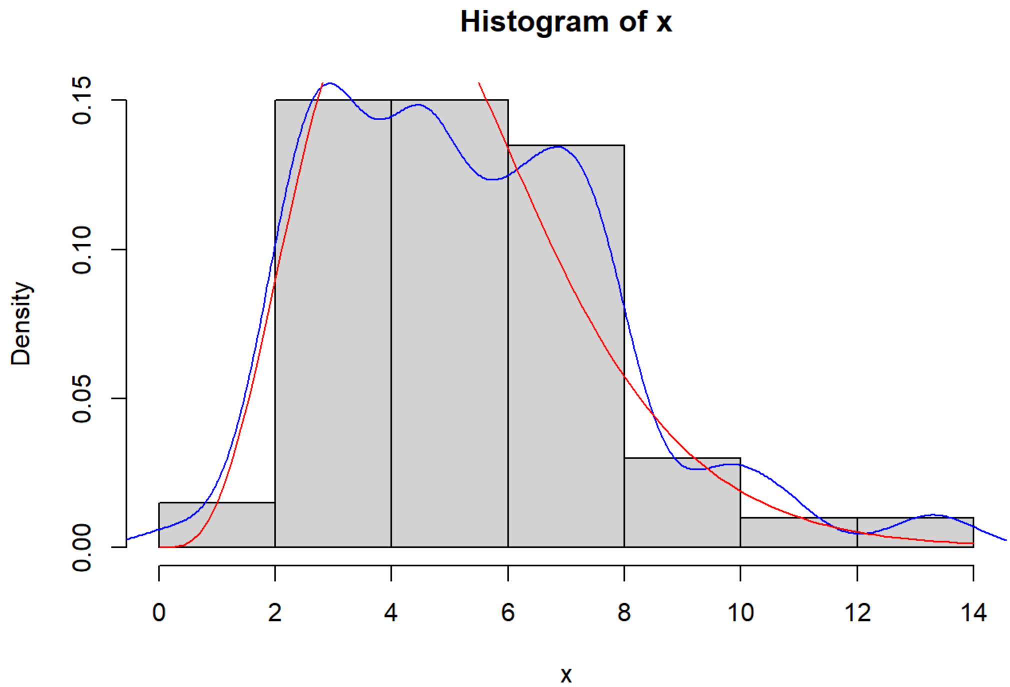
Коефициентът на изостреност /ексцес/ на експоненциалното разпределение е положителен 6. Коефициентът на изостреност /ексцес/ в извадката ще е също положителен приблизително 6, т.е. имаме leptokurtic разпределение (наднормален ексцес).
Разпределението ни е едномодално.

d) $X \in \Gamma(\alpha = 5, \beta = 1), \theta = \dfrac{1}{\beta} = 1$

```
> x <- rgamma(100, shape = 5, rate = 1)
> boxplot(x, horizontal = TRUE)
```




```
> hist(x, probability = TRUE)
> lines(density(x, bw = "SJ"), col = "blue")
> curve(dgamma(x, shape = 5, rate = 1), add = TRUE, col = "red")
```



```
> skewness(x)
[1] 0.8066322
```

Коефициентът на асиметрия на гама разпределението зависи от параметъра α и е $\dfrac{2}{\sqrt{\alpha}} = \dfrac{2}{\sqrt{5}} \approx 0.89$. За това получените коефициенти в случая ще бъдат приблизително 0.89, т.е. имаме разпределение с дясна асиметрия. Т.к. точното разпределение на емперичния коефициент на асметрия в случая не е известно, смисълът на думата приблизително можем да го анализираме само при големи извадки.

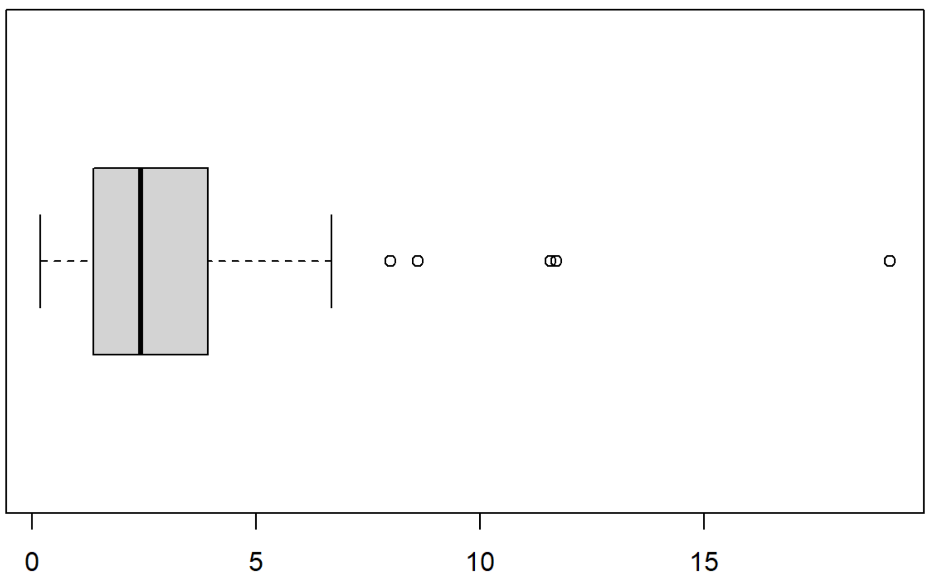
```
> kurtosis(x)
[1] 0.8165433
```

Коефициентът на изостреност /ексцес/ на гама разпределението зависи от параметъра α и е $\dfrac{6}{\alpha} = \dfrac{6}{5} = 1.2$. За това получените коефициенти в случая ще бъдат приблизително 1.2, т.е. имаме leptokurtic разпределение (наднормален ексцес).

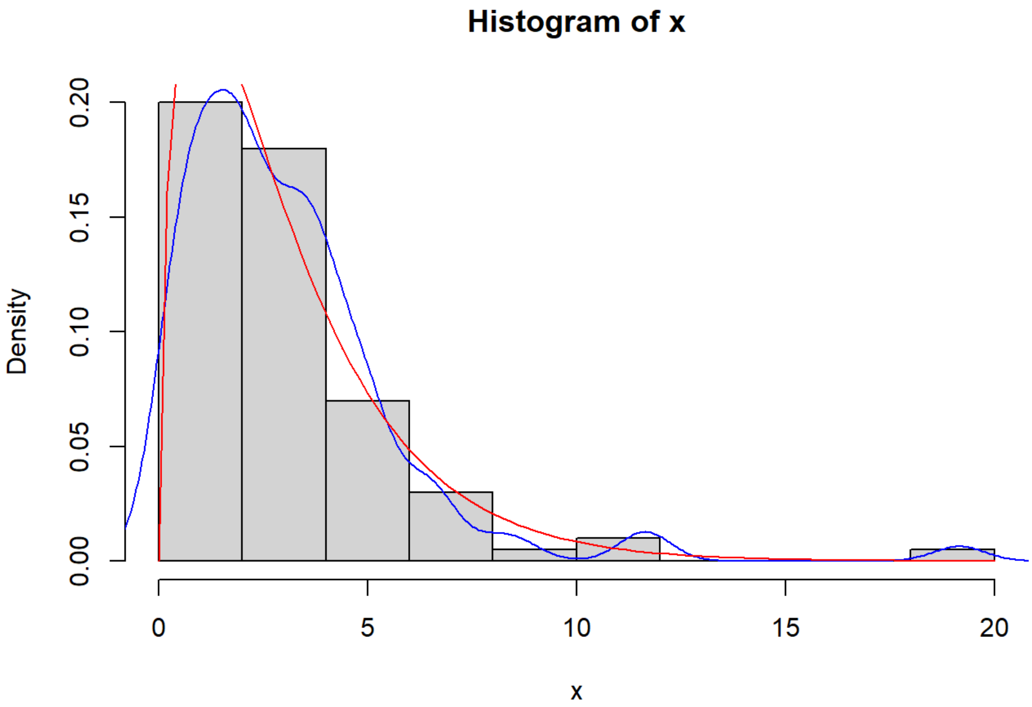
Разпределението ни е едномодално.

е) $X \in \mathscr{X}^2(3)$

```
> x <- rchisq(100, df = 3)
> boxplot(x, horizontal = TRUE)
```




```
> hist(x, probability = TRUE)
> lines(density(x, bw = "SJ"), col = "blue")
> curve(dchisq(x, df = 3), add = TRUE, col = "red")
```




```
> skewness(x)
[1] 2.735068
```

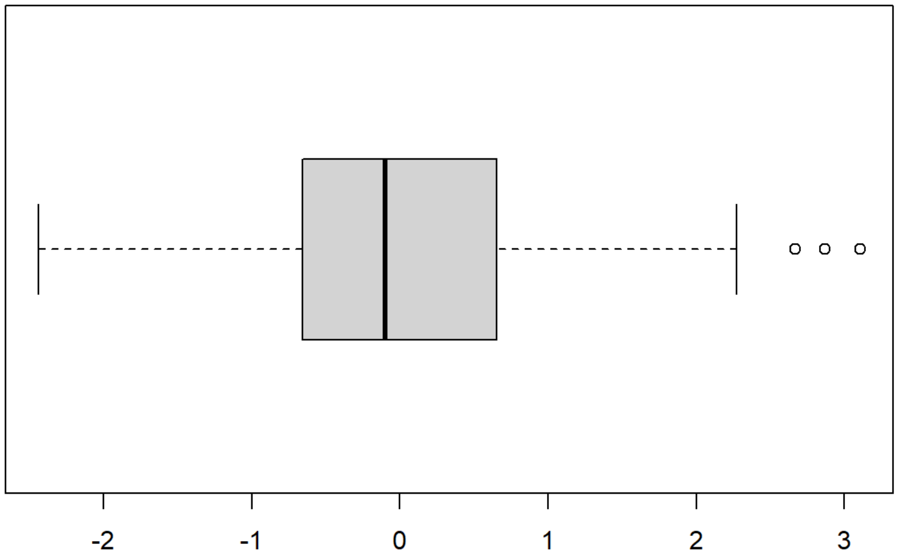
Коефициентът на асиметрия на $\mathscr{X}^2$ разпределението зависи от степените на свобода $\sqrt{\frac{8}{n}} = \sqrt{\frac{8}{3}} \approx 1.63$. За това получените коефициенти в случая ще бъдат приблизително 1.63, т.е. имаме разпределение с дясна асиметрия. Т.к. точното разпределение на емперичния коефициент на асметрия в случая не е известно, смисълът на думата приблизително можем да го анализираме само при големи извадки.

```
> kurtosis(x)
[1] 11.94065
```

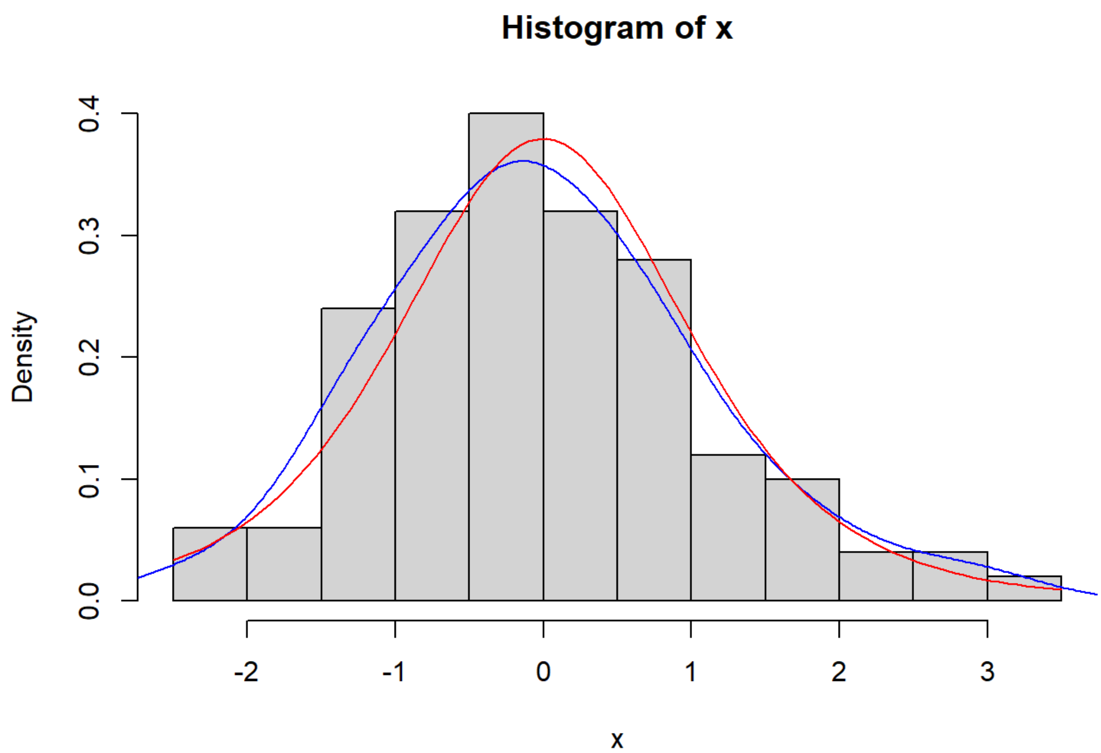
Коефициентът на изостреност /ексцес/ на $\mathscr{X}^2$ разпределението зависи от степените на свобода n и е $\frac{12}{n} = \frac{12}{3} = 4$. За това получените коефициенти в случая ще бъдат приблизително 4, т.е. имаме leptokurtic разпределение (наднормален ексцес).
Разпределението ни е едномодално.

f) $X \in t(5)$

```
> x <- rt(100, df = 5)
> boxplot(x, horizontal = TRUE)
```




```
> hist(x, probability = TRUE)
> lines(density(x, bw = "SJ"), col = "blue")
> curve(dt(x, df = 5), add = TRUE, col = "red")
```




```
> skewness(x)
[1] 0.4129929
```

Коефициентът на асиметрия на Student-t разпределението с повече от 3 степени на свобода $\nu > 3$ е 0. В противен случай той не съществува. Коефициентът на асиметрия в извадката е приблизително 0, т.е. разпределението е симетрично. Т.к. точното разпределение на емперичния коефициент на асметрия в случая не е известно, смисълът на думата приблизително можем да го анализираме само при големи извадки.

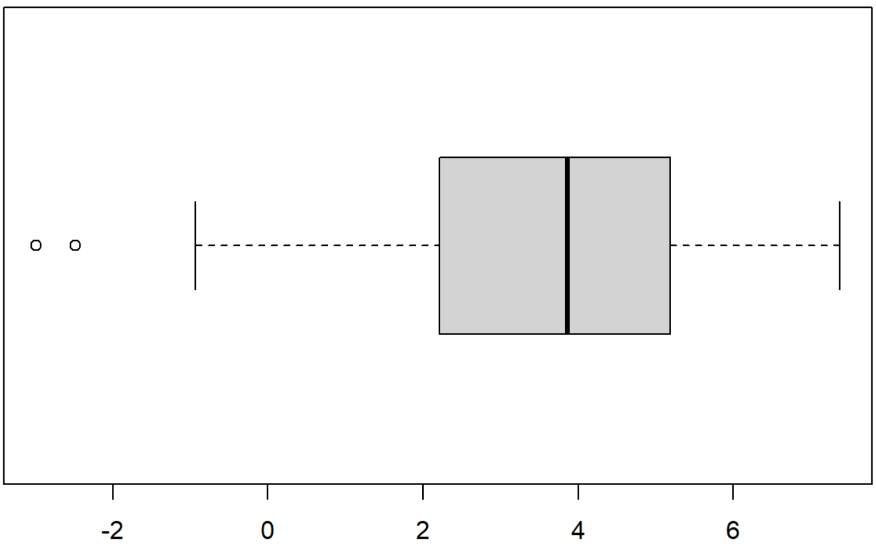
```
> kurtosis(x)
[1] 0.3645355
```

Коефициентът на изостреност /ексцес/ на Student-t разпределението с повече от 4 степени на свобода $\nu > 4$ е $\frac{6}{\nu - 4} = \frac{6}{5-4} = 6$. В противен случай той не съществува. Коефициентът на изостреност /ексцес/ в извадката е приблизително 6, т.е. имаме leptokurtic разпределение (наднормален ексцес).
Разпределението ни е едномодално.

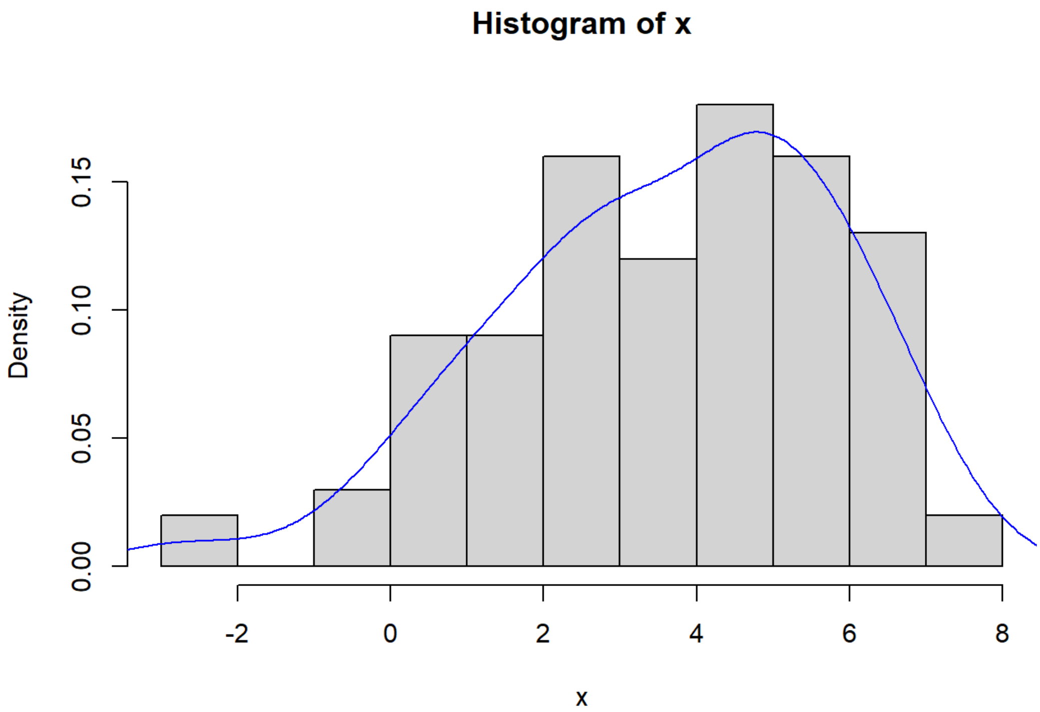
g) X е смес от две разпределения $N(1,2)$ и $N(5,2)$ с вероятност за първото $p = 0.4$.

$I_A \in Bernoulli(0.4)$
$X = I_A \times N(1,2) + I_{\overline{A}} \times N(5,2)$

```
> i <- rbinom(100, size = 1, prob = 0.4)
> n1 <- rnorm(100, mean = 1, sd = sqrt(2))
> n2 <- rnorm(100, mean = 5, sd = sqrt(2))
> x <- i * n1 + (1 - i) * n2
> boxplot(x, horizontal = TRUE)
```



```
> hist(x, probability = TRUE)
> lines(density(x, bw = "SJ"), col = "blue")
```




```
> skewness(x)
[1] -0.6004053
> kurtosis(x)
[1] 0.0963946
```

т.к. теоретичните стойности на коефициентите на асиметрия и ексцес в случая не са известни ще ги определим само емпирично.

Коефициентът на асиметрия в извадката е приблизително 0, т.е. имаме почти симетрично емпирично разпределение.

Коефициентът на изостреност /ексцес/ е отрицателен, т.е. имаме platykurtic емпирично разпределение (поднормален ексцес).

Разпределението ни е двумодално.

Задача 2

Нека $X_1, X_2, \ldots, X_n$ са независими сл.в. зададени както в Зад.1. Какво можете да кажете за разпределението на $Y = X_1 + X_2 + \ldots + X_n$.

Разгледайте случаите $n = 2,\ 10,\ 100$.

Решение:

а) $X_1, X_2, \ldots, X_n \in N(5{,}2)$.

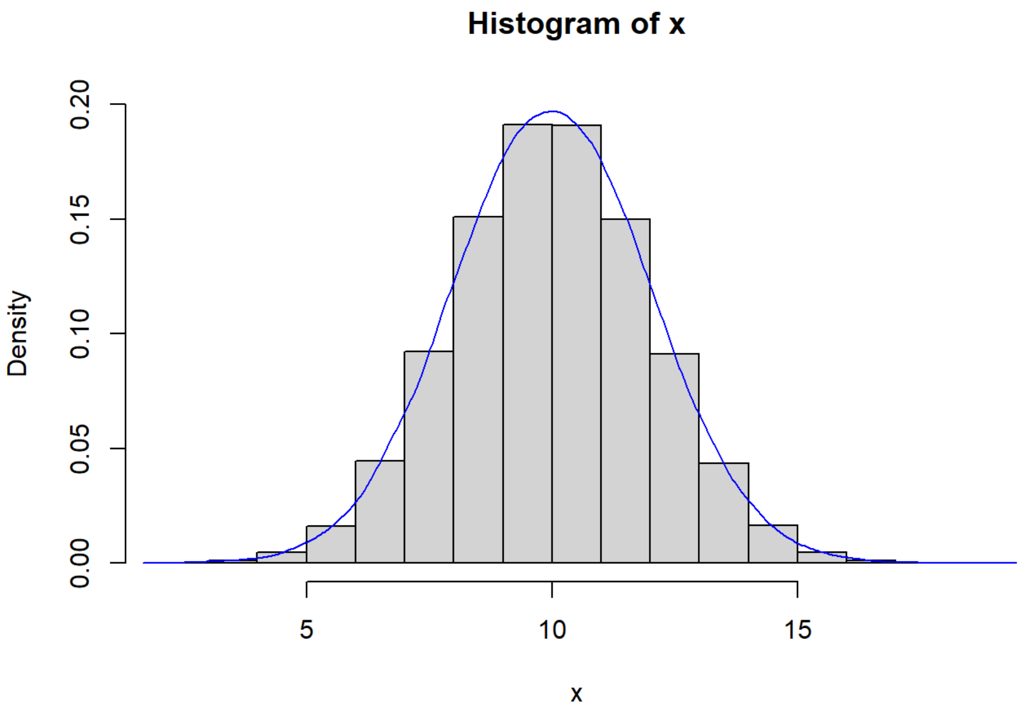
От $X_i \in N(\mu_i, \sigma_i^2)$ и X_i независими за $i = 1, \ldots, n$, следва

$$X_1 + X_2 + \ldots + X_n \in N(\mu_1 + \mu_2 + \ldots + \mu_n, \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2)$$

В нашия случай $Y = X_1 + X_2 + \ldots + X_n \in N(5n, 2n)$

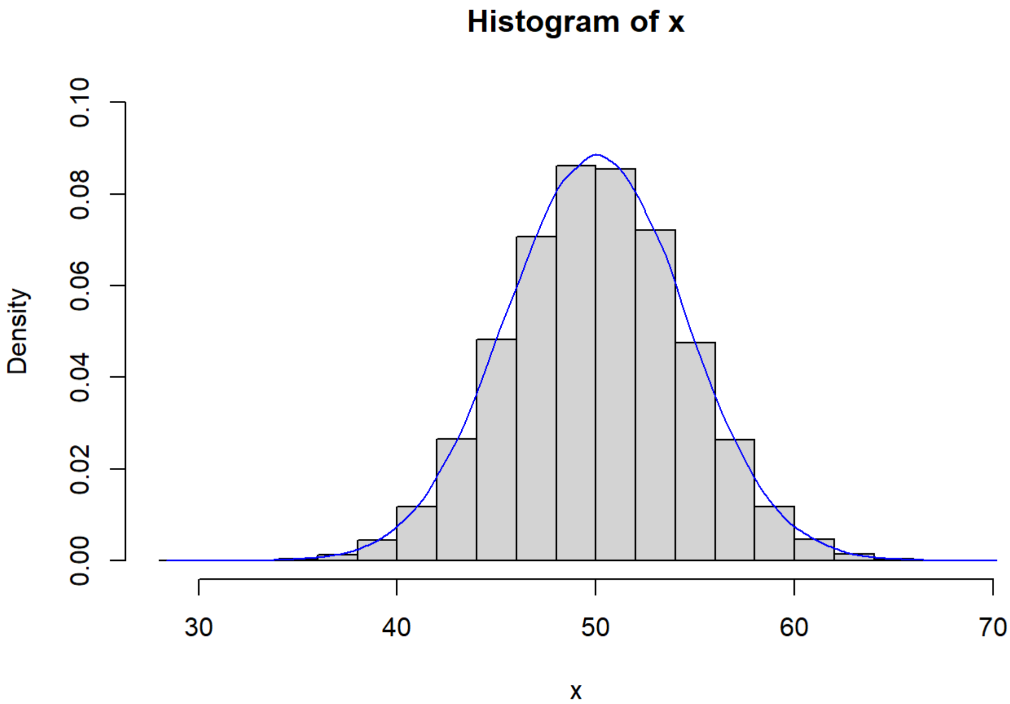
При $n = 2$, $Y = X_1 + X_2 \in N(10, 4)$

```
> n <- 2
> x <- rnorm(10^5, mean = 5*n, sd = sqrt(2*n))
> hist(x, probability = TRUE, ylim = c(0, 0.2))
> lines(density(x, bw = "SJ"), col = "blue")
```



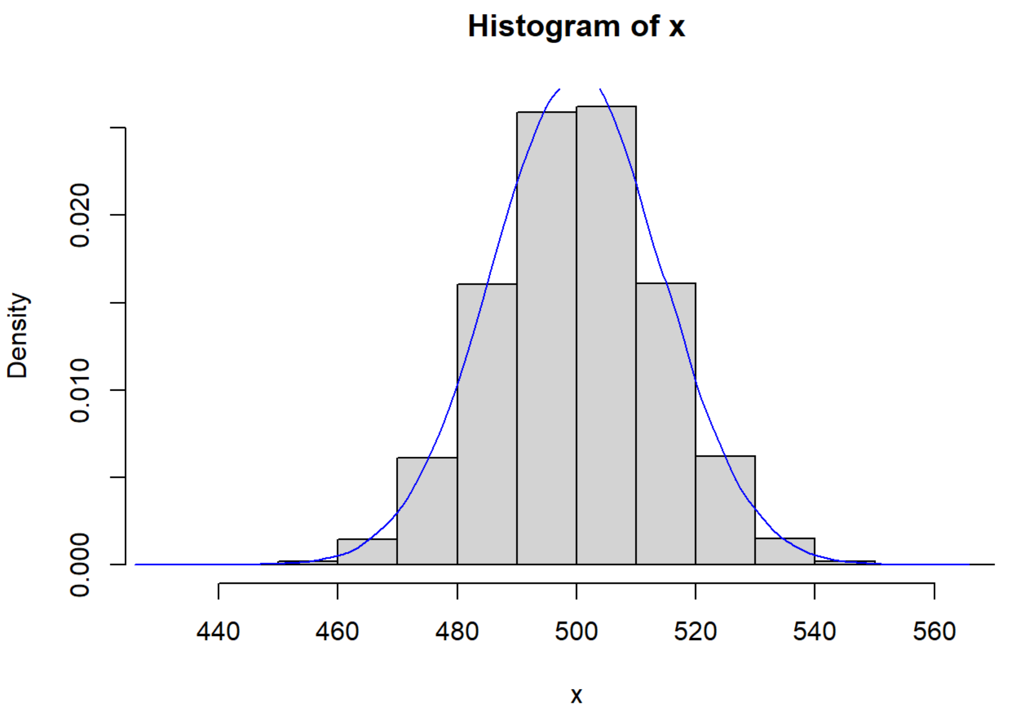
При $n = 10: \; Y_1 = X_1 + \ldots + X_{10} \in N(50, 20)$

```
> n <- 10
> x <- rnorm(10^5, mean = 5*n, sd = sqrt(2*n))
> hist(x, probability = TRUE, ylim = c(0, 0.1))
> lines(density(x, bw = "SJ"), col = "blue")
```



При $n = 100: \ \ Y = X_1 + \ldots + X_{100} \in N(500,\, 200)$

```
> n <- 100
> x <- rnorm(10^5, mean = 5*n, sd = sqrt(2*n))
> hist(x, probability = TRUE)
> lines(density(x, bw = "SJ"), col = "blue")
```



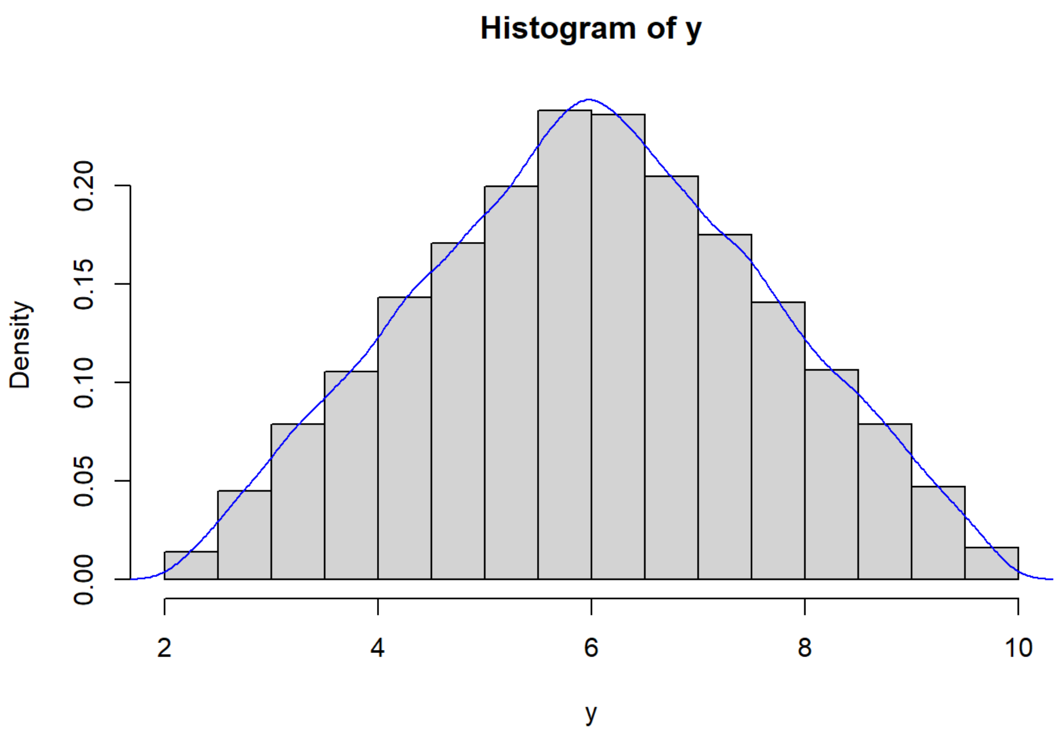
b) $X_1, X_2, \ \dots, X_n \in Unif(1{,}5)$

$$\mathbb{E}[X_1 + X_2 + \dots + X_n] = m\mathbb{E}[X_1] = n \times \frac{1+5}{2} = 3n$$
$$\mathbb{D}[X_1 + X_2 + \dots + X_n] = n\mathbb{D}[X_1] = n \times \frac{(5-1)^2}{12} = \frac{4}{3}n$$

При $n\ = 2: \ \ Y = X_1 + X_2$

$$\mathbb{E}[X_1 + X_2] = 2\mathbb{E}[X_1] = 2 \times 3 = 6$$
$$\mathbb{D}[X_1 + X_2] = 2\mathbb{D}[X_1] = 2\frac{4}{3} = \frac{8}{3} \approx 2.67$$

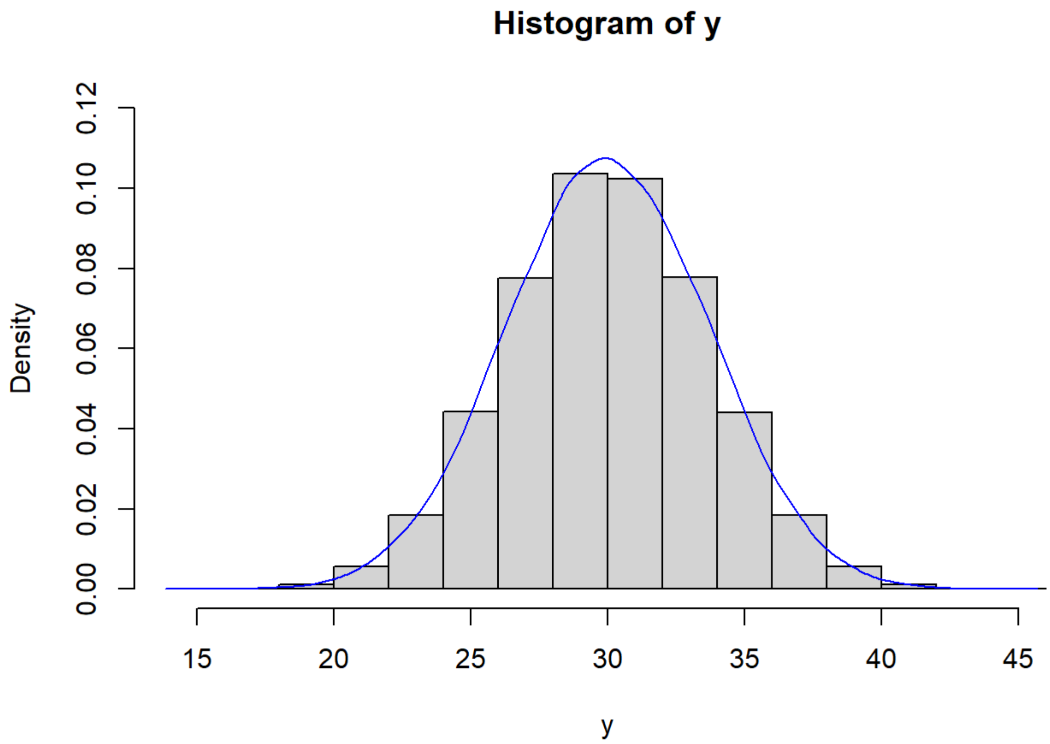
```
> x1 <- runif(50000, min = 1, max = 5)
> x2 <- runif(50000, min = 1, max = 5)
> y <- x1 + x2
> mean(y)
[1] 6.010569
> var(y)
[1] 2.648775
> hist(y, probability = TRUE)
> lines(density(y, bw = "SJ"), col = "blue")
```



При $n\ = 10: \ \ Y = X_1 + \dots + X_{10}$

$$\mathbb{E}[X_1 + X_2 + \dots + X_{10}] = 10\mathbb{E}[X_1] = 10 \times 3 = 30$$
$$D[X_1 + X_2 + \dots + X_{10}] = 10D[X_1] = 10 \times \frac{4}{3} = \frac{40}{3} \approx 13.33$$

```
> y <- 0
> for(i in 1:(10^5)){
+   y[i] <- sum(runif(10, min = 1, max = 5))
+ }
> mean(y)
[1] 29.99417
> var(y)
[1] 13.29523
> hist(y, probability = TRUE, ylim = c(0, 0.12))
> lines(density(y, bw = "SJ"), col = "blue")
```

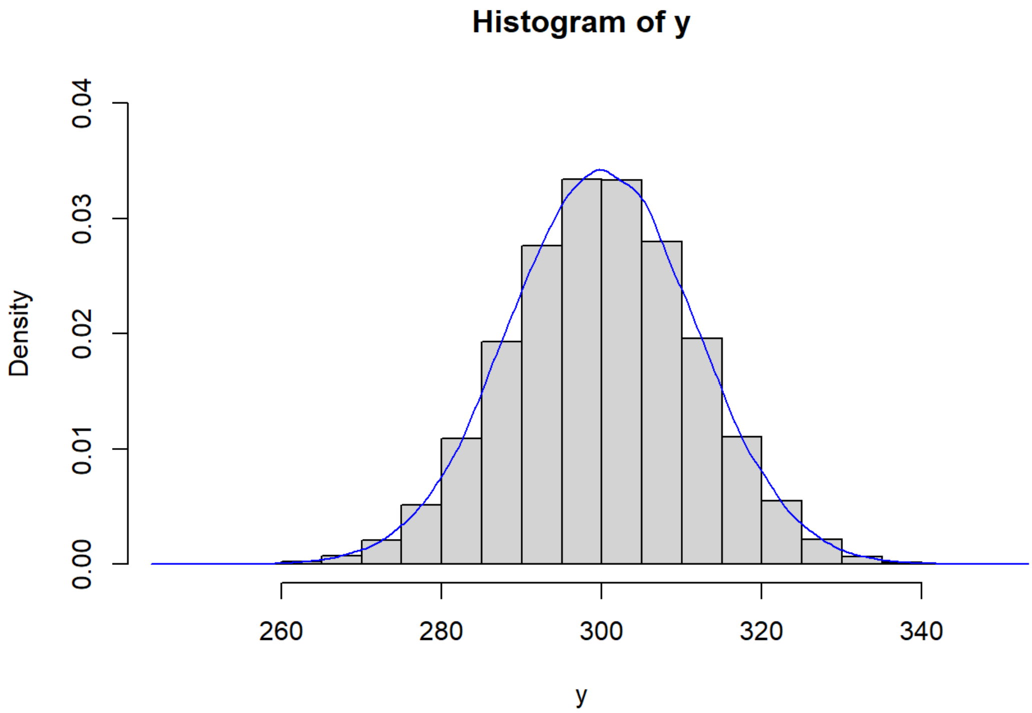


При $n = 100: \;\; Y = X_1 + \ldots + X_{100}$

$$\mathbb{E}[X_1 + X_2 + \ldots + X_{100}] = 100\mathbb{E}[X_1] = 100 \times 3 = 300$$

$$\mathbb{D}[X_1 + X_2 + \ldots + X_{100}] = 10\mathbb{D}[X_1] = 100 \times \frac{4}{3} = \frac{400}{3} \approx 133.33$$

```
> y <- 0
> for(i in 1:(10^5)){
+   y[i] <- sum(runif(100, min = 1, max = 5))
+ }
> mean(y)
[1] 300.0705
> var(y)
[1] 133.6929
> hist(y, probability = TRUE, ylim = c(0, 0.04))
> lines(density(y, bw = "SJ"), col = "blue")
```



т.к. $100 >> 30$ е голямо можем да използваме централна гранична теорема (ЦГТ). От централна гранична теорема (ЦГТ), ако X_i еднакво разпределени, независими и с **крайна дисперсия** за $i = 1, \ldots, n$, то при увеличаване на обема на извадката

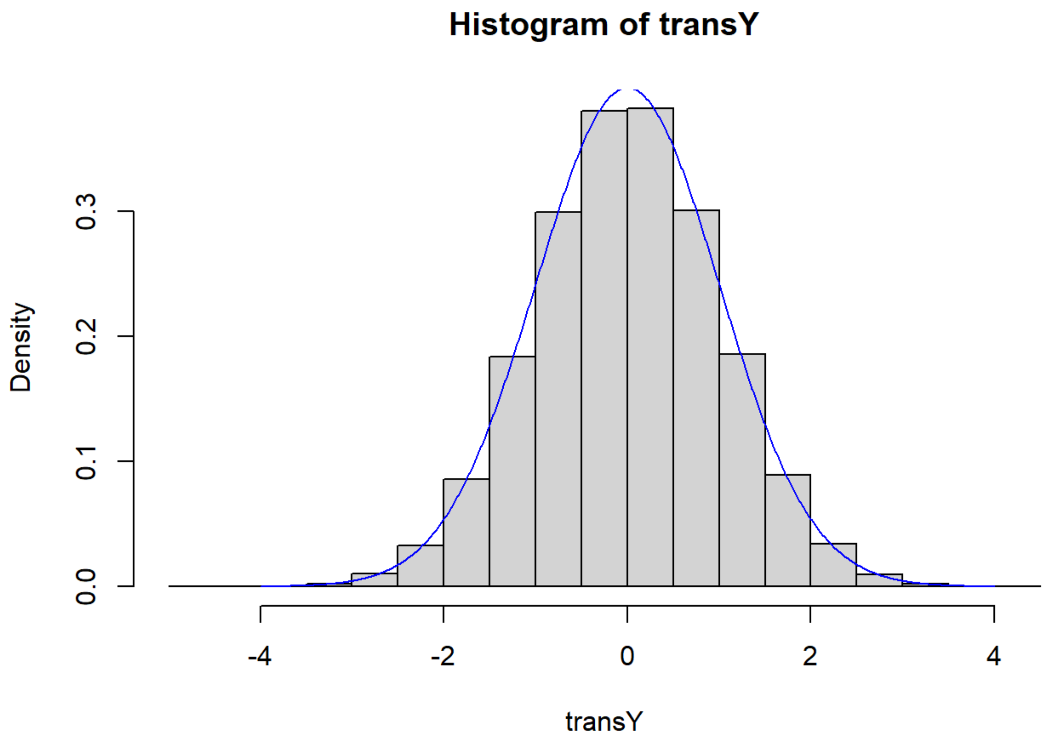
$$\frac{X_1 + X_2 + \ldots + X_n - n\mathbb{E}[X]}{\sqrt{n\mathbb{D}[X]}} \xrightarrow{d} N(0,1)$$

В случая, $\mathbb{E}[X_i] = \dfrac{a+b}{2} = \dfrac{1+5}{2} = 3$, $\mathbb{D}[X_i] = \dfrac{(b-a)^2}{12} = \dfrac{(5-1)^2}{12} = \dfrac{4}{3}$

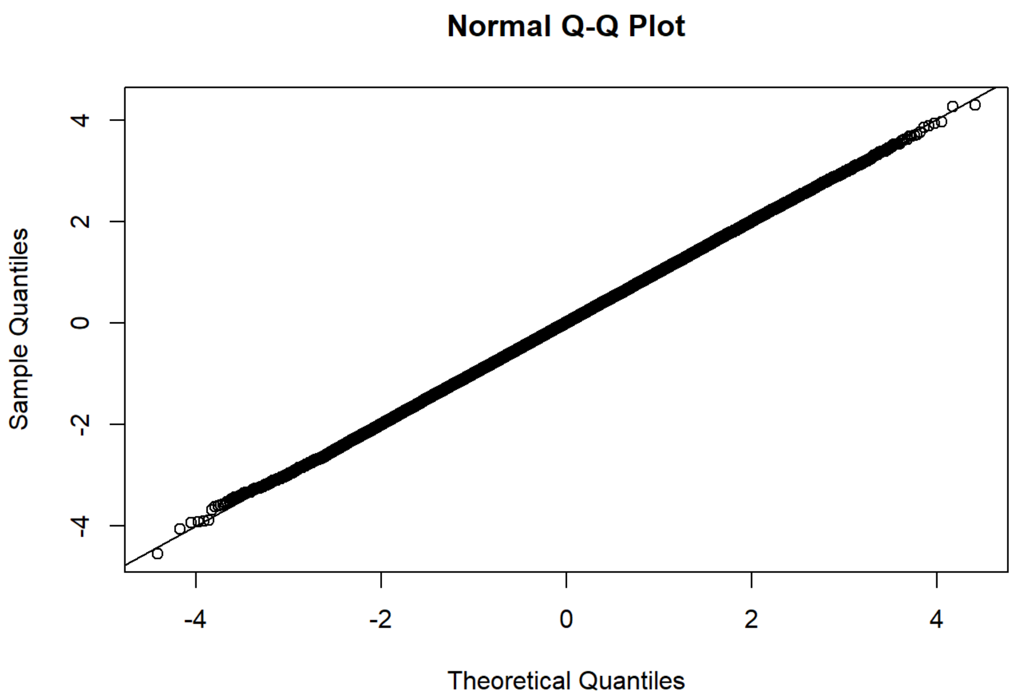
$Y = X_1 + X_2 + \ldots + X_n$

$$\frac{Y - n\mathbb{E}[X]}{\sqrt{n\mathbb{D}[X]}} \xrightarrow{d} N(0,1), \quad \frac{Y - 3n}{\sqrt{\frac{4}{3}n}} \xrightarrow{d} N(0,1)$$

```
> transY <- (y - 3*n) / sqrt((4*n)/3)
> hist(transY, probability = TRUE)
> xCoord <- seq(-4, 4, 0.01)
> lines(xCoord, dnorm(xCoord, 0, 1), col = "blue")
```



> qqnorm(transY)
> qqline(transY)



c) $X_1, X_2, \ldots, X_n \in Exp(3)$

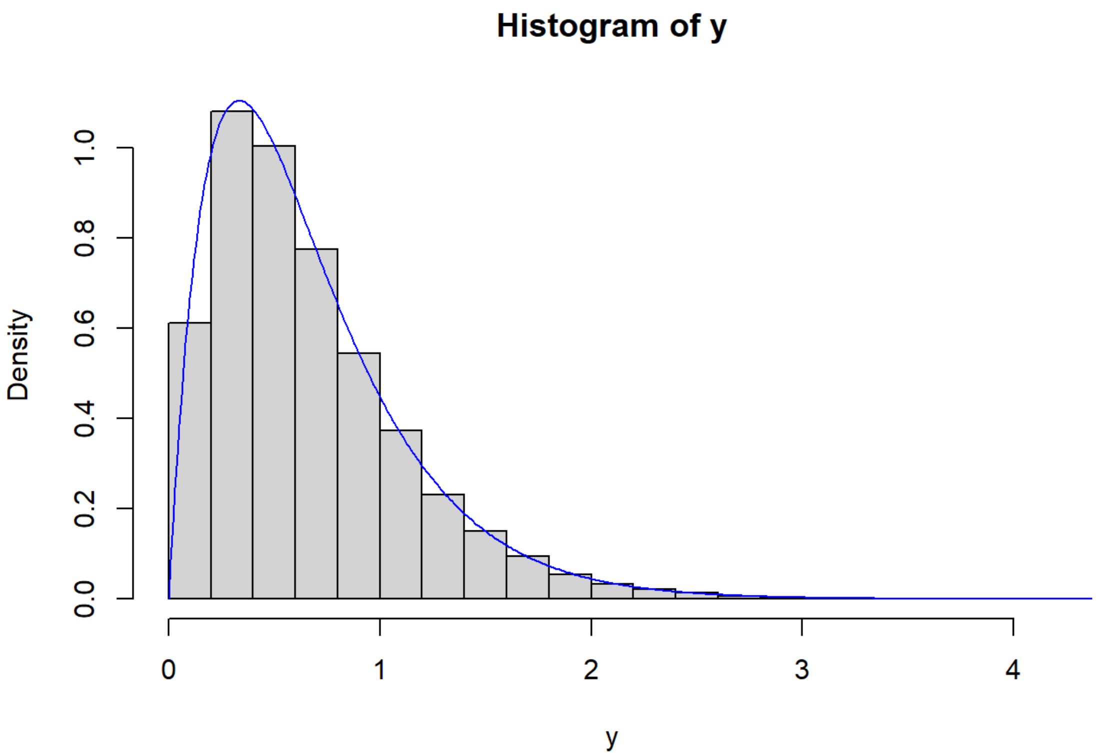
Тъй като $Exp(\lambda) = \Gamma(1,\lambda)$ и сума на независими гама случайни величини с един и същ втори параметър е отново гама със същия втори параметър, а първите параметри се сумират, то

$$Y = X_1 + X_2 + \ldots + X_n \in \Gamma(n,3)$$
$$f_Y(x) = \frac{3^n}{\Gamma(n)} x^{n-1} e^{-3x} = \frac{3^n}{(n-1)!} x^{n-1} e^{-3x}, x > 0$$
$$\mathbb{E}[X_1 + X_2 + \ldots + X_n] = \frac{n}{3}$$
$$\mathbb{D}[X_1 + X_2 + \ldots + X_n] = \frac{n}{3^2} = \frac{n}{9}$$

При $n = 2:\ \ Y = X_1 + X_2 \in \Gamma(2,3)$

$$f_Y(x) = \frac{3^2}{(2-1)!} x^{2-1} e^{-3x} = 9xe^{-3x}, x > 0$$
$$\mathbb{E}[X_1 + X_2] = \frac{2}{3} \approx 0.67$$
$$\mathbb{D}[X_1 + X_2] = \frac{2}{3^2} \approx 0.22$$

```
> x1 <- rexp(50000, rate = 3)
> x2 <- rexp(50000, rate = 3)
> y <- x1 + x2
> mean(y)
[1] 0.6635124
> var(y)
[1] 0.2188843
> hist(y, probability = TRUE)
> xCoord <- seq(0, 5, 0.01)
> lines(xCoord, dgamma(xCoord, shape = 2, rate = 3), col = "blue")
```



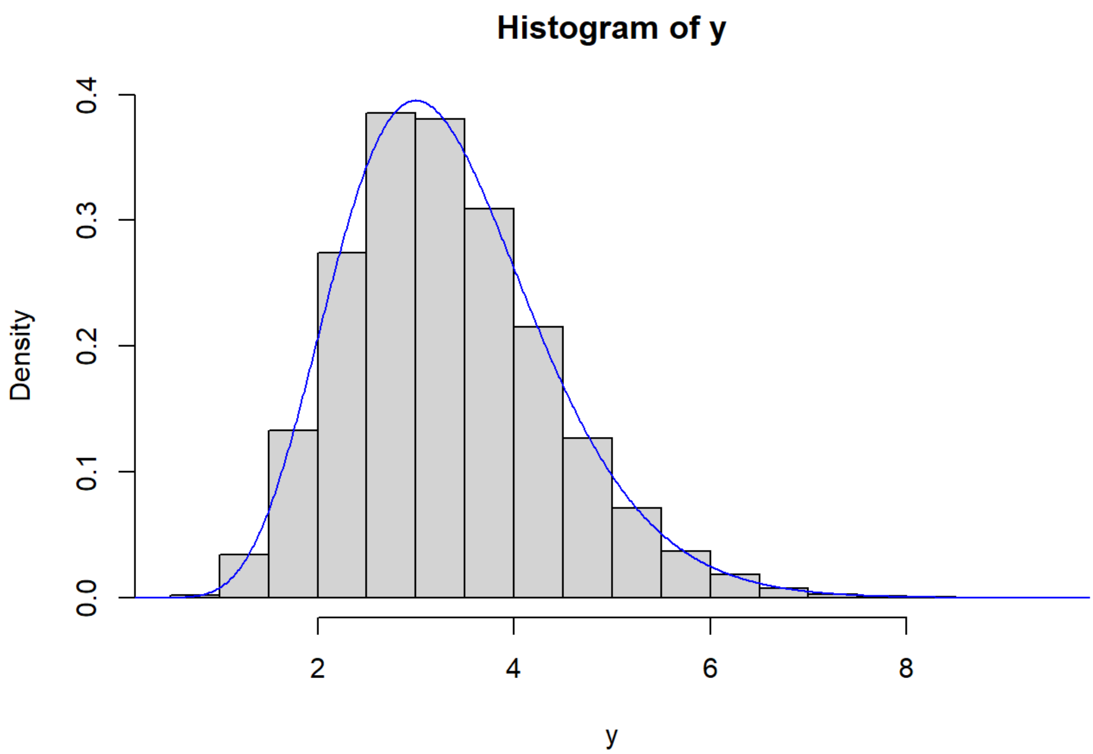
При $n = 10: \;\; Y = X_1 + \ldots + X_{10} \in \Gamma(10{,}3)$

$$f_Y(x) = \frac{10^2}{(10-1)!}x^{10-1}e^{-3x} = \frac{100}{9!}x^9e^{-3x}, x > 0$$

$$\mathbb{E}[X_1 + \ldots + X_{10}] = \frac{10}{3} \approx 3.33$$

$$\mathbb{D}[X_1 + \ldots + X_{10}] = \frac{10}{3^2} \approx 1.11$$

```
> y <- rgamma(50000, shape = 10, rate = 3)
> mean(y)
[1] 3.329143
> var(y)
[1] 1.10743
> hist(y, probability = TRUE)
> xCoord <- seq(0, 15, 0.01)
> lines(xCoord, dgamma(xCoord, shape = 10, rate = 3), col = "blue")
```



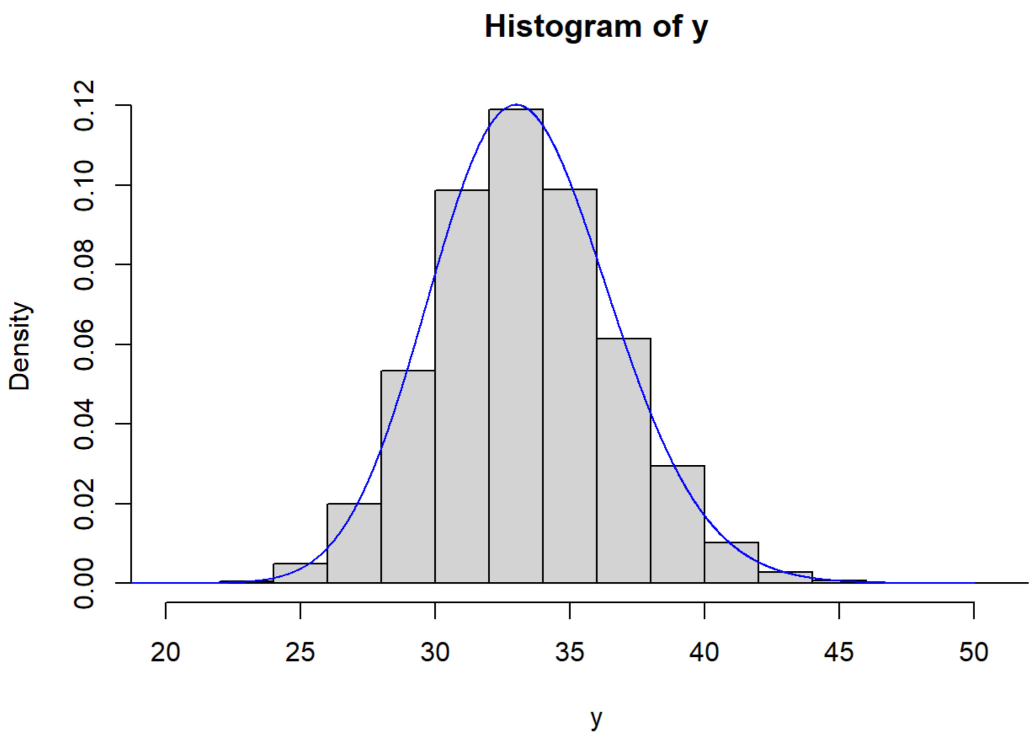
При $n = 100: \;\; Y = X_1 + \ldots + X_{100} \in \Gamma(100{,}3)$

$$f_Y(x) = \frac{100^2}{(100-1)!}x^{100-1}e^{-3x} = \frac{100^2}{99!}x^{99}e^{-3x}, x > 0$$

$$\mathbb{E}[X_1 + \ldots + X_{100}] = \frac{100}{3} \approx 33.33$$

$$\mathbb{D}[X_1 + \ldots + X_{100}] = \frac{100}{3^2} \approx 11.11$$

```
> y <- rgamma(50000, shape = 100, rate = 3)
> mean(y)
[1] 33.32296
> var(y)
[1] 11.19246
> hist(y, probability = TRUE)
> xCoord <- seq(0, 50, 0.01)
> lines(xCoord, dgamma(xCoord, shape = 100, rate = 3), col = "blue")
```



т.к. 100 е голямо можем да използваме централна гранична теорема (ЦГТ).
От централна гранична теорема (ЦГТ), ако X_i еднакво разпределени, независими и с крайна дисперсия за $i = 1, \dots, n$, то при увеличаване на обема на извадката

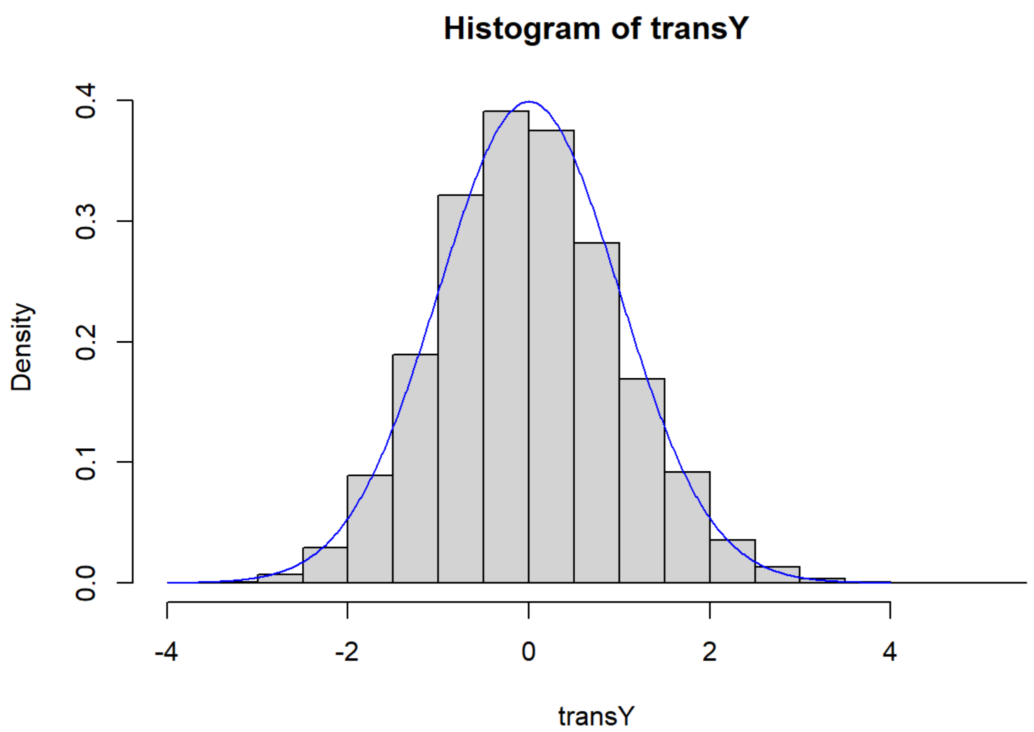
$$\frac{X_1 + X_2 + \dots + X_n - n\mathbb{E}[X]}{\sqrt{n\mathbb{D}[X]}} \xrightarrow{d} N(0,1)$$

В случая

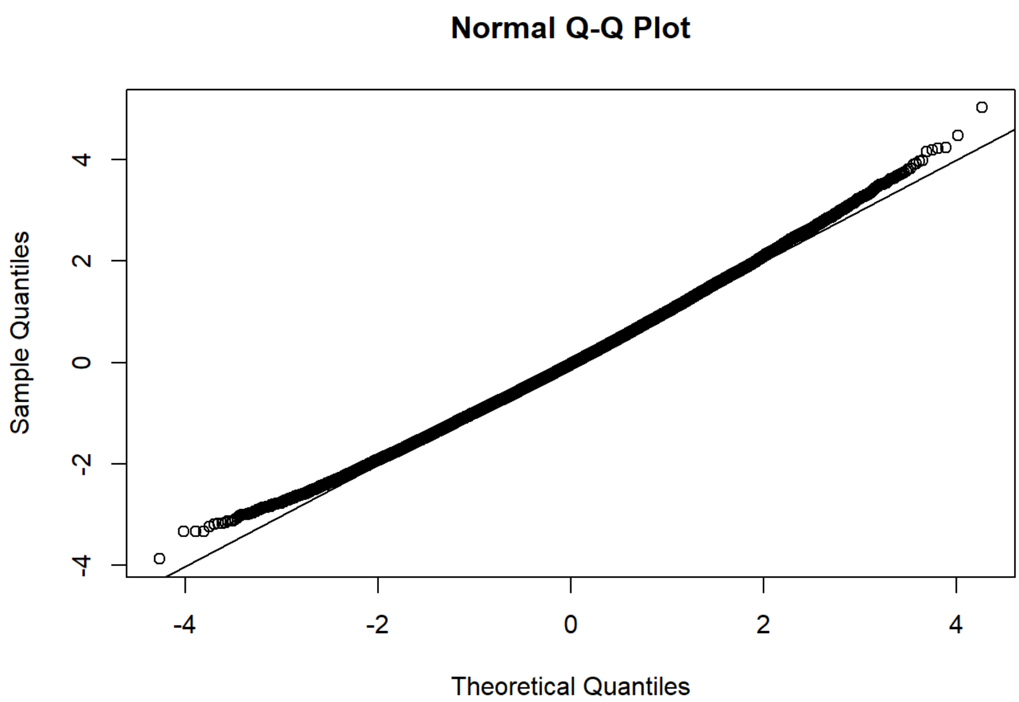
$$\mathbb{E}[X_1] = \frac{1}{3}, \mathbb{D}[X_i] = \frac{1}{3^2}.$$

$$Y = X_1 + X_2 + \dots + X_n, \frac{Y - n\frac{1}{3}}{\sqrt{n\frac{1}{3^2}}} \xrightarrow{d} N(0,1)$$

```
> transY <- (y - n/3) / sqrt(n/9)
> hist(transY, probability = TRUE)
> xCoord <- seq(-4, 4, 0.01)
> lines(xCoord, dnorm(xCoord, 0, 1), col = "blue")
```



> qqnorm(transY)
> qqline(transY)



d) $X_1, X_2, \ldots, X_n \in \Gamma(5,1)$
Тъй като сума на независими гама случайни величини с един и същ втори параметър е отново гама със същия втори параметър, а първите параметри се сумират, то

$Y = X_1 + X_2 + \ldots + X_n \in \Gamma(5n,\, 1)$

$$f_Y(x) = \frac{1}{\Gamma(5n)} x^{5n-1} e^{-x} = \frac{1}{(5n-1)!} x^{5n-1} e^{-x},\, x > 0$$

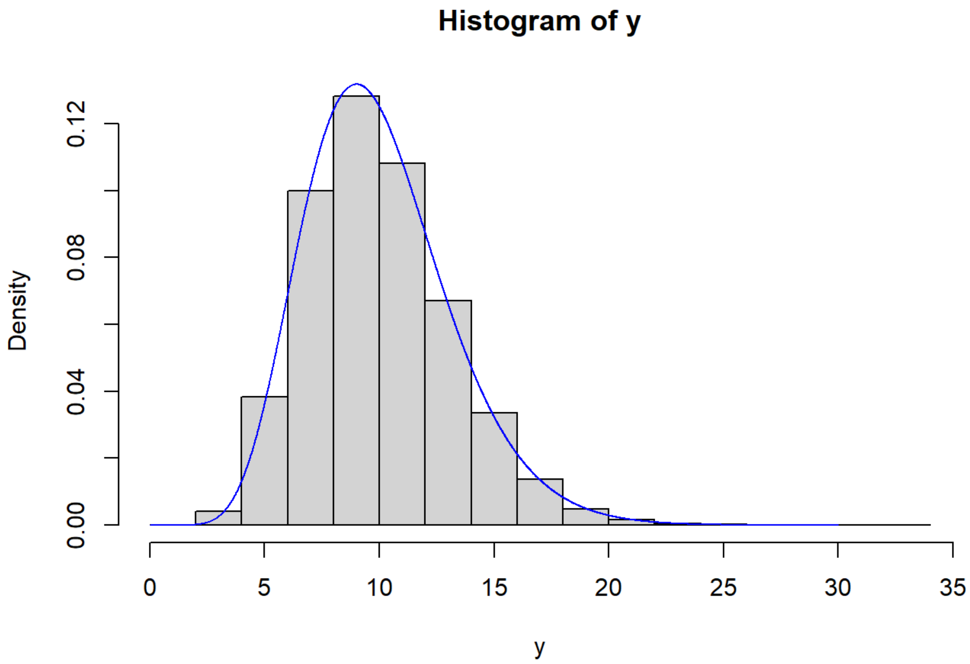
$\mathbb{E}[X_1 + X_2 + \ldots + X_n] = 5n$
$\mathbb{D}[X_1 + X_2 + \ldots + X_n] = 5n$

При $n = 2: \;\; Y = X_1 + X_2 \in \Gamma(10{,}1)$

$$f_Y(x) = \frac{1}{\Gamma(10)} x^{10-1} e^{-x} = \frac{1}{9!} x^9 e^{-x},\, x > 0$$

$\mathbb{E}[X_1 + X_2] = 10$
$\mathbb{D}[X_1 + X_2] = 10$

```
> x1 <- rgamma(50000, shape = 5, rate = 1)
> x2 <- rgamma(50000, shape = 5, rate = 1)
> y <- x1 + x2
> mean(y)
[1] 9.999332
> var(y)
[1] 9.965328
> hist(y, probability = TRUE)
> xCoord <- seq(0, 30, 0.01)
> lines(xCoord, dgamma(xCoord, shape = 10, rate = 1), col = "blue")
```

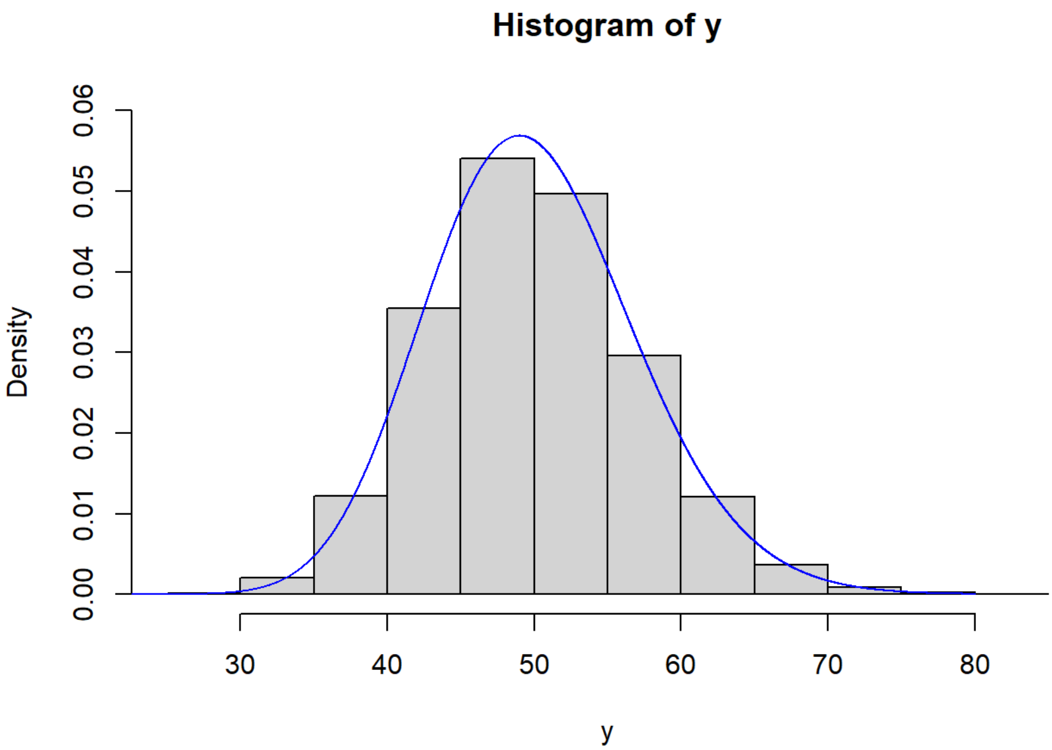


При $n = 10: \ Y = X_1 + \ldots + X_{10} \in \Gamma(50,\ 1)$

$$f_Y(x) = \frac{1}{\Gamma(50)} x^{50-1} e^{-x} = \frac{1}{49!} x^{49} e^{-x}, x > 0$$

$\mathbb{E}[X_1 + + \ldots + X_{10}] = 50$
$\mathbb{D}[X_1 + \ldots + X_{10}] = 50$

```
> y <- rgamma(50000, shape = 50, rate = 1)
> mean(y)
[1] 49.96254
> var(y)
[1] 50.45873
> hist(y, probability = TRUE, ylim = c(0, 0.06))
> xCoord <- seq(0, 80, 0.01)
> lines(xCoord, dgamma(xCoord, shape = 50, rate = 1), col = "blue")
```

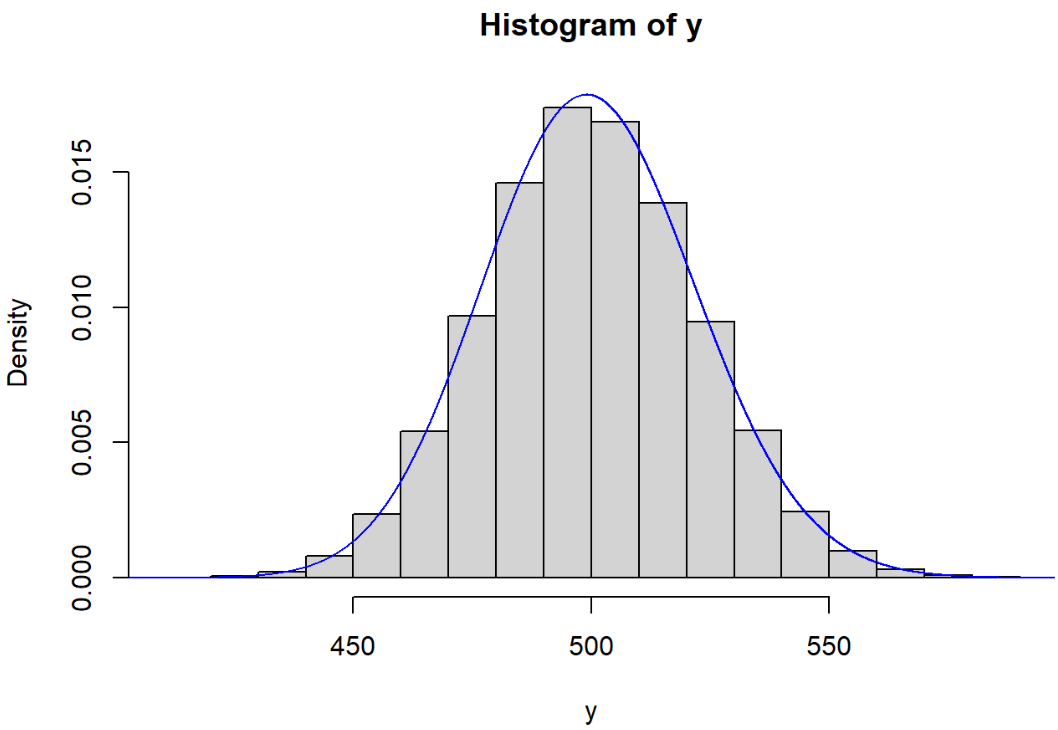


При $n = 100: \ Y = X_1 + \ldots + X_{100} \in \Gamma(500{,}1)$

$$f_Y(x) = \frac{1}{\Gamma(500)} x^{500-1} e^{-x} = \frac{1}{499!} x^{499} e^{-x}, x > 0$$

$\mathbb{E}[X_1 + \ldots + X_{100}] = 500$
$\mathbb{D}[X_1 + \ldots + X_{100}] = 500$

```
> y <- rgamma(50000, shape = 500, rate = 1)
> mean(y)
[1] 500.0568
> var(y)
[1] 502.6545
> hist(y, probability = TRUE)
> xCoord <- seq(0, 600, 0.01)
> lines(xCoord, dgamma(xCoord, shape = 500, rate = 1), col = "blue")
```



т.к. 100 е голямо можем да използваме централна гранична теорема (ЦГТ).

От централна гранична теорема (ЦГТ), ако X_i еднакво разпределени, независими и с крайна дисперсия за $i = 1, \ldots, n$, то при увеличаване на обема на извадката

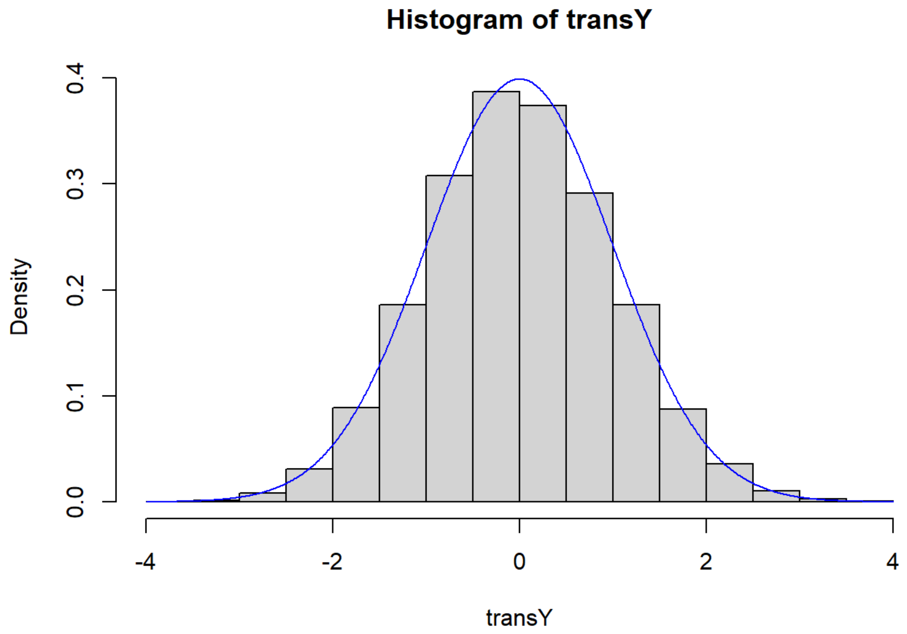
$$\frac{X_1 + X_2 + \ldots + X_n - n\mathbb{E}[X]}{\sqrt{n\mathbb{D}[X]}} \xrightarrow{d} N(0,1)$$

В случая

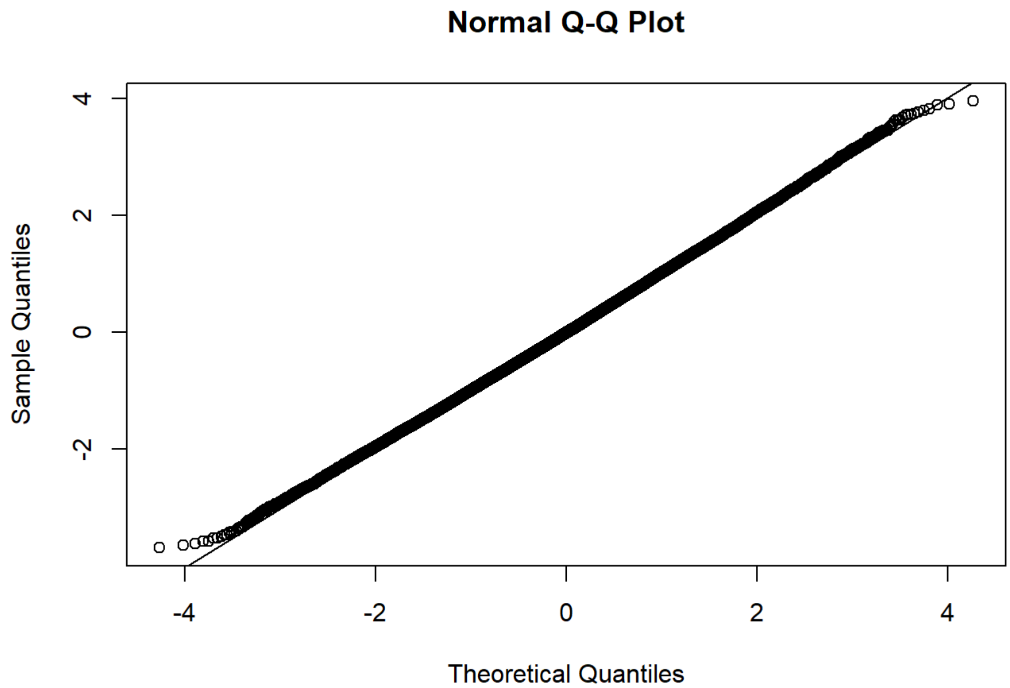
$$\mathbb{E}[X_i] = 5,\ \mathbb{D}[X_i] = 5,\ Y = X_1 + X_2 + \ldots + X_n$$

$$\frac{Y - 5n}{\sqrt{5n}} \xrightarrow{d} N(0,1)$$

```
> transY <- (y - 5*n) / sqrt(5*n)
> hist(transY, probability = TRUE)
> xCoord <- seq(-4, 4, 0.01)
> lines(xCoord, dnorm(xCoord, 0, 1), col = "blue")
```




```
> qqnorm(transY)
> qqline(transY)
```



e) $X_1, X_2, \ldots, X_n \in \mathcal{X}^2(5)$

Тъй като $\mathcal{X}^2 \equiv \Gamma(\frac{5}{2}, \frac{1}{2})$ и сума на независими гама случайни величини с един и същ втори параметър е отново гама със същия втори параметър, а първите параметри се сумират, то

$Y = X_1 + X_2 + \ldots + X_n \in \Gamma(\frac{5n}{2}, \frac{1}{2})$

$$f_X(x) = \frac{\left(\frac{1}{2}\right)}{\Gamma\left(\frac{5n}{2}\right)} x^{\frac{5}{2}n-1} e^{-\frac{x}{2}}, x > 0$$

$$\mathbb{E}[X_1 + X_2 + \ldots + X_n] = \frac{\frac{5n}{2}}{\left(\frac{1}{2}\right)^2} = 5n$$

$$\mathbb{D}[X_1 + X_2 + \ldots + X_n] = \frac{\frac{5n}{2}}{\left(\frac{1}{2}\right)^2} = 10n$$

При $n = 2$

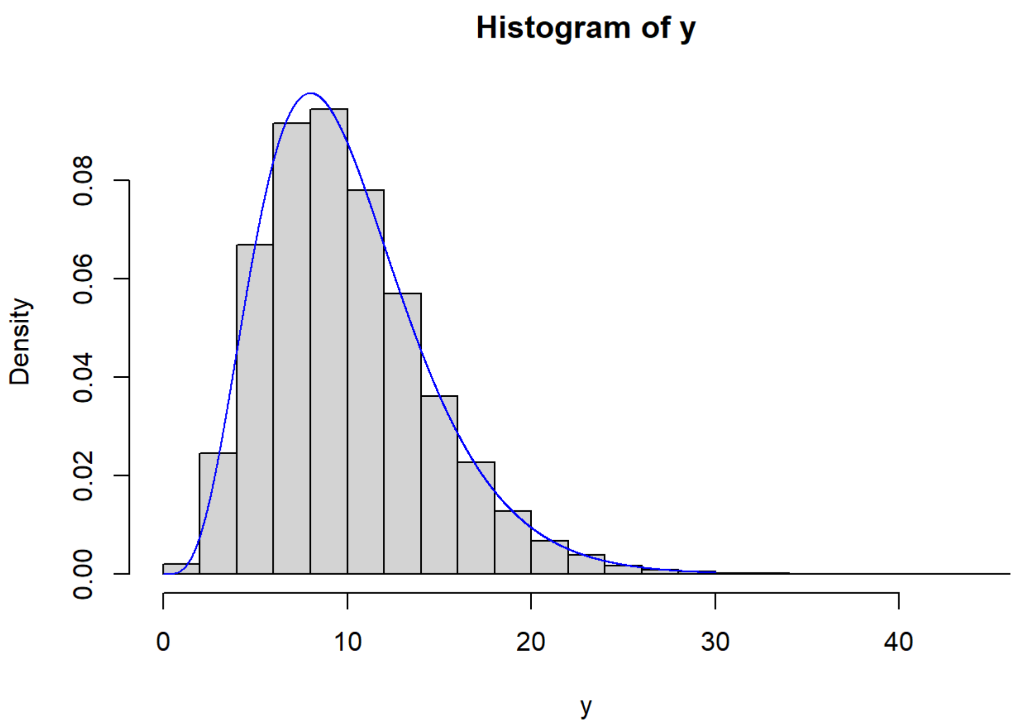
$$Y = X_1 + X_2 \in \Gamma\left(5, \frac{1}{2}\right)$$

$$f_Y(x) = \frac{\left(\frac{1}{2}\right)^5}{\Gamma(5)} x^{5-1} e^{-\frac{x}{2}} = \frac{1}{2^5 4!} x^4 e^{-\frac{x}{2}}, x > 0$$

$$\mathbb{E}[X_1 + X_2] = \frac{5}{\frac{1}{2}} = 10$$

$$\mathbb{D}[X_1 + X_2] = \frac{5}{\left(\frac{1}{2}\right)^2} = 20$$

```
> x1 <- rgamma(50000, shape = 5/2, rate = 1/2)
> x2 <- rgamma(50000, shape = 5/2, rate = 1/2)
> y <- x1 + x2
> mean(y)
[1] 9.989032
> var(y)
[1] 19.94206
> hist(y, probability = TRUE)
> xCoord <- seq(0, 30, 0.01)
> lines(xCoord, dgamma(xCoord, shape = 5, rate = 1/2), col = "blue")
```



При $n = 10$

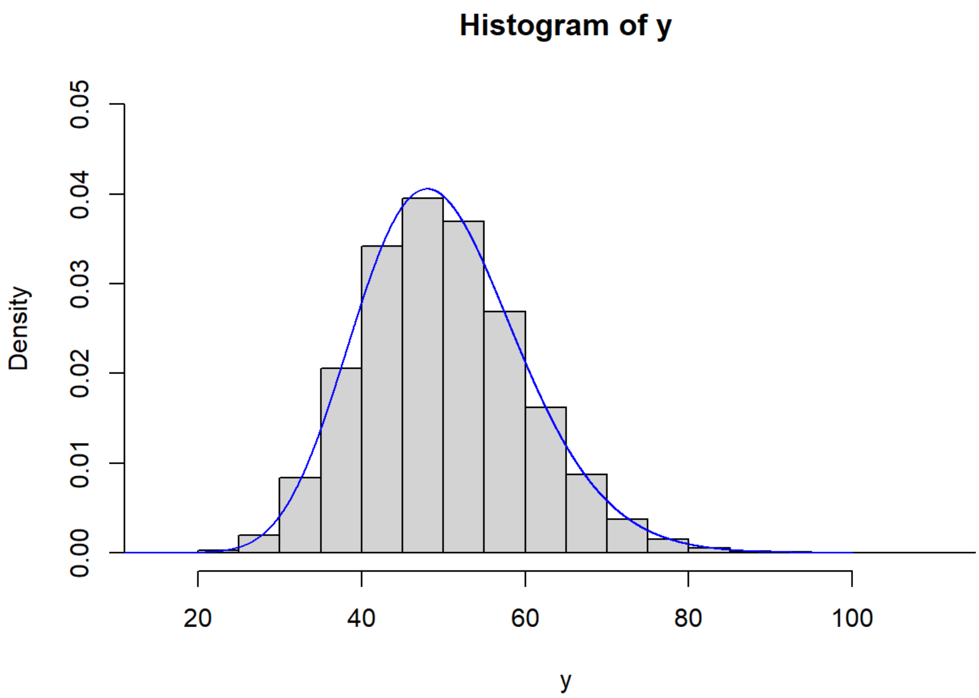
$$Y = X_1 + \ldots + X_{10} \in \Gamma\left(10 \times \frac{5}{2}, \frac{1}{2}\right) \equiv \Gamma\left(25, \frac{1}{2}\right)$$

$$f_Y(x) = \frac{\left(\frac{1}{2}\right)^{25}}{\Gamma(25)} x^{25-1} e^{-\frac{x}{2}} = \frac{\left(\frac{1}{2}\right)^{25}}{24!} x^{24} e^{-\frac{x}{2}}, x > 0$$

$$\mathbb{E}[X_1 + \ldots + X_{10}] = \frac{25}{\frac{1}{2}} = 50$$

$$\mathbb{D}[X_1 + \ldots + X_{10}] = \frac{25}{\left(\frac{1}{2}\right)^2} = 100$$

```
> y <- rgamma(50000, shape = 25, rate = 1/2)
> mean(y)
[1] 49.98172
> var(y)
[1] 99.28311
> hist(y, probability = TRUE, ylim = c(0, 0.05))
> xCoord <- seq(0, 100, 0.01)
> lines(xCoord, dgamma(xCoord, shape = 25, rate = 1/2), col = "blue")
```



При $n = 100$

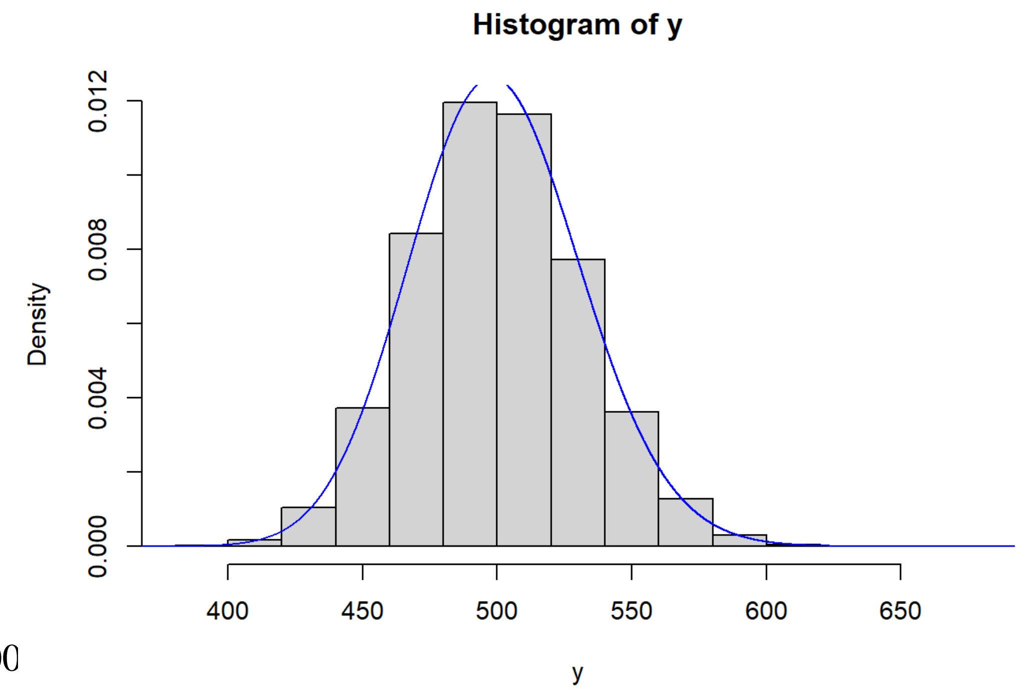
$$Y = X_1 + \ldots + X_{100} \in \Gamma\left(100 \times \frac{5}{2}, \frac{1}{2}\right) \equiv \Gamma\left(250, \frac{1}{2}\right)$$

$$f_Y(x) = \frac{\left(\frac{1}{2}\right)^{250}}{\Gamma(250)} x^{250-1} e^{-\frac{x}{2}} = \frac{1}{249\,!} x^{249} e^{-\frac{x}{2}}, x > 0$$

$$\mathbb{E}[X_1 + \ldots + X_{100}] = \frac{250}{\frac{1}{2}} = 500$$

$$\mathbb{D}[X_1 + \ldots + X_{100}] = \frac{250}{\left(\frac{1}{2}\right)^2} = 1000$$

```
> y <- rgamma(50000, shape = 250, rate = 1/2)
> mean(y)
[1] 500.0983
> var(y)
[1] 998.2995
> hist(y, probability = TRUE)
> xCoord <- seq(0, 800, 0.01)
> lines(xCoord, dgamma(xCoord, shape = 250, rate = 1/2), col = "blue")
```



т.к. 100 … ЦГТ).
От цент… исими и
с крайна дисперсия за $i = 1, \dots, n$, то при увеличаване на обема на извадката

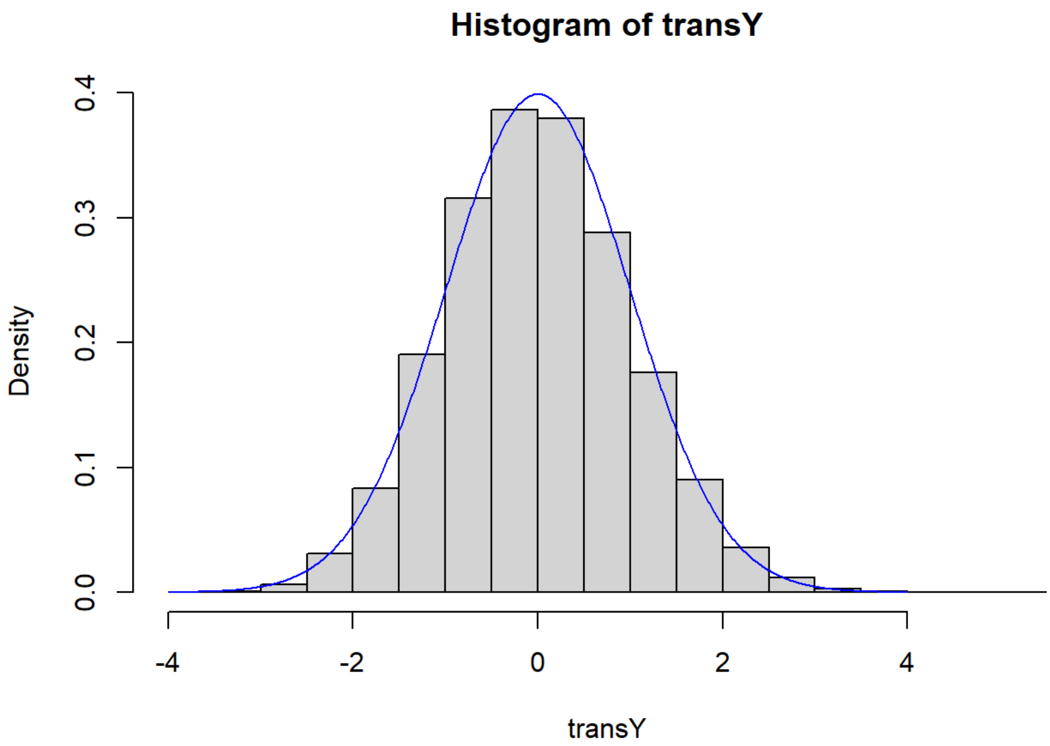
$$\frac{X_1 + X_2 + \dots + X_n - n\mathbb{E}[X]}{\sqrt{nD[X]}} \xrightarrow{d} N(0,1)$$

В случая

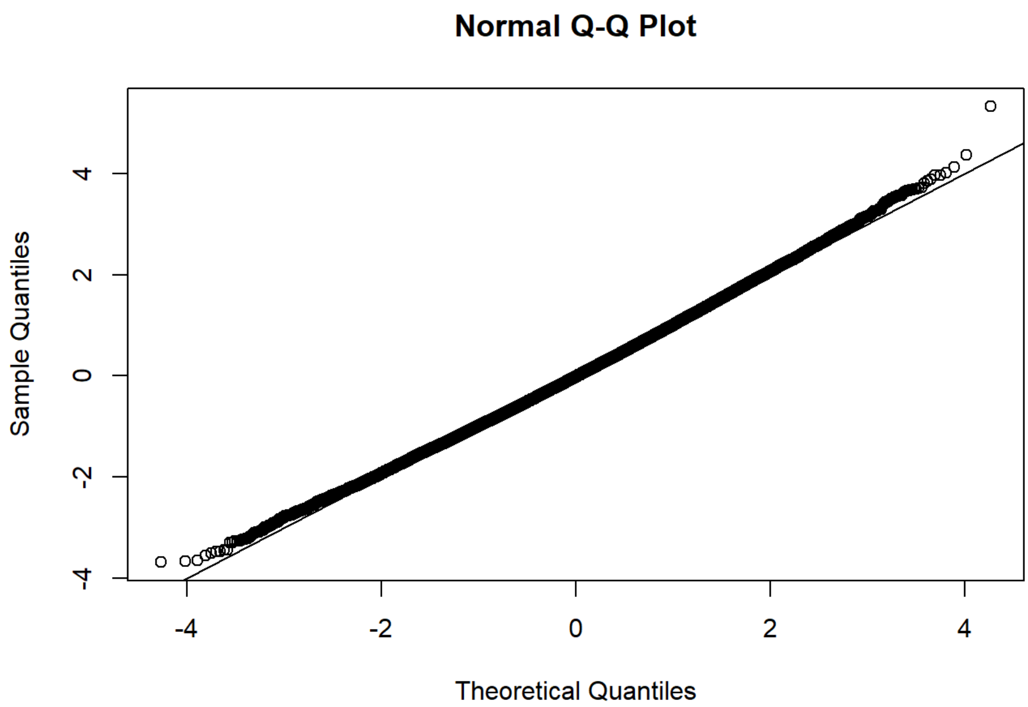
$$\mathbb{E}[X_i] = \frac{\frac{5}{2}}{\frac{1}{2}} = 5,\ \mathbb{D}[X_i] = \frac{\frac{5}{2}}{\left(\frac{1}{2}\right)^2} = 10$$

$$Y = X_1 + X_2 + \dots + X_n,\ \frac{Y - 5n}{\sqrt{10n}} \xrightarrow{d} N(0,1)$$

```
> transY <- (y - 5*n) / sqrt(10*n)
> hist(transY, probability = TRUE)
> xCoord <- seq(-4, 4, 0.01)
> lines(xCoord, dnorm(xCoord, 0, 1), col = "blue")
```



> qq
> qq



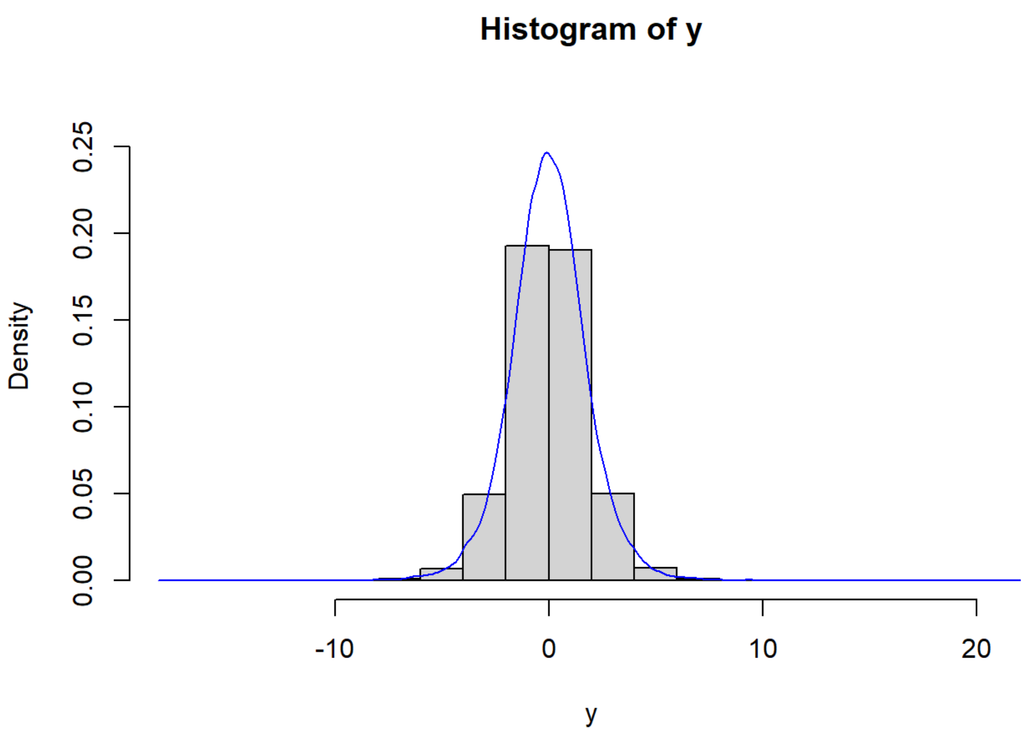
f) $X_1, X_2, \ldots, X_n \in t(5)$

$$\mathbb{E}[X_1 + X_2 + \ldots + X_n] = m\mathbb{E}[X_1] = n \times 0 = 0$$
$$\mathbb{D}[X_1 + X_2 + \ldots + X_n] = n\mathbb{D}[X_1] = n \times \frac{5}{5-2} = \frac{5}{3}n$$

При $n = 2: \;\; Y = X_1 + X_2$

$$\mathbb{E}[X_1 + X_2] = 2\mathbb{E}[X_1] = 2 \times 0 = 0$$
$$\mathbb{D}[X_1 + X_2] = 2\mathbb{D}[X_1] = 2 \times \frac{5}{3} = \frac{10}{3} \approx 3.33$$

```
> x1 <- rt(50000, df = 5)
> x2 <- rt(50000, df = 5)
> y <- x1 + x2
> mean(y)
[1] 0.0003633749
> var(y)
[1] 3.332129
> hist(y, probability = TRUE, ylim = c(0,0.27))
> lines(density(y, bw = "SJ"), col = "blue")
```

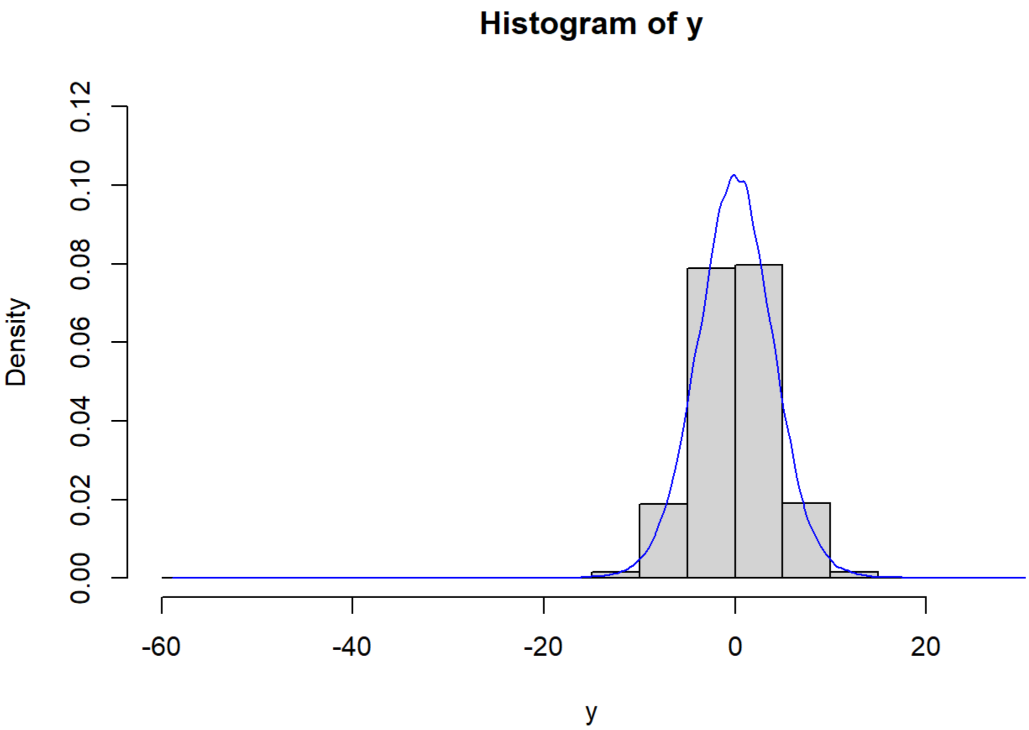


При $n = 10: \;\; Y = X_1 + \ldots + X_{10}$

$$\mathbb{E}[X_1 + \ldots + X_{10}] = 10\mathbb{E}[X_1] = 10 \times 0 = 0$$
$$\mathbb{D}[X_1 + \ldots + X_{10}] = 10D[X_1] = 10 \times \frac{5}{3} = \frac{50}{3} \approx 16.67$$

```
> y <- 0
```

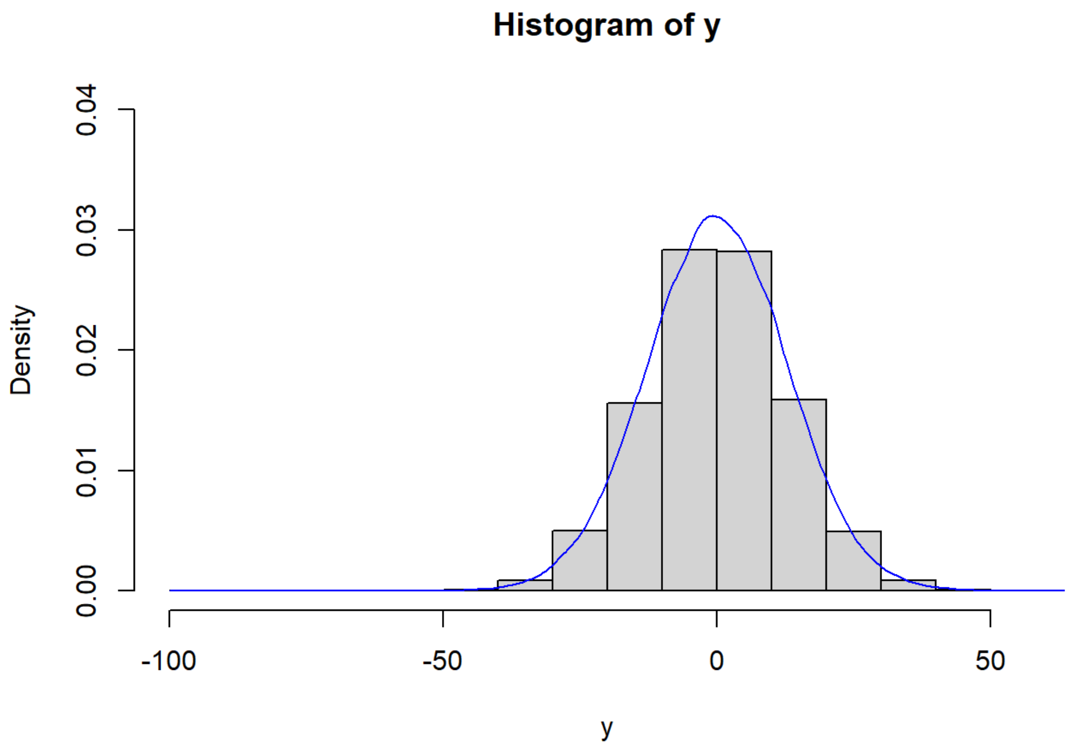
```
> for(i in 1:(10^5)){
+   y[i] <- sum(rt(10, df = 5))
+ }
> mean(y)
[1] 0.008303705
> var(y)
[1] 16.56679
> hist(y, probability = TRUE, ylim = c(0, 0.12))
> lines(density(y, bw = "SJ"), col = "blue")
```



При $n = 100: \; Y = X_1 + .. + X_{100}$

$$\mathbb{E}[X_1 + \ldots + X_{100}] = 100\mathbb{E}[X_1] = 100 \times 0 = 0$$
$$\mathbb{D}[X_1 + \ldots + X_{100}] = 10\mathbb{D}[X_1] = 100 \times \frac{5}{3} = \frac{500}{3} \approx 166.67$$

```
> y <- 0
> for(i in 1:(10^5)){
+   y[i] <- sum(rt(100, df = 5))
+ }
> mean(y)
[1] 0.01873994
> var(y)
[1] 165.7692
> hist(y, probability = TRUE, ylim = c(0, 0.04))
> lines(density(y, bw = "SJ"), col = "blue")
```



т.к. [illegible] Т).
От централна гранична теорема (ЦГТ), ако X_i еднакво разпределени, независими и с крайна дисперсия за $i = 1, \ldots, n$, то при увеличаване на обема на извадката

$$\frac{X_1 + X_2 + \ldots + X_n - n\mathbb{E}[X]}{\sqrt{n\mathbb{D}[X]}} \xrightarrow{d} N(0,1)$$

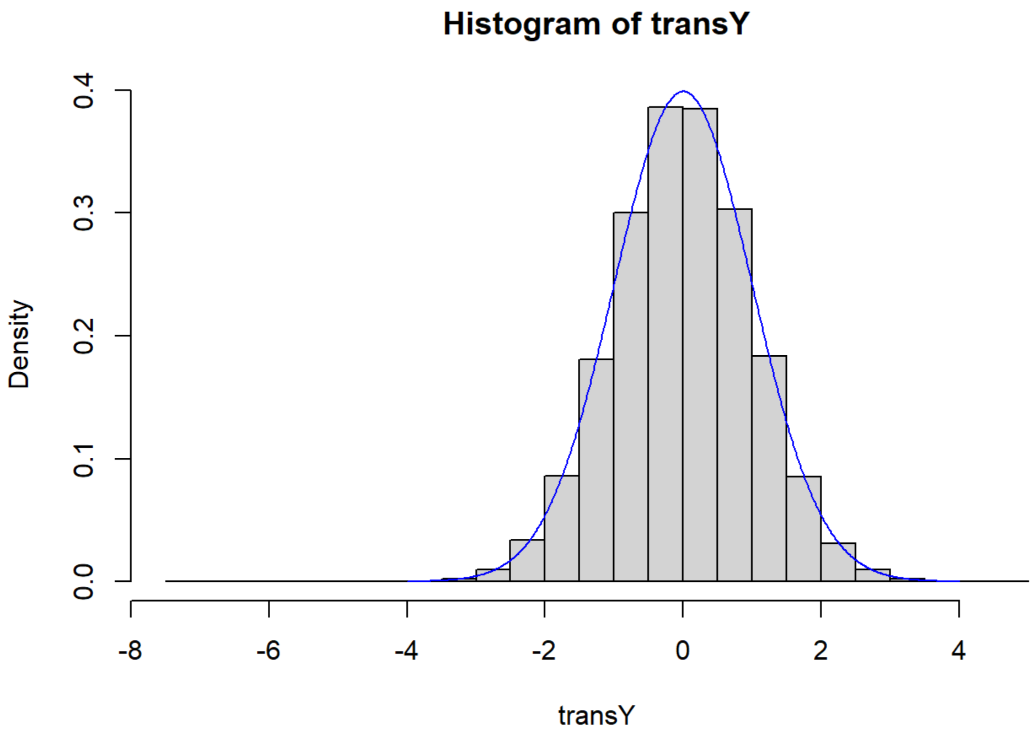
В случая

$\mathbb{E}[X_i] = 0,\ \mathbb{D}[X_i] = \dfrac{5}{3}$

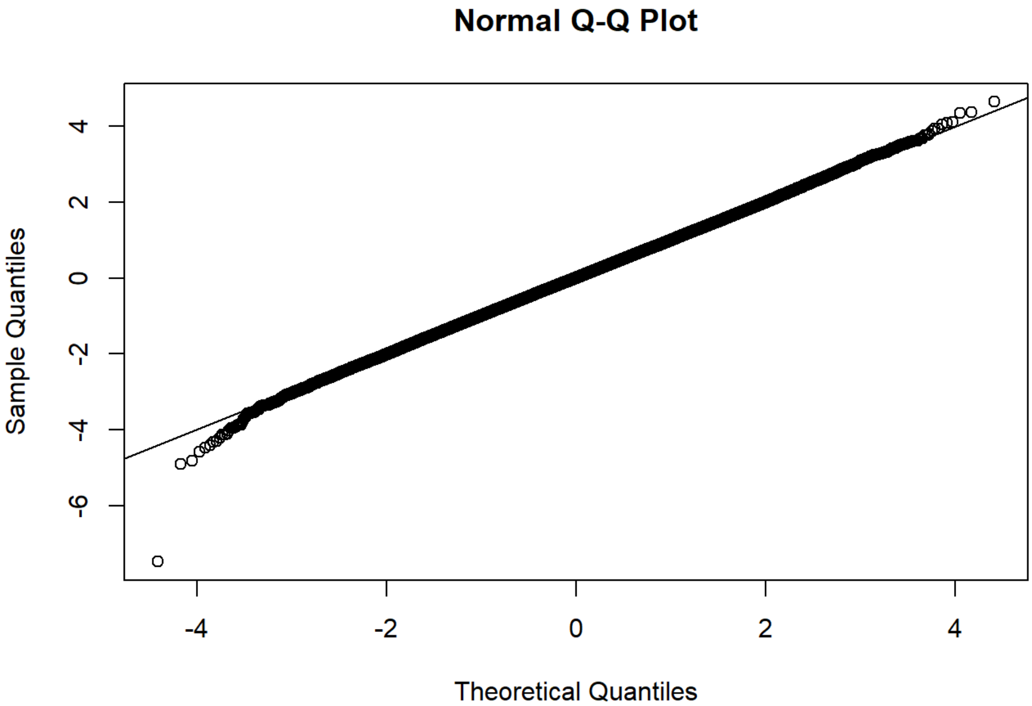
$Y = X_1 + X_2 + \ldots + X_n$

$$\frac{Y - n\mathbb{E}[X]}{\sqrt{n\mathbb{D}[X]}} \xrightarrow{d} N(0,1),\ \frac{Y - 0 \times n}{\sqrt{\frac{5}{3}n}} \xrightarrow{d} N(0,1)$$

```
> transY <- (y - 0*n) / sqrt((5*n)/3)
> hist(transY, probability = TRUE)
> xCoord <- seq(-4, 4, 0.01)
> lines(xCoord, dnorm(xCoord, 0, 1), col = "blue")
```



> qqnorm(transY)
> qqline(transY)



g) X е смес от две разпределения $N(1,2)$ и $N(5,2)$ с вероятност за първото $p = 0.4$

$$I_A \in Bernoulli(0.4)$$
$$X = I_A \times N(1,2) + I_{\overline{A}} \times N(5,2)$$
$$X \overset{d}{=} X_1 \overset{d}{=} X_2 \overset{d}{=} \dots \overset{d}{=} X_n$$

и тези случайни величини са независими

$$\mathbb{E}[X_1 + X_2 + \ldots + X_n] = n\mathbb{E}[X] = n(p \times 1 + (1-p) \times 5)$$
$$= n(0.4 + (1 - 0.4) \times 5) = 3 \times 4n$$

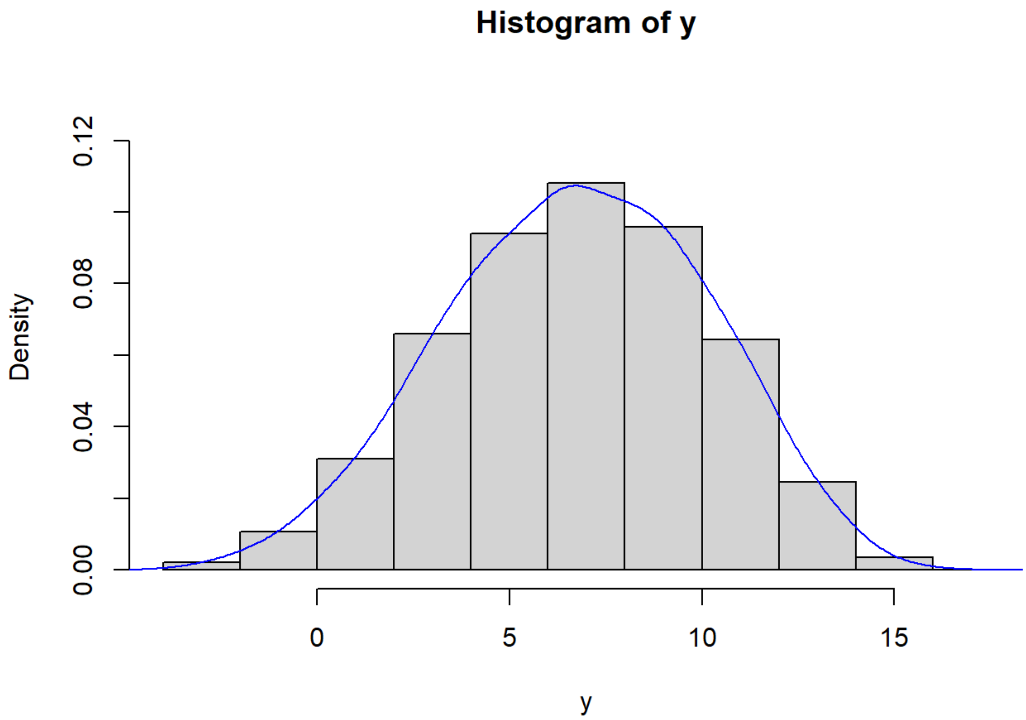
$$\begin{aligned}
\mathbb{E}[X_1^2] &= \mathbb{E}[[I_A N(1,2) + I_{\overline{A}} N(5,2)]^2 = \\
&= \mathbb{E}[[I_A N(1,2)]^2 + 2I_A N(1,2) I_{\overline{A}} N(5,2) + [I_{\overline{A}} N(5,2)]^2] = \\
&= \mathbb{E}[I_A [N(1,2)]^2 + 2I_A N(1,2) I_{\overline{A}} N(5,2) + I_{\overline{A}} [N(5,2)]^2] = \\
&= \mathbb{E}[I_A]\mathbb{E}[[N(1,2)]^2] + 2\mathbb{E}[0 \times N(1,2) \times N(5,2)] + \mathbb{E}[I_{\overline{A}}]\mathbb{E}[[N(5,2)]^2 = \\
&= \mathbb{E}[I_A]\mathbb{E}[[N(1,2)]^2] + 2\mathbb{E}[0] + \mathbb{E}[0] + \mathbb{E}[I_{\overline{A}}]\mathbb{E}[[N(5,2)]^2] = \\
&= [(2 + 1^2) + 0 + (1-p)(2 + 5^2) = \\
&= 3p + 27(1-p) = 27 - 24p = 27 - 24 \times 0.4 = 17.4
\end{aligned}$$

$$\mathbb{D}[X_1] = \mathbb{E}[X_1^2] - [\mathbb{E}[X]]^2 = 17.4 - 3.4^2 = 5.842$$
$$\mathbb{D}[X - 1 + X_2 + \ldots + X_n] = n\mathbb{D}[X_1] = 5.842n$$

При $n = 2: \;\; Y = X_1 + X_2$

$$\mathbb{E}[X_1 + X_2] = 2\mathbb{E}[X_1] = 2 \times 3.4 = 6.8$$
$$\mathbb{D}[X_1 + X_2] = 2\mathbb{D}[X_1] = 2 \times 5.842 = 11.684$$

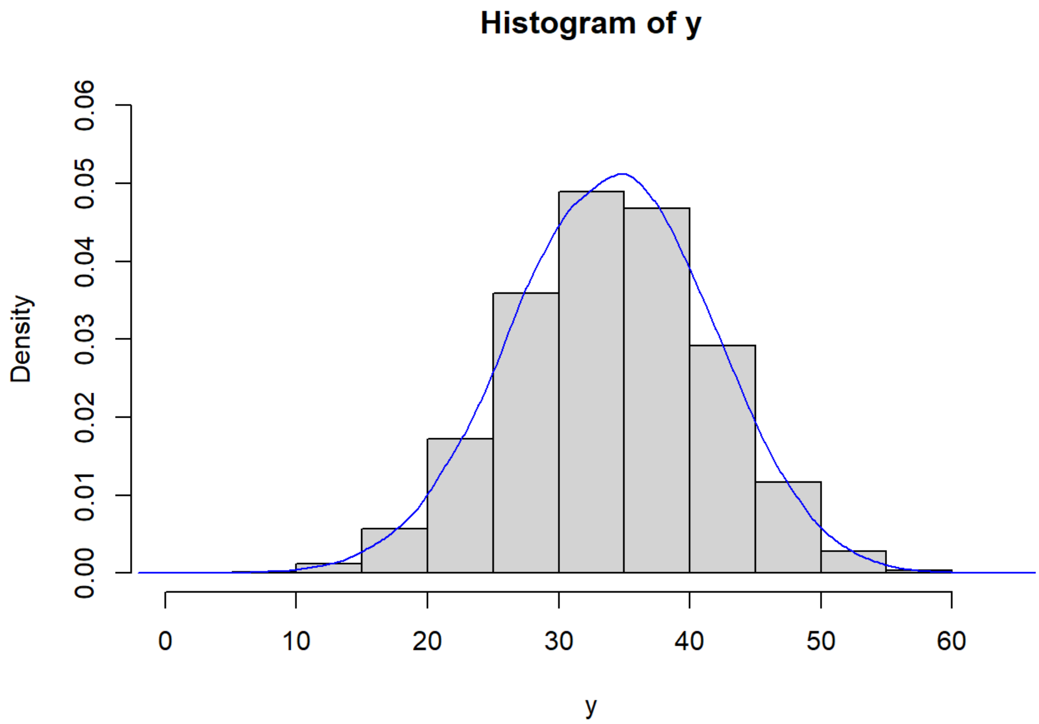
```
> i <- rbinom(5000, size = 1, prob = 0.4)
> n1 <- rnorm(5000, mean = 1, sd = sqrt(2))
> n2 <- rnorm(5000, mean = 5, sd = sqrt(2))
> x1 <- i * n1 + (1 - i) * n2
> i <- rbinom(5000, size = 1, prob = 0.4)
> n1 <- rnorm(5000, mean = 1, sd = sqrt(2))
> n2 <- rnorm(5000, mean = 5, sd = sqrt(2))
> x2 <- i * n1 + (1 - i) * n2
> y <- x1 + x2
> mean(y)
[1] 6.745864
> var(y)
[1] 11.59401
> hist(y, probability = TRUE, ylim = c(0,0.13))
> lines(density(y, bw = "SJ"), col = "blue")
```



При n

$$\mathbb{E}[X_1 + \ldots + X_{10}] = 10\mathbb{E}[X_1] = 10 \times 3.4 = 34$$
$$\mathbb{D}[X_1 + \ldots + X_{10}] = 10\mathbb{D}[X_1] = 10 \times 5.842 = 58.42$$

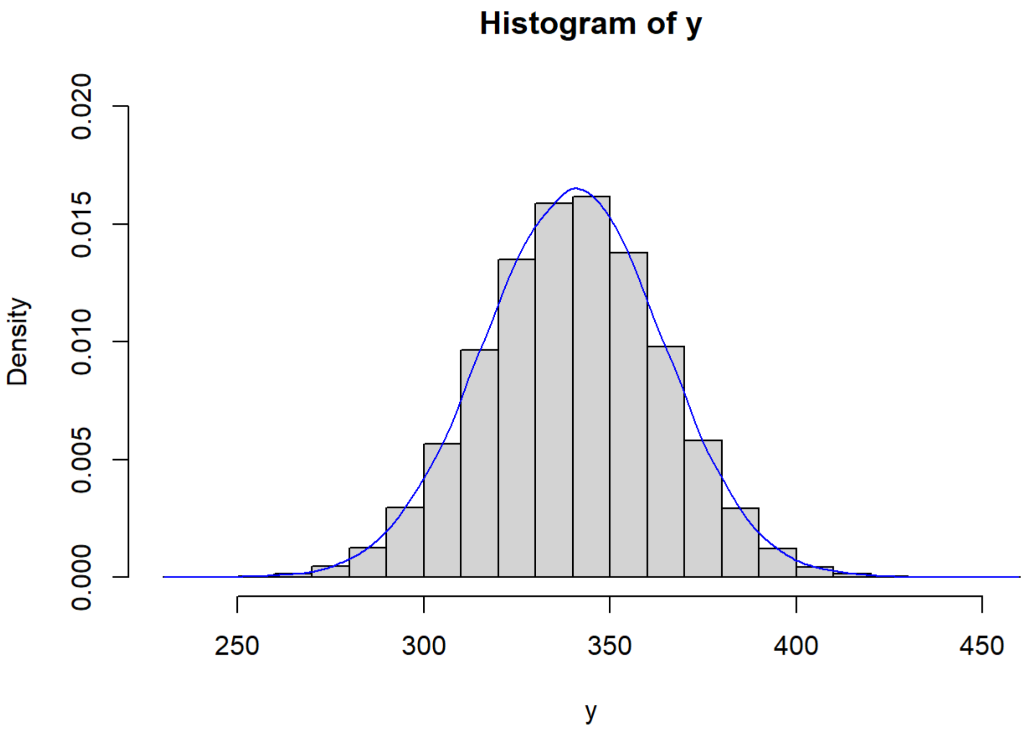
```
> y <- 0
> for(i in 1:(10^5)){
+   b <- rbinom(10, size = 1, prob = 0.4)
+   n1 <- rnorm(10, mean = 1, sd = sqrt(2))
+   n2 <- rnorm(10, mean = 5, sd = sqrt(2))
+   x <- b * n1 + (1 - b) * n2
+   y[i] <- sum(x)
+ }
> mean(y)
[1] 34.00682
> var(y)
[1] 58.62883
> hist(y, probability = TRUE, ylim = c(0, 0.06))
> lines(density(y, bw = "SJ"), col = "blue")
```



При $n = 100: \;\; Y = X_1 + \ldots + X_{100}$

$$\mathbb{E}[X_1 + \ldots + X_{100}] = 10\mathbb{E}[X_1] = 100 \times 3.4 = 340$$
$$\mathbb{D}[X_1 + \ldots + X_{100}] = 100\mathbb{D}[X_1] = 100 \times 5.842 = 584.2$$

```
> y <- 0
> for(i in 1:(10^5)){
+   b <- rbinom(100, size = 1, prob = 0.4)
+   n1 <- rnorm(100, mean = 1, sd = sqrt(2))
+   n2 <- rnorm(100, mean = 5, sd = sqrt(2))
+   x <- b * n1 + (1 - b) * n2
+   y[i] <- sum(x)
+ }
> mean(y)
[1] 340.0929
> var(y)
[1] 580.7164
> hist(y, probability = TRUE, ylim = c(0, 0.02))
> lines(density(y, bw = "SJ"), col = "blue")
```



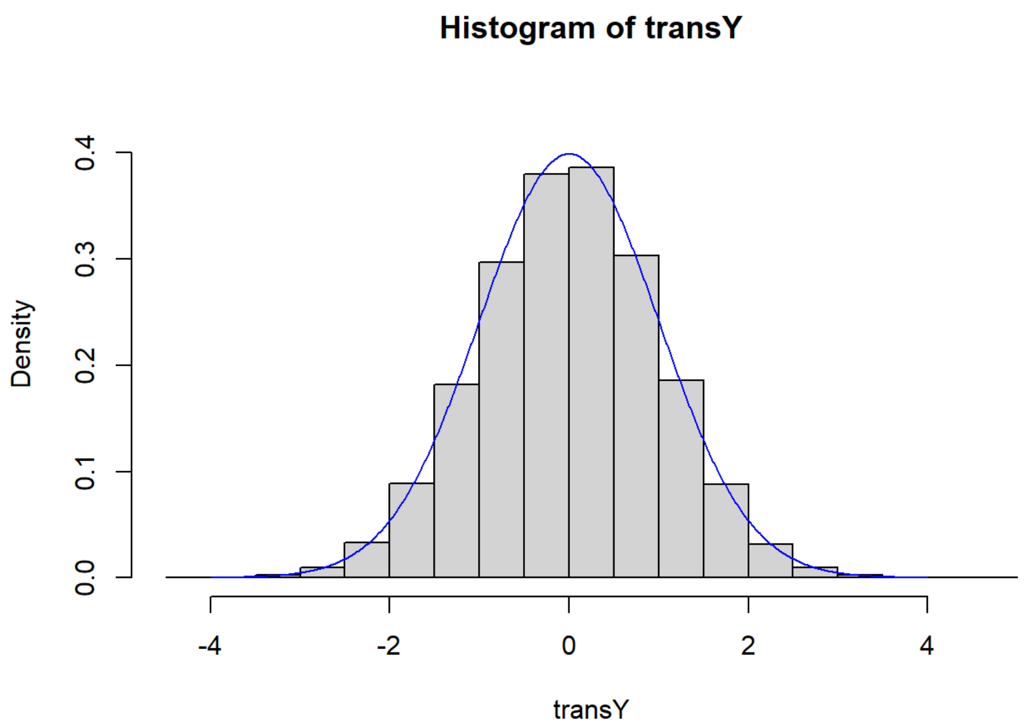
т.к. $100 >> 30$ е голямо можем да използваме централна гранична теорема (ЦГТ). От централна гранична теорема (ЦГТ), ако X_i еднакво разпределени, независими и с крайна дисперсия за $i = 1, \dots, n$, то при увеличаване на обема на извадката

$$\frac{X_1 + X_2 + \dots + X_n - n\mathbb{E}[X]}{\sqrt{n\mathbb{D}[X]}} \xrightarrow{d} N(0,1)$$

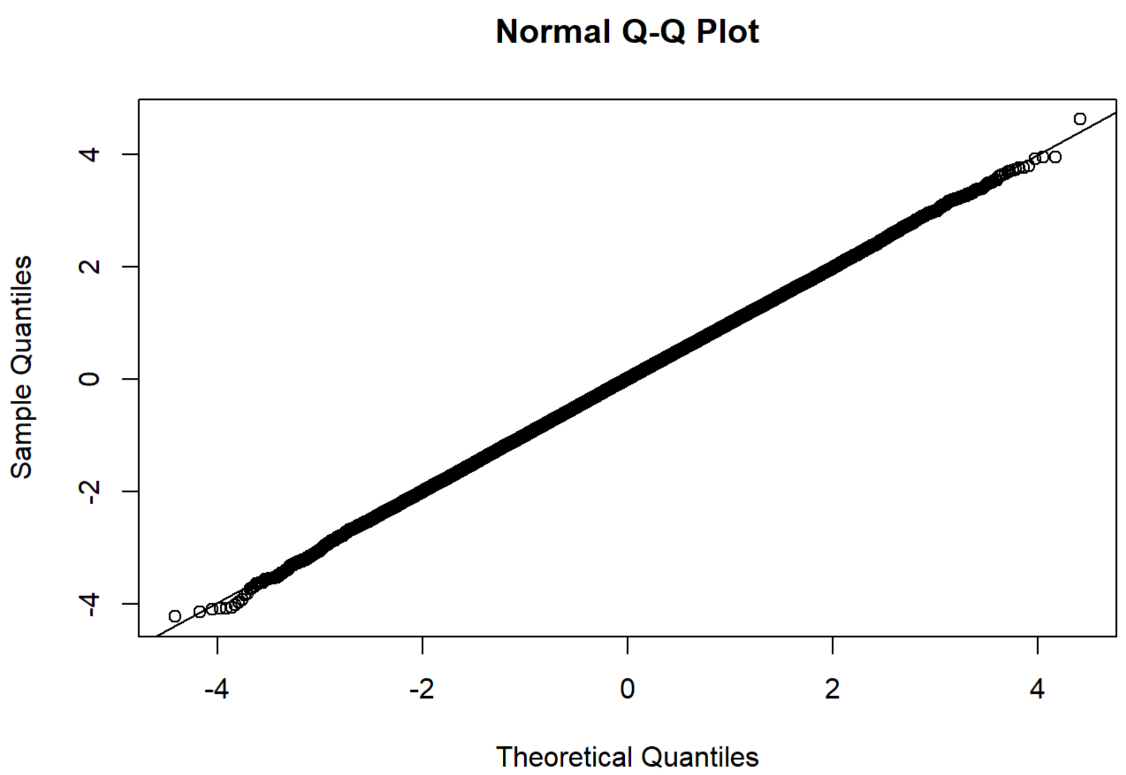
В случая

$$\mathbb{E}[X_i] = 0,\ \mathbb{D}[X_i] = \frac{5}{3},\ Y = X_1 + X_2 + \dots + X_n,$$
$$\frac{Y - n\mathbb{E}[X]}{\sqrt{n\mathbb{D}[X]}} \xrightarrow{d} N(0,1),\ \frac{Y - 3.4n}{\sqrt{5{,}842n}} \xrightarrow{d} N(0,1)$$

```
> transY <- (y - 3.4 * n) / sqrt(5.842 * n)
> hist(transY, probability = TRUE, ylim = c(0, 0.44))
> xCoord <- seq(-4, 4, 0.01)
> lines(xCoord, dnorm(xCoord, 0, 1), col = "blue")
```




```
> qqnorm(transY)
> qqline(transY)
```



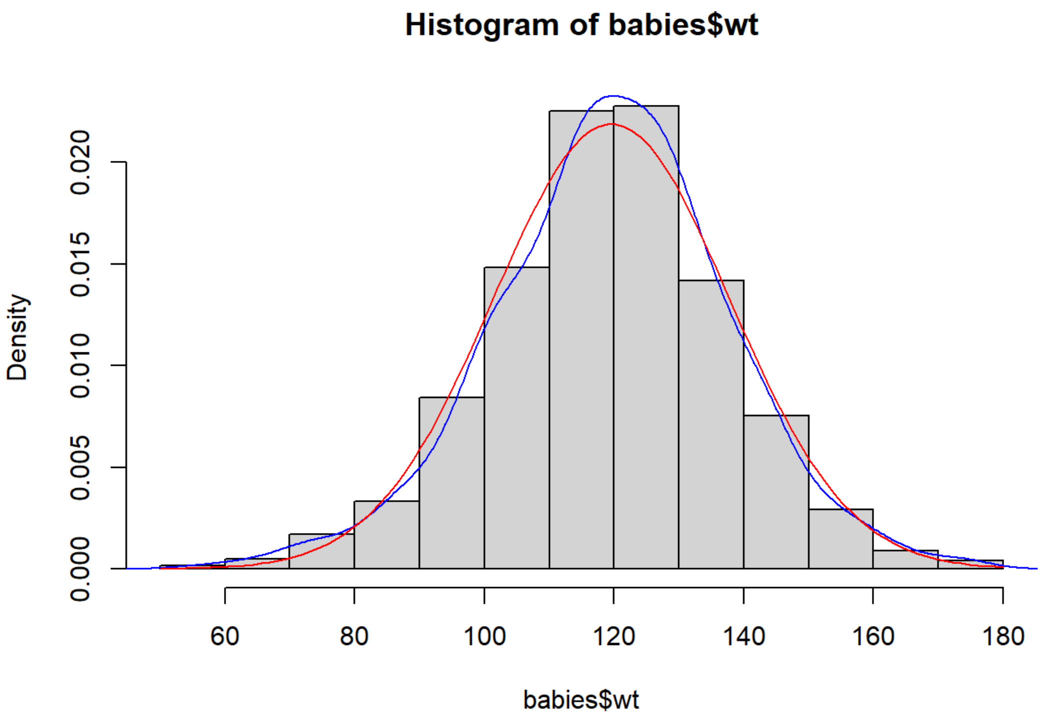
Задача 3

Определете дали са нормално разпределени наблюденията:

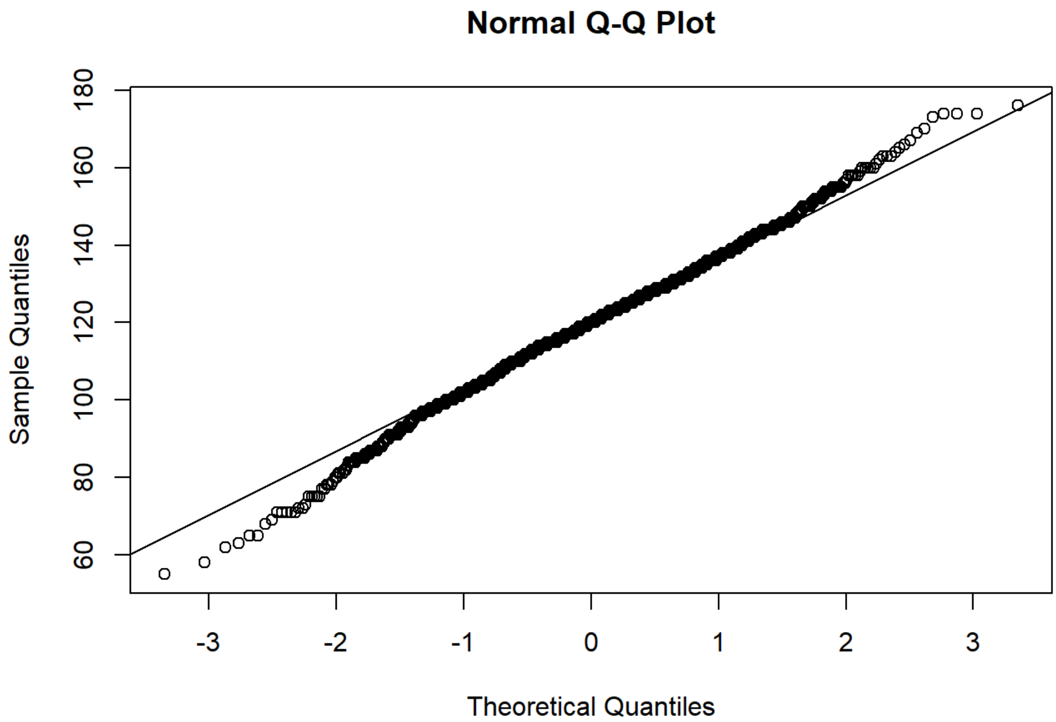
а) теглото на бебетата дадени в babies от пакета UsingR;

```
> library(UsingR)
```

```
> hist(babies$wt, probability = TRUE)
> lines(density(babies$wt, bw = "SJ"), col = "blue")
> x = babies$wt
> curve(dnorm(x, mean(babies$wt), sd(babies$wt)), add = TRUE, col = "red")
```




```
> qqnorm(babies$wt)
> qqline(babies$wt)
```

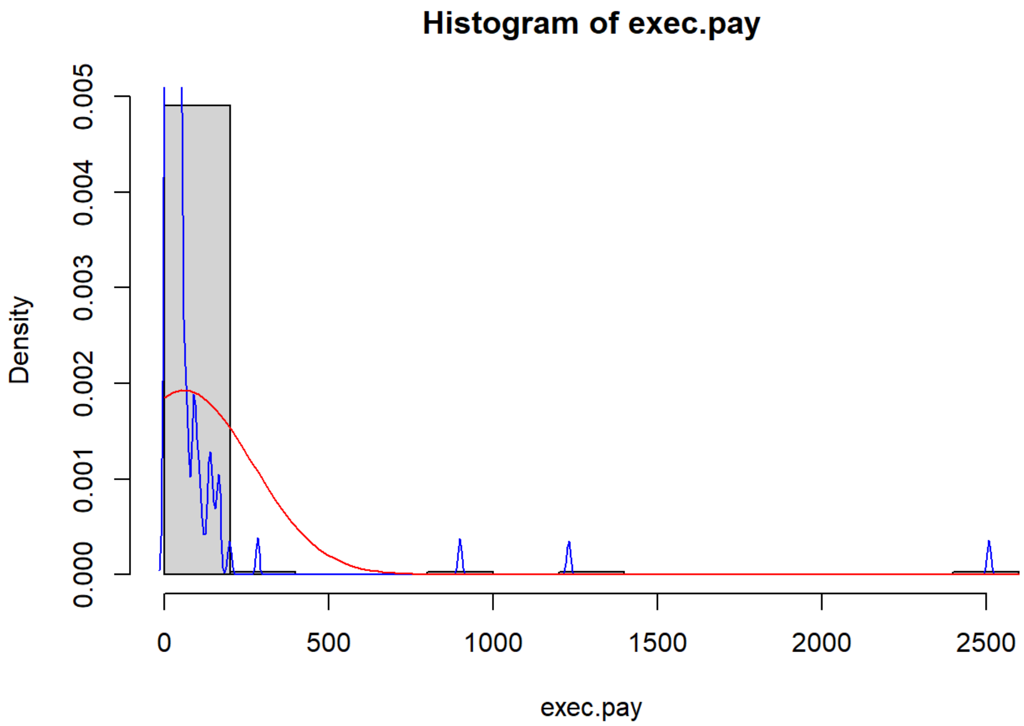



```
> shapiro.test(babies$wt)

	Shapiro-Wilk normality test

data:  babies$wt
W = 0.99559, p-value = 0.001192
```

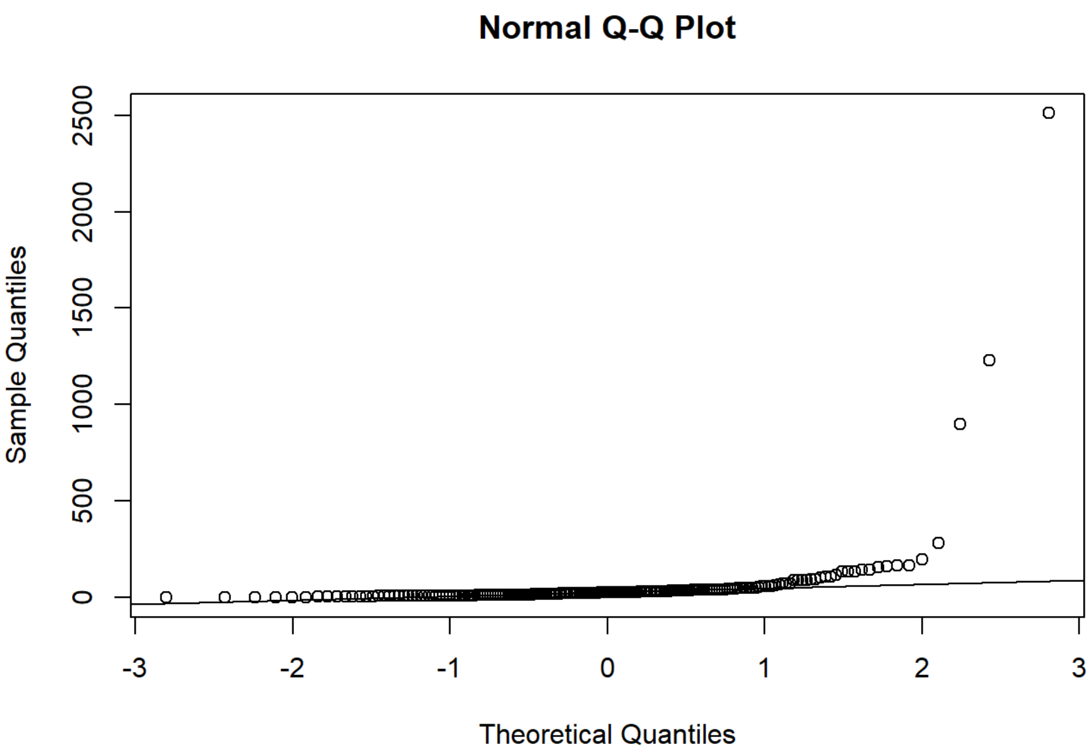
Имаме $p-value = 0.001192 < 0.05 = \alpha$, което означава, че разпределението ни е значително по-различно от нормалното разпределение, т.е. нямаме основание да допуснем, че данните ни са нормално разпределени.

b) exac.pay от пакета UsingR;

```
> hist(exec.pay, probability = TRUE)
> lines(density(exec.pay, bw = "SJ"), col = "blue")
> x = exec.pay
> curve(dnorm(x, mean(exec.pay), sd(exec.pay)), add = TRUE, col = "red")
```



> qqnorm(exec.pay)
> qqline(exec.pay)



> shapiro.test(exec.pay)

Shapiro-Wilk normality test

data: exec.pay
W = 0.19352, p-value < 2.2e-16

Имаме $p-value < 2.2e-16 < 0.05 = \alpha$, което означава, че разпределението ни е значително по-различно от нормалното разпределение, т.е. нямаме основание да допуснем, че данните ни са нормално разпределени.

Задача 4
Размерът на пъпешите е нормално разпределена сл.в. с очакване 25см. и дисперсия 36. Пъпешите по-малки от 20см. са трето качество, а останалите се разделят на две равни по брой групи, като по-големите са първо качество, а по-малките второ. Каква част от пъпешите са трето качество. Колко голям трябва да е пъпеша за да бъде първо качество.

Решение:

$X \in N(25, 6^2)$
$(X < 20)$ – 3-то качество

$$\frac{\#(X \geq 20)}{2} = \#(\text{1-во качество}) = \#(\text{2-ро качество})$$

```
> pnorm(q = 20, mean = 25, sd = 6)
[1] 0.2023284
```

$\mathbb{P}(X < 20) = 0.2023284$, т.е. 20.23% от пъпешите са 3-то качество.

```
> qnorm(p = 0.2023284 + (1 - 0.2023284)/2, mean = 25, sd = 6)
[1] 26.53817
```

$$q = F_{\overleftarrow{X}}\left(0.2023284 + \frac{1 - 0.2023284}{2}\right) = 26.53817,$$

т.е. за да бъде 1-во качество пъпешът трябва да бъде над 26.54см.

Задача 5

Нека сл.в. X е гамма разпределена с параметри 2 и 0.5. Определете:

a) $\mathbb{P}(X < 1)$;

```
> pgamma(1, shape = 2, rate = 0.5)
[1] 0.09020401
```

b) $\mathbb{P}(X > 2)$;

```
> pgamma(2, shape = 2, rate = 0.5, lower.tail = FALSE)
[1] 0.7357589
```

c) c, така че $\mathbb{P}(X > c) = 0.35$;

```
> qgamma(0.35, shape = 2, rate = 0.5, lower.tail = FALSE)
[1] 4.437689
```

d) Q_1, M, Q_3

```
> qgamma(0.25, shape = 2, rate = 0.5)
[1] 1.922558
> qgamma(0.5, shape = 2, rate = 0.5)
[1] 3.356694
> qgamma(0.75, shape = 2, rate = 0.5)
[1] 5.385269
```

Sources

[1] Monika Petkova's notes on R programming language @ FMI, Sofia University