Exercise 6

Applied Statistics 2019, IT University of Copenhagen

1. Sum of Bernoulli Distributed Variables (T)

Let X and Y be two independent random variables where $X \sim \text{Ber}(p)$ and $Y \sim \text{Ber}(q)$. Let $Z = X + Y$. Investigate how Z is distributed by deriving the probability mass function for Z.

Solution: Since X and Y are two independent discrete random variables, the probability mass function p_Z of $Z = X + Y$ is given by

$$p_Z(c) = \sum_j p_X(c - b_j) p_Y(b_j)$$

cf. page 152 (Dekking et al.), where the sum runs over all possible values b_j of Y. Since Y follows a Bernoulli distribution, we know that the possible values of b_j are 0 and 1, we can also express this as

$$\begin{aligned} p_Z(c) &= p_X(c)p_Y(0) + p_X(c-1)p_Y(1) && (1)\\ &= p_X(c)(1-q) + p_X(c-1)q. && (2) \end{aligned}$$

Since Z is the sum of 2 Bernoulli distributed random variables, the possible values of Z are $\{0, 1, 2\}$. We also know, the probability mass function p_X of X is given by $p_X(0) = P(X = 1) = p$ and $p_X(1) = P(X = 0) = 1 - p$ cf. page 45 (Dekking et al.). From this, we can now find $p_Z(c)$ for all possible values of c:

$$\begin{aligned} p_Z(0) &= p_X(0)(1-q) + p_X(-1)q = (1-p)(1-q) && (3)\\ p_Z(1) &= p_X(1)(1-q) + p_X(0)q = p(1-q) + (1-p)q && (4)\\ p_Z(2) &= p_X(2)(1-q) + p_X(1)q = pq, && (5) \end{aligned}$$

which can also be expressed as

$$p_Z(c) = P(Z = c) = \begin{cases} (1-q)(1-p) & \text{if } c = 0,\\ p(1-q) + (1-p)q & \text{if } c = 1,\\ pq & \text{if } c = 2. \end{cases}$$

2. Casino La Cella Fortuna (T)

The casino La bella Fortuna is for sale and you think you might want to buy it, but you want to know how much money you are going to make. All the present owner can tell you is that the roulette game Red or Black is played about 1000 times a night, 365 days a year. Each time it is played you have probability 19/37 of winning the player's bet of 1 EUR and probability 18/37 of having to pay the player 1 EUR. Explain in detail why the law of large numbers can be used to determine the income of the casino, and determine how much it is.

The Law of Large Numbers: the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become even closer as more trials are performed.

Solution: We cannot know for certain, what the income of the casino will be. In fact, all the roulette games could be lost and the casino would have no income. However, the law of large numbers comes to our rescue. In broad terms, the law of large numbers says that for n independent random variables with the same expectation μ and variance σ^2, there is 0 probability that the absolute difference between the average of the random variables $\bar{X}_n$ and the expectation will be larger than 0 when $n \to \infty$ cf. page 185 (Dekking et al.).

In our case, we have a roulette game with 2 outcomes: either we (the casino) win or the player wins with probabilities $\frac{19}{37}$ and $\frac{18}{37}$ respectively, where we either acquire or lose a dollar. We can let X_i be the Bernoulli distributed random variable of the ith roulette game. The expected value of X_i is

$$\mathrm{E}[X_i] = 1 \cdot \frac{19}{37} + (-1) \cdot \frac{18}{37} = \frac{1}{37}$$

cf. page 90 (Dekking et al.), and since X_i are identically distributed, this will be the same for all values of i. That is, we are expected to win $\frac{1}{37}$ EUR from a roulette game. We also know that each roulette game is independent from each other. The law of large numbers now tells us that there's a 0 probability that the average of these random variables will differ from the expected value, as the number of random variables approaches infinity - or rather: as the number of random variables increases, the average of them will approach the expected value. Since the expected value tells us that we earn $\frac{1}{37}$ EUR from a roulette game, the law of large numbers tells us that if we play a sufficiently large number of roulette games, we will earn just about the same per game as the expected value.

We were informed that the casino plays 1000 roulette games each day, and with an expected value of $\frac{1}{37}$ per game, we would earn

```
1000*365*(1/37)
```

```
## [1] 9864.865
```

EUR every year. Since 365000 is a fairly large number in this context, the law of large numbers tells us that we're safe to assume this is the amount of money we will make in a year. To test this hypothesis, we can simulate the roulette games and find our income:

```
roulette_avg <- function(n_games) {
  samp <- sample(c(-1, 1), n_games, replace=TRUE, prob=c(18/37, 19/37))
  return(sum(samp))
}
roulette_avg(365000)
```

```
## [1] 10146
```

3. Central Limit Theorem (T)

Let $X_1, X_2, \ldots$ be a sequence of independent $N(0,1)$ distributed random variables. For $n = 1, 2, \ldots$, let Y_n be the random variable, defined by $Y_n = X_1^2 + \ldots + X_n^2$.

(a) Show that $E[X_i^2] = 1$.

Solution: The variance of any random variable X can be written as

$$\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2$$

cf. page 97 (Dekking et al.). Solving for the wanted quantity gives us

$$\mathrm{E}\left[X^2\right] = \mathrm{Var}(X) + \mathrm{E}\left[X\right]^2,$$

where we know the variance is 1 and the expected value is 0, which gives us $\mathrm{E}\left[X^2\right] = 1 + 0^2 = 1$.

(b) One can show—using integration by parts—that $E[X_i^4] = 3$. Deduce from this that $\mathrm{Var}(X_i^2) = 2$.

Solution: Using the same formula for variance as above but substituting X for X^2 yields

$$\mathrm{Var}\left(X^2\right) = \mathrm{E}\left[\left(X^2\right)^2\right] - \mathrm{E}\left[X^2\right]^2 = \mathrm{E}\left[X^4\right] - \mathrm{E}\left[X^2\right]^2.$$

Using the given result $\mathrm{E}\left[X^4\right] = 3$, and the result from the previous exercise $\mathrm{E}\left[X^2\right] = 1$, we see

$$\mathrm{Var}\left(X^2\right) = \mathrm{E}\left[X^4\right] - \mathrm{E}\left[X^2\right]^2 = 3 - 1^2 = 2.$$

(c) Use the central limit theorem to approximate $P(Y_{100} > 110)$.

Solution: In the last 2 exercises we found that $\mathrm{E}(X^2) = 1$ and $\mathrm{Var}(X^2) = 2$. Since these are independent identically distributed random variables with positive variance, the central limit theorem applies cf. page 197 (Dekking et al.). Note that

$$\begin{aligned}
P(Y_{100} > 110) &= P\left(\frac{Y_{100}}{100} > \frac{110}{100}\right) && (6)\\
&= P\left(\sqrt{100}\frac{\frac{Y_{100}}{100} - 1}{\sqrt{2}} > \sqrt{100}\frac{\frac{11}{10} - 1}{\sqrt{2}}\right) && (7)\\
&= P\left(Z_{100} > \frac{1}{\sqrt{2}}\right) && (8)\\
&= 1 - P\left(Z_{100} \leq \frac{1}{\sqrt{2}}\right) && (9)\\
&\approx 1 - \Phi\left(\frac{1}{\sqrt{2}}\right). && (10)
\end{aligned}$$

We can evaluate the last equation with R:

```
pnorm(1/sqrt(2), 0, 1, lower.tail=FALSE)
```

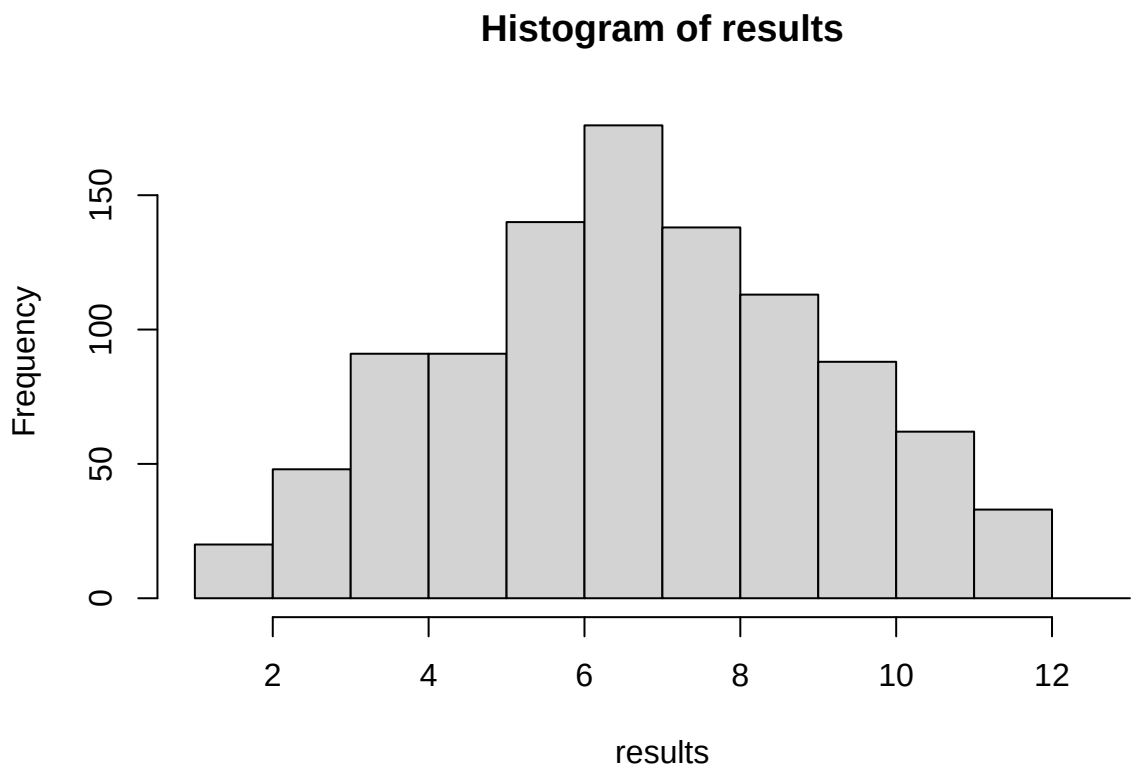
```
## [1] 0.2397501
```

4. Sum of Random Variables (R)

Assume you have n dice you throw simultaneously. The sum of outcomes can be characterised by the formula $X = X_1 + X_2 + ... + X_n$. Simulate the sum for thousand times using a couple of different values for n. Plot the histogram of the outcomes (`hist`) for each values of n you chose. What is the distribution like when $n = 2$? What can you see when you increase n?

Solution:

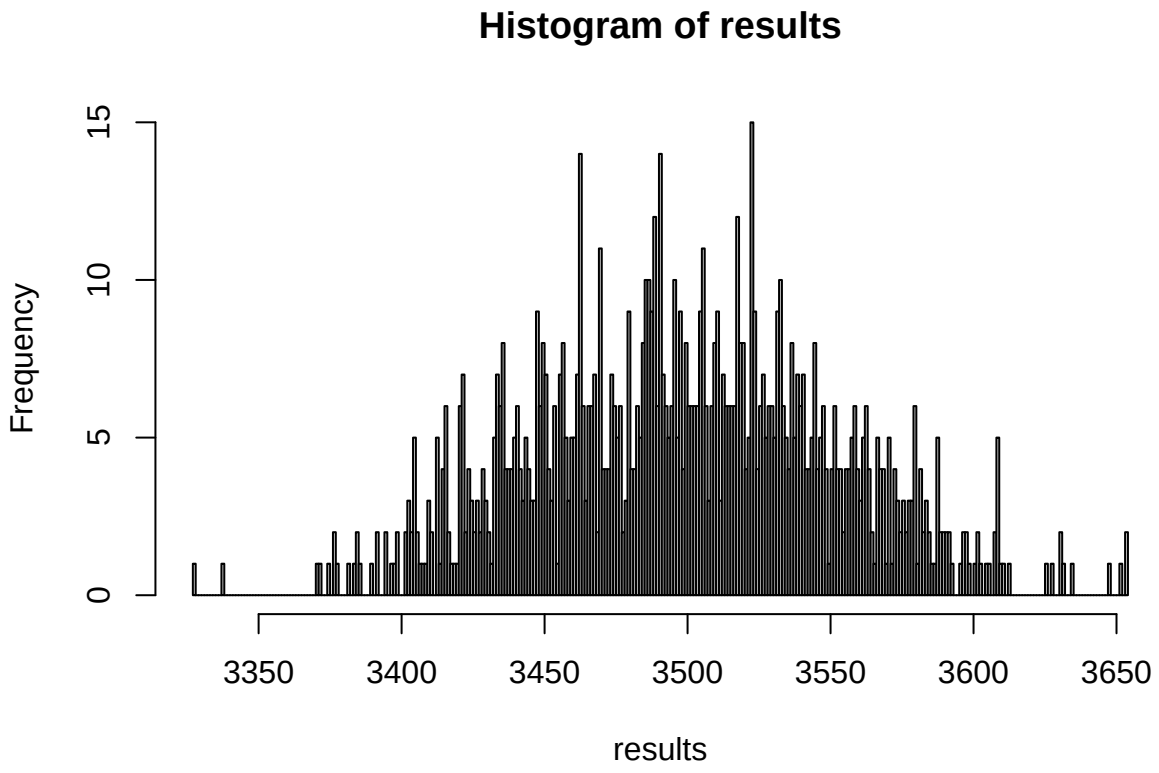
```
throw_dice <- function(n_dice) {
  results = 0
  for(i in (1:1000)){
    results[i] <- sum(sample(c(1, 2, 3, 4, 5, 6), n_dice,
                             replace=TRUE))
  }
  return(results)
}

results = throw_dice(2)
hist(results, breaks=seq(min(results)-1, max(results)+1, 1))
```




```
#seq(start, stop, step)
```

It looks like a triangular distribution. Increasing the number of dice will make the distribution tend towards a normal distribution.

```
results = throw_dice(1000)
hist(results, breaks=seq(min(results), max(results), 1))
```



5. Michelson Experiment (R)

In 1879, A.A. Michelson made a famous experiment to determine the speed of light, the data set is available as `Michelson(HistData)`. The velocity estimates are $v = 299000\text{km/s} + \delta v$ where δv is the velocity value saved in the data set.

(a) Study the examples shown by typing '?Michelson'.

Solution:

```
library(HistData)
```

```
## Warning: package 'HistData' was built under R version 4.0.3
```

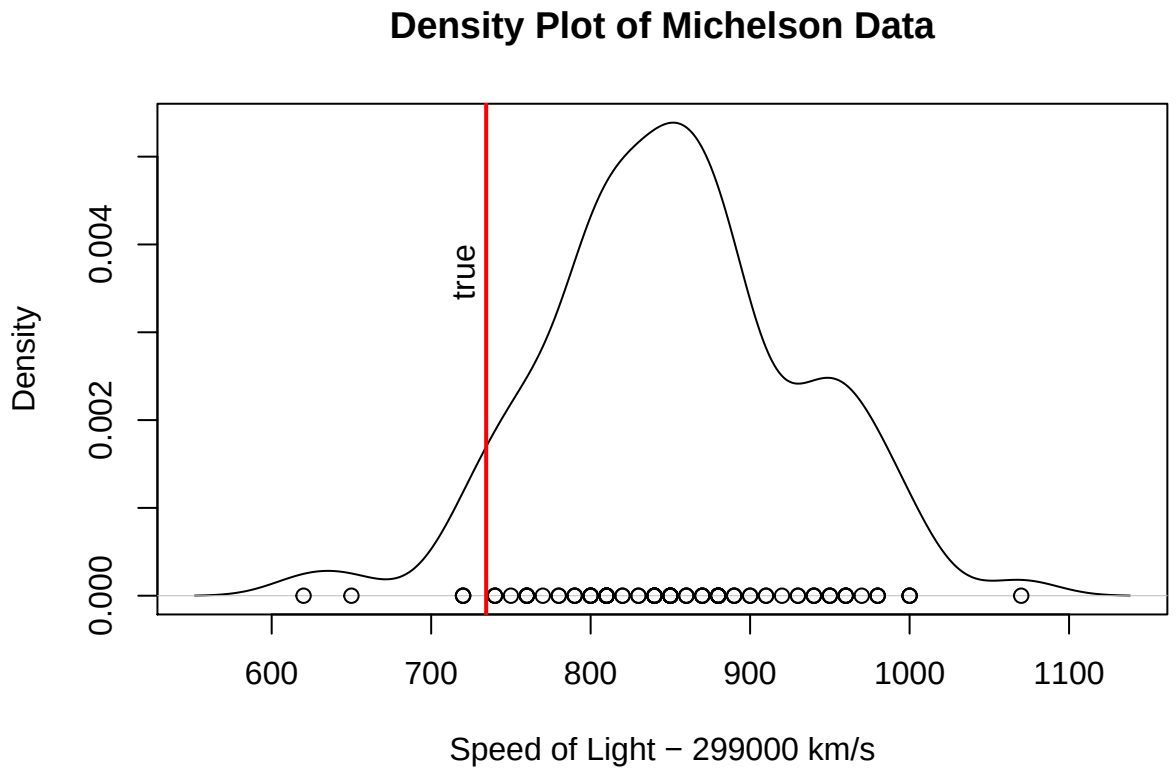
```
?Michelson
```

```
## starting httpd help server ... done
```

(b) Make density plot of the measurements together with data together with the individual estimates and true speed of light.

Solution:

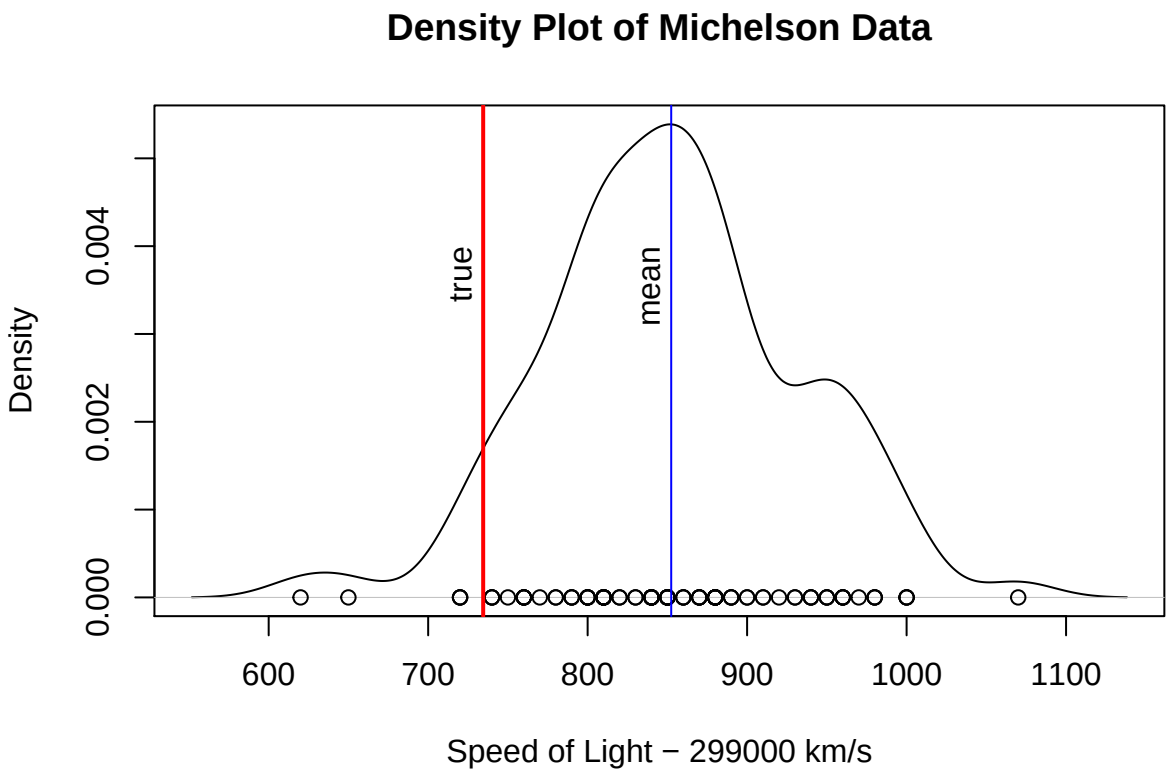
```
plot(density(Michelson$velocity),
     xlab = 'Speed of Light - 299000 km/s',
     main = 'Density Plot of Michelson Data')
points(x=Michelson$velocity, y=integer(length(Michelson$velocity)))
abline(v=734.5, lwd=2, col='red')
text(734.5, .004, "true", srt=90, pos=2)
```



(c) Assume that the velocity estimates are independent and identically distributed. Compute an estimate for the speed of light and compare to the true value known today. Add an estimate of it to the plot.

Solution: We'll take the mean of the measurements to be our estimate.

```
plot(density(Michelson$velocity),
     xlab = 'Speed of Light - 299000 km/s',
     main = 'Density Plot of Michelson Data')
points(x=Michelson$velocity, y=integer(length(Michelson$velocity)))
abline(v=mean(Michelson$velocity), col='blue')
abline(v=734.5, lwd=2, col='red')
text(mean(Michelson$velocity), .004, "mean", srt=90, pos=2)
text(734.5, .004, "true", srt=90, pos=2)
```



(d) Use Chebyshev's inequality to find out an upper bound for the probability that the speed of light was equal or further away from the value known today.

Solution: Using Chebyshev's inequality (pg. 183, Dekking et al.) we can calculate an upper bound for the probability:

```
mich_a = mean(Michelson$velocity - 734.5)
mich_var = var(Michelson$velocity)
prob = mich_var/mich_a^2
prob
```

```
## [1] 0.4490995
```

The interpretation of this is that there was at most a 45% chance that the true speed of light was at least as far as the the value known today or further away from the estimated mean. In other words: given the measurements, there is 55% chance that the true speed of light is in the region from the estimated mean to the true speed of light (in both directions).

(e) After looking closer at the data set do you have any reservations about the assumptions made above?

The estimates were likely not independent, contained systematic error, that led to a shifted mean.

6. Central Limit Theorem (R)

As a simulated experiment, draw n independent samples from Gamma distribution with shape parameter $a = 7.1$ and scale parameter $s = 1.44$, that is, $X_i \sim \Gamma(a, s)$, $i = 1, 2, \ldots, n$. Repeat the experiment m times

(a) Use the central limit theorem to compare the sample mean and variance of the standardised average to the theoretic values.

Solution:

```
gamma_sample <- function(n, m) {
  std_avg = 0
  gamma_exp_val = 7.1*1.44
  gamma_sd = sqrt(7.1*1.44^2)
  for(i in (1:m)) {
    trial_mean = mean(rgamma(n, shape=7.1, scale=1.44))
    std_avg[i] <- sqrt(n)*(trial_mean - gamma_exp_val)/gamma_sd
  }
  return(std_avg)
}
std_avg = gamma_sample(3000, 3000)
mean(std_avg)
```

```
## [1] -0.005408663
```

```
sd(std_avg)
```

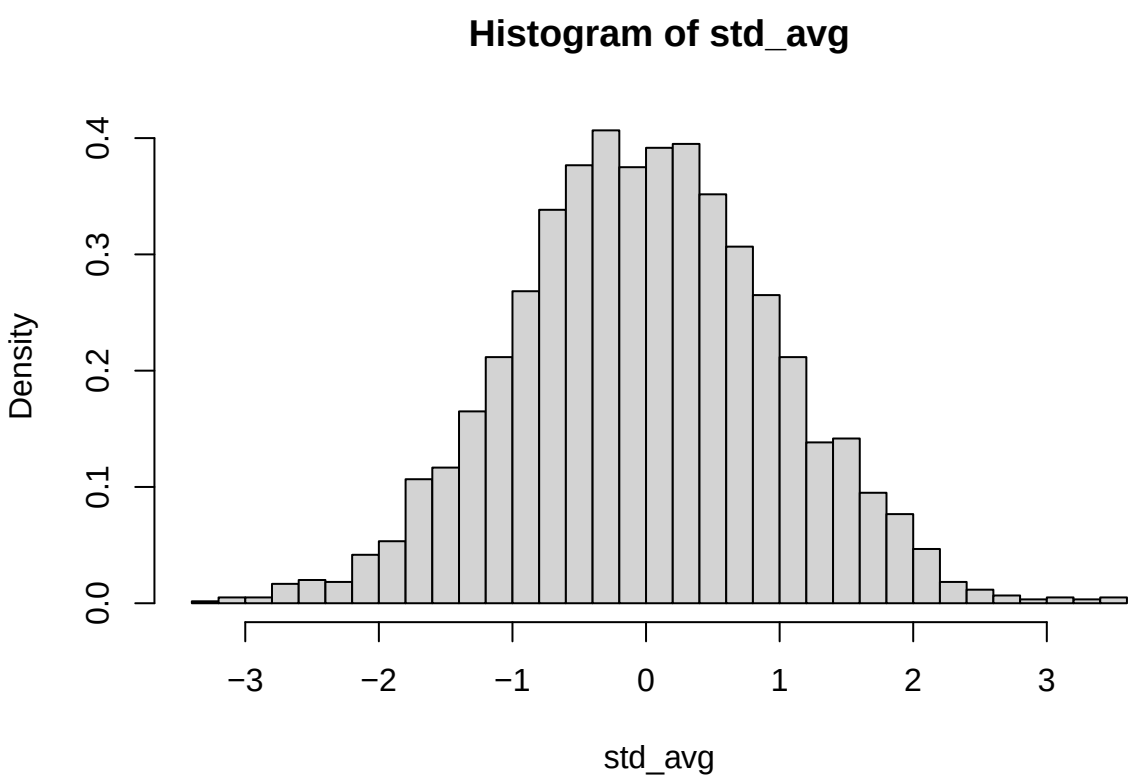
```
## [1] 0.987466
```

The theoretic value of the mean and variance should be 0 and 1 respectively.

(b) Plot the histogram for the standardized average. How far are you from the theoretic density that you will get on the limit?

Solution: Pretty close I'd say.

```
hist(std_avg, freq=F, breaks=25)
```


Histogram of std_avg
Density
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
std_avg