# DASS. Statistical laws

*Alla Tambovtseva*

### Random samples generation

Now we will generate random (*pseudo-random* to be more precise as you remember) samples from different distributions. Let's start from a normal distribution. Suppose that we want to take a sample from $N(\mu = 2, \sigma = 3)$. To do this we can use the function `rnorm()` (`r` stands for *random* and then goes the name of the distribution), indicate a sample size and specify the parameters of a distribution:

```
# N(mu=2, sigma=3) n = 10000
x <- rnorm(n = 10000, mean = 2, sd = 3)

# look at several values at the beginning
head(x)
```

```
## [1]  0.4018579  1.6753856 -1.5092802  0.8660410  3.3107953 -0.5999873
```

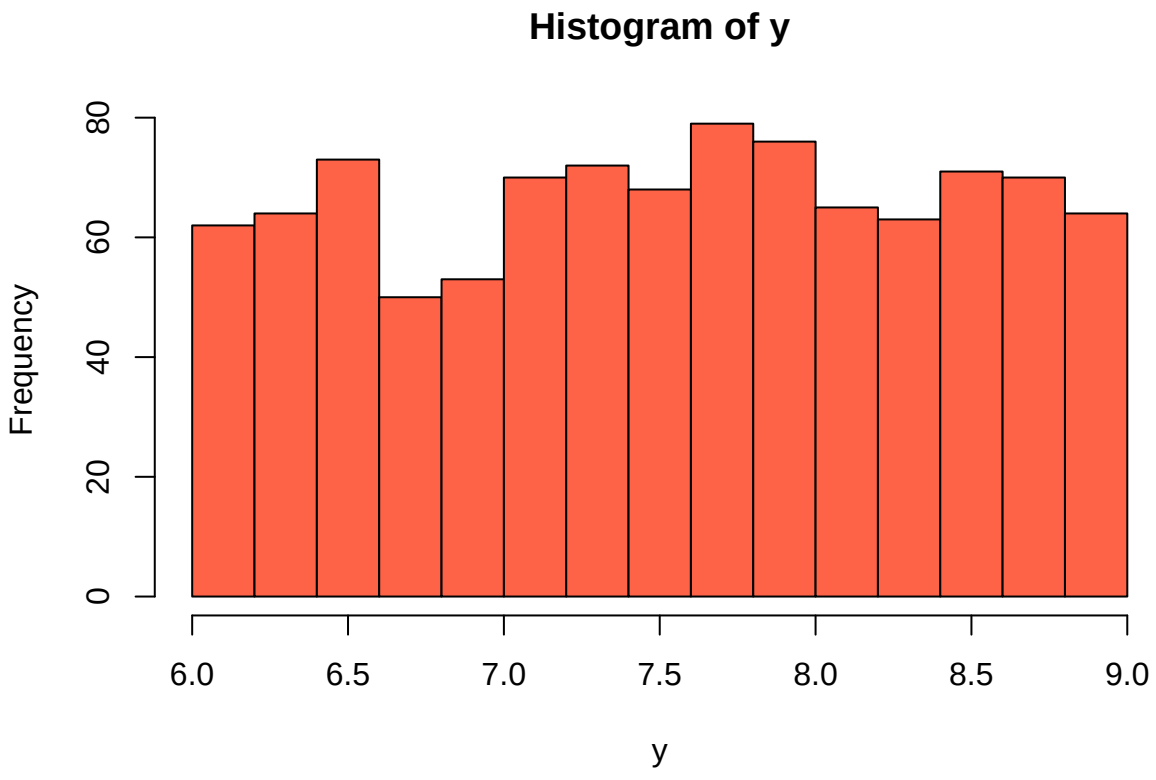
Of course, you will get a different sample due to the randomness inherited in computer algorithms. It is ok, then we will look how to set a starting point of an algorithm and make our code reproducible, i.e. returning the same results all the time.

Now consider a different distribution, a uniform one. There are two parameters, the endpoints of an interval on which a corresponding random variable is defined (here `min` and `max`).

```
y <- runif(n = 1000, min = 6, max = 9)
```

To see why it is called *uniform*, we can plot a histogram:

```
hist(y, col='tomato')
```



As it can be seen, the frequencies in $(6, 9]$ are approximately the same, so that is why we talk about uniformity.

### Law of Large Numbers

Now let's make an illustration for the Law of Large Numbers. We want to show that as a sample size increases, a sample mean approaches to a population mean. Firstly, we generate a population. Strictly speaking, usually we do not know a population size and we are not interested in it, but here we will just take a very large sample from a certain distribution and call it a population.

```
# w is the population
w <- rnorm(n = 50000, mean = 5, sd = 1.5)
```

Now we will take 1000 samples of a small size, calculate their means and plot the distribution of a sample mean.

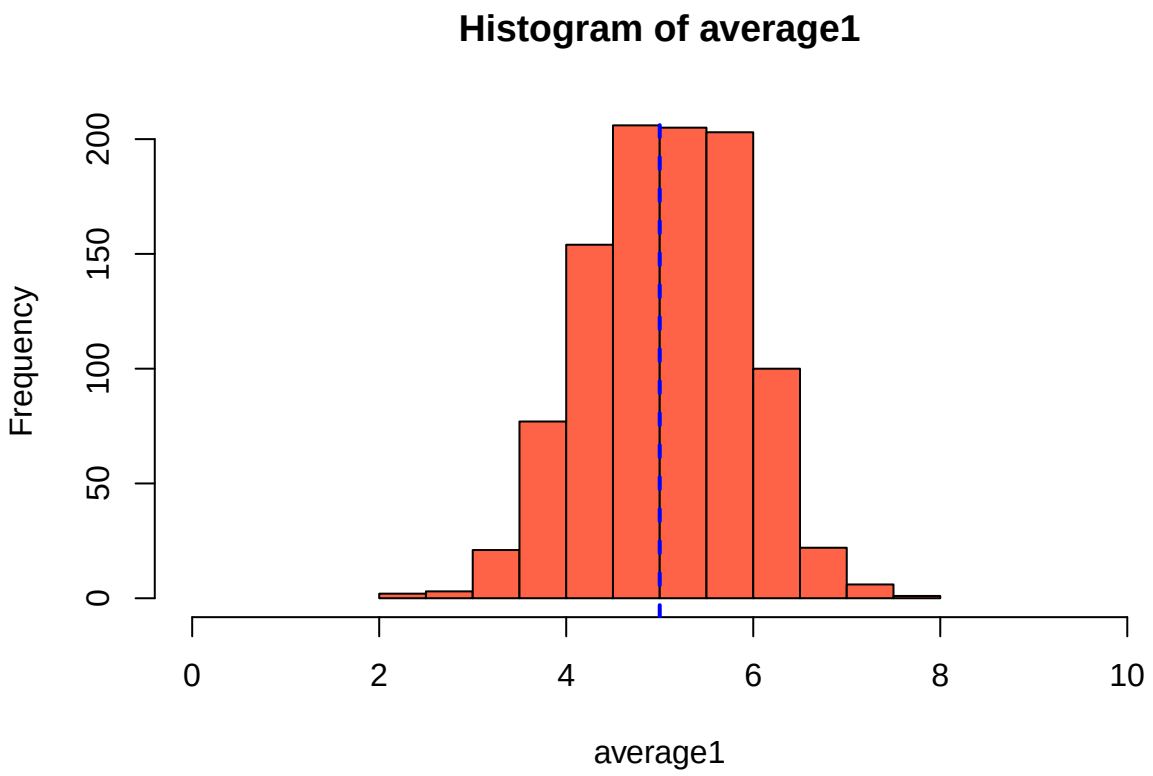
```
# we take a sample of size 3
n1 <- 3

# create a vector of 1000 NA's
average1 <- rep(NA, 1000)

# take 1000 different samples
# calculate their means
# and add to the vector average1
for (i in 1:1000){
  m <- mean(sample(w, n1))
  average1[i] <- m
}
```

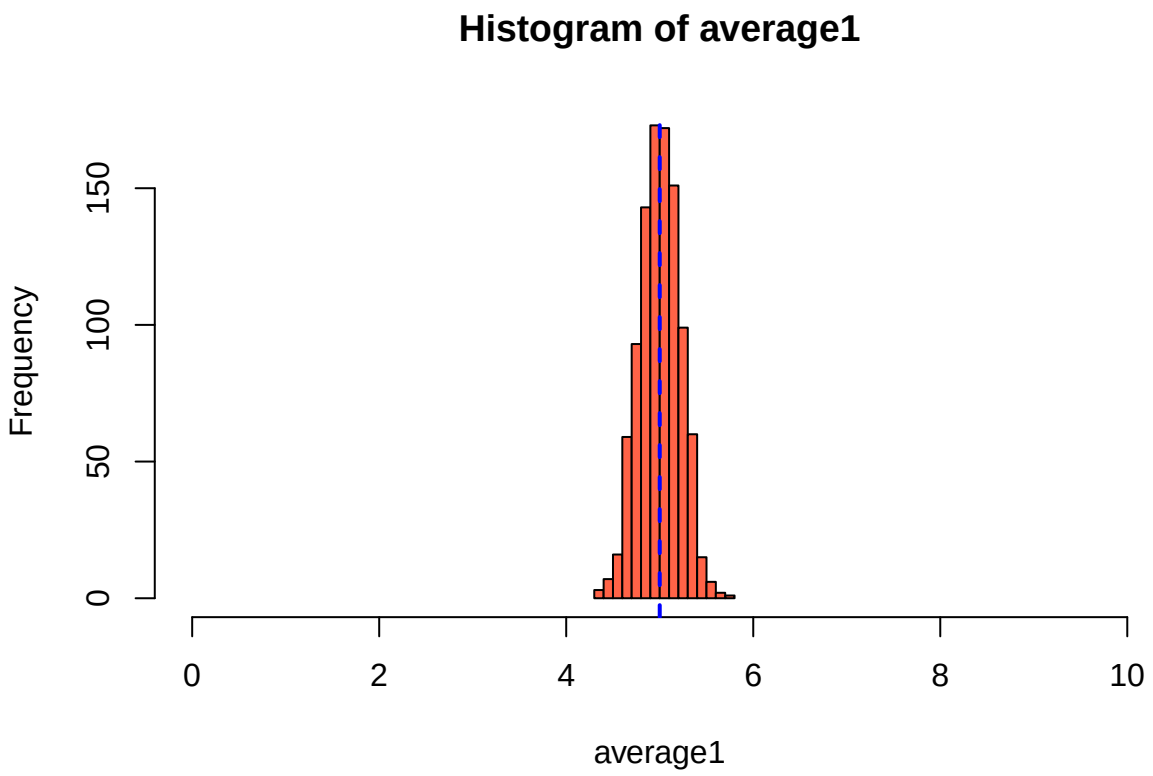
Now we can plot a histogram for sample means. Let's add a dashed vertical line that will show $\mu$, a population mean that is equal to 5 (we set it to 5 before).

```
# xlim - the limits for the x-axis
hist(average1, col="tomato", xlim=c(0, 10))
# v - vertical, lwd - line width, lty - line type
abline(v = 5, col = "blue", lwd = 2, lty = 2)
```



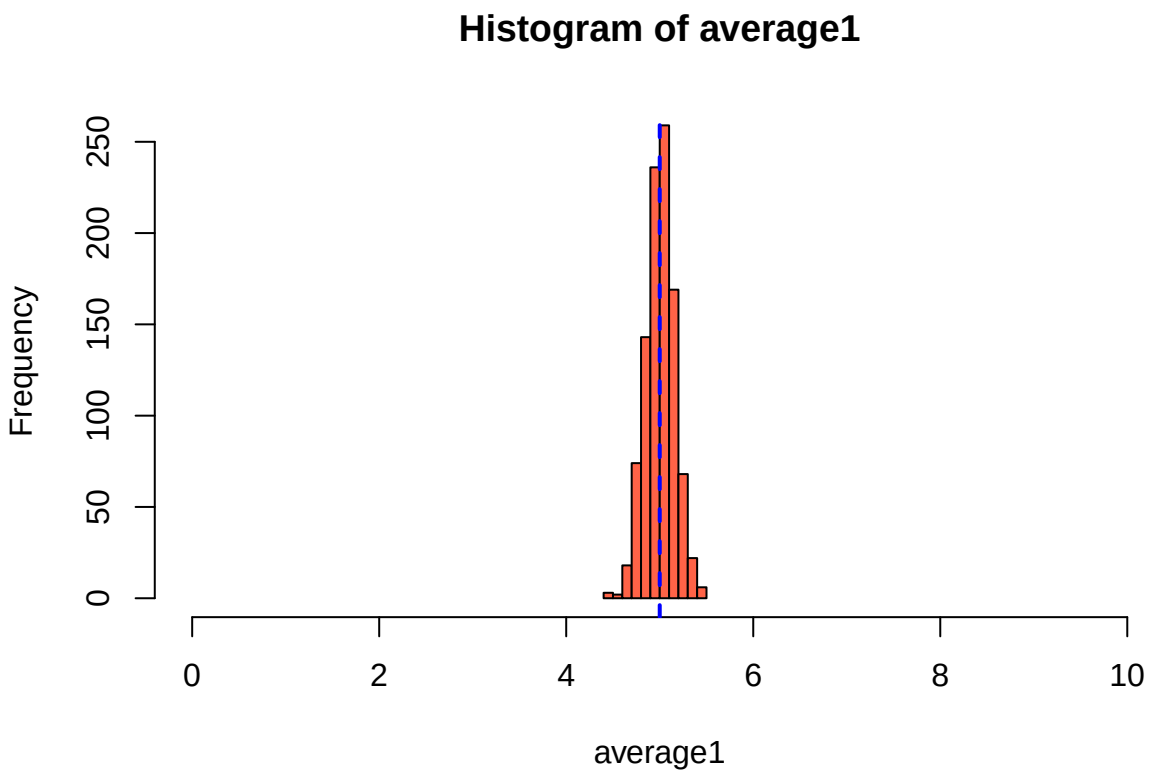

It is hard to make conclusions right now. It is better to plot histograms for cases with different sample sizes and then compare the results. Now the sample size is 50 (the number of samples is still 1000):

```
w <- rnorm(n = 50000, mean = 5, sd = 1.5)
n1 <- 50
average1 <- rep(NA, 1000)
for (i in 1:1000){
  m <- mean(sample(w, n1))
  average1[i] <- m
}
hist(average1, col="tomato", xlim=c(0, 10))
abline(v = 5, col = "blue", lwd = 2, lty = 2)
```




What can we see? This distribution is more concentrated around the expected value, so with the increase of a sample size, the variability of a sample mean reduces. Hence, a single value of a sample mean in general is closer to a population mean (expected value of a distribution). Finally, to highlight the difference, let's plot a histogram for the case of samples with `n = 100`:

```
w <- rnorm(n = 50000, mean = 5, sd = 1.5)
n1 <- 100
average1 <- rep(NA, 1000)
for (i in 1:1000){
  m <- mean(sample(w, n1))
  average1[i] <- m
}
hist(average1, col="tomato", xlim=c(0, 10))
abline(v = 5, col = "blue", lwd = 2, lty = 2)
```




Well, here it is very narrow as expected.

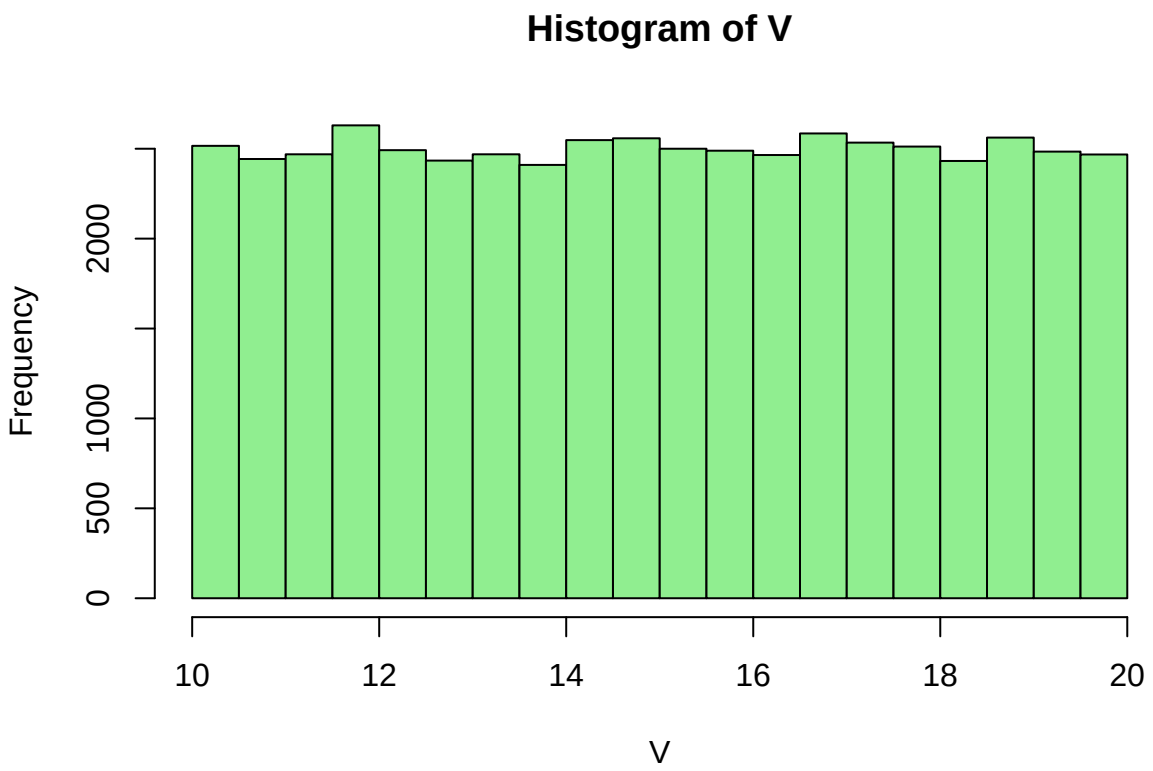
### Central Limit Theorem

Generate a population (here a very large sample) from the uniform distribution with `min=10`, `max=20`:

```
# set seed, a starting point for a reproducible code
# here it is 1234
set.seed(1234)
V <- runif(n = 50000, min=10, max=20)
```

Plot a histogram for the population:

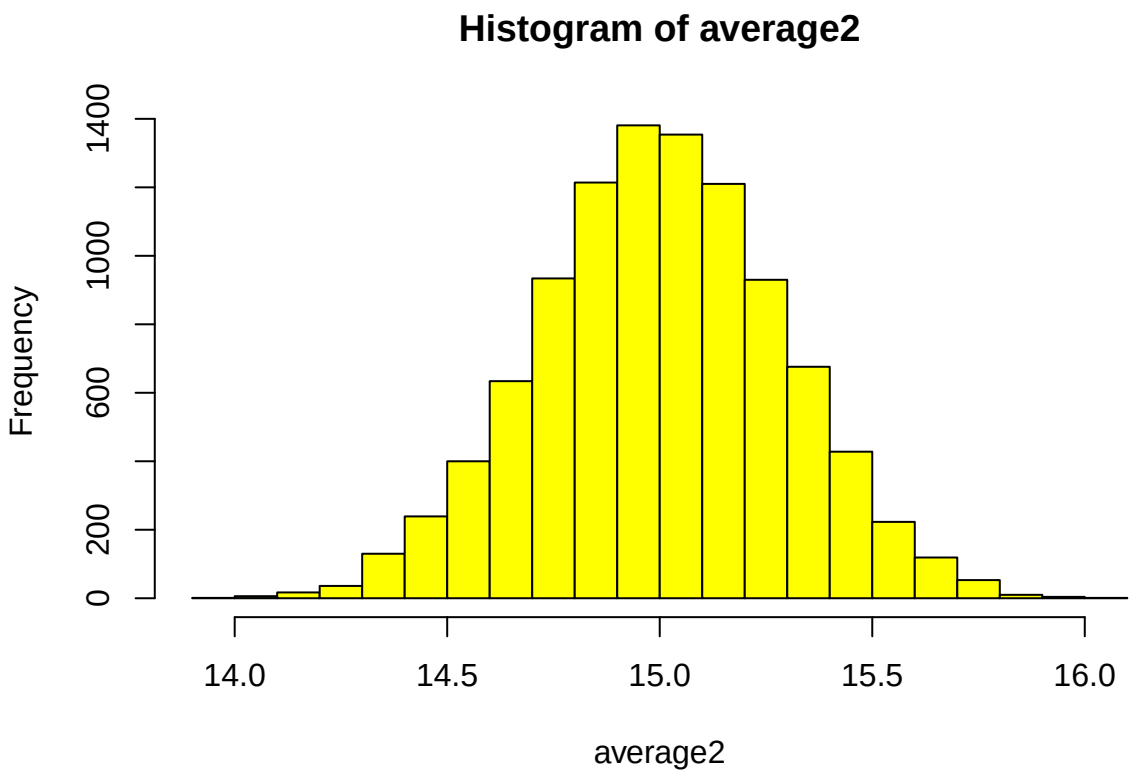
```
hist(V, col= 'lightgreen')
```




Does not look like a normal distribution, does it? And now take 10000 samples and plot a histogram for the series of their means:

```
set.seed(12)
average2 <- rep(NA, 10000)
for (i in 1:10000){
  m <- mean(sample(V, 100))
  average2[i] <- m
}

hist(average2, col = "yellow")
```




It is normal! Magic? No, just a central limit theorem. Finally, we can check whether the parameters of this normal distribution coincide with those promised by the theorem.

The expected value for such a uniform distribution is $\frac{min+max}{2} = \frac{10+20}{2} = 15$. The variance is $\frac{(max-min)^2}{12} = 8.33$ (do not think a lot about these formulas, they are just for illustration). The standard deviation is $\sqrt{8.33} = 2.89$.

So, as the central limit theorem says, the distribution of a sample mean is $N(\mu, \frac{\sigma}{\sqrt{n}})$. Here it should be $N(15, \frac{2.89}{100})$, so $N(15, 0.289)$ approximately. Let's check:

```
mean(average2)
```

```
## [1] 15.00102
```

```
sd(average2)
```

```
## [1] 0.2869689
```

Great! It works!