Chapter 5 - Sampling distribution

Dr. Alessandro Ruggieri

We can use pseudo-random number generators and data visualization packages to illustrate two important theorems in probability theory, i.e. the Central Limit Theorem and the Law of Large Number.

Central Limit Theorem

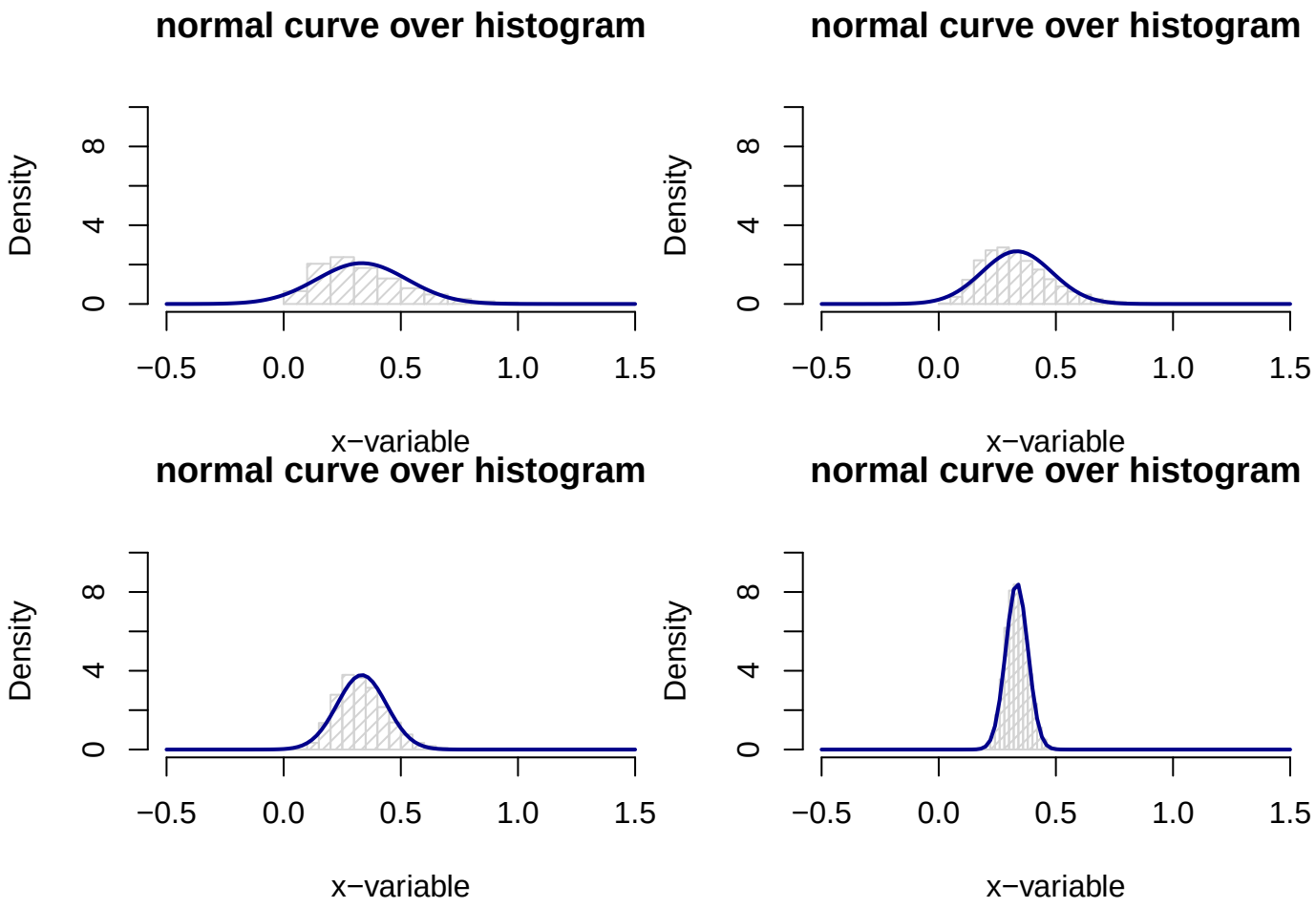
The Central Limit Theorem states the sample mean of a random sample of n observations drawn from a population with any probability distribution will be approximately normally distributed, if n is large.

To illustrate the theorem numerically, we consider a random variable distributed as exponential with scale parameters equal to 3. This random variable has a mean (mu) and standard deviation (sigma) equal to mu=sigma=1/3.

We take a random sample of size n from this distribution and compute the sample mean x. We repeat this process 10,000 times and store the sample mean in each iteration. Finally, we plot the histogram of 10,000 means and compare it to a normal distribution with mean equal to mu and standard deviation equal to sigma/n.

```
# Define sample size, mean and standard deviation for the average of an exponential r.v.
samplesize<-c(3,5,10,50)
mu<-1/3
sigma<-mu/sqrt(samplesize)

# pre-allocate variables used in the first loop
j<-1

# loop over sample size
for (value in samplesize)  {

# pre-allocate variables used in the second loop
n<-value
i<-0

maxi<-10000
xbar <- seq(0,0, len=maxi)

 # loop over repetitions
  while (i < maxi) {
    # sample n observation from exponential r.v.
    x<-rexp(n, r=1/mu)
    # compute sample mean
    xbar[i]<-mean(x)
    # update counter
    i<-i+1
  }
```

```r
# histogram
hist(xbar, density=20, breaks=20, prob=TRUE, xlab="x-variable", xlim=c(-0.5, 1.5), ylim=c(0, 10),
     main="normal curve over histogram")
curve(dnorm(x, mean=mu, sd=sigma[j]), col="darkblue", lwd=2, add=TRUE, yaxt="n")

# update counter
j<-j+1
}
```



As n increases, the normal distribution becomes a better approximation of the sampling distribution of the mean (regardless what is the the distribution of the underlying random variable).

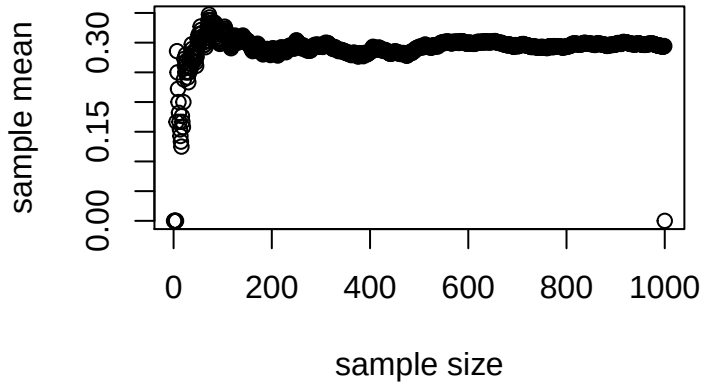
Law of Large Number

The Law of Large Number states that - given a random sample of size n taken from a population mean - the sample mean will "approach" the population mean as n increases, regardless of the underlying probability distribution of the data.

To illustrate the theorem numerically, we are going to sample one observation from a bernoulli random variable with probability p(0=0.3), and we will repeat this 1,000 times. Each time, we will compute the sample mean using observations sampled up to that point.

Finally, we are plotting the sequence of sample mean constructed iteratively.

```r
# Define probability for a bernoulli r.v.
p<-0.3

# Allocate objects
i<- 1
maxi <-1000
x    <-seq(0,0, len=maxi)
xbar <-seq(0,0, len=maxi)


# Generate sequentially 1,000 random numbers from bernoulli r.v.
while (i < maxi) {
x[i]<-rbinom(1, n=1, p=0.3)
# Compute sequentially the sample mean.
xbar[i] <- mean(x[1:i])
i<-i+1
}

# plot mean for each sample
t=1:maxi
plot(t,xbar, ylab="sample mean", xlab="sample size")
```



As the sample size becomes large, the variance of the sample mean will become smaller until eventually the distribution becomes degenerate around the population mean.