BB503/BB602 - R Training - Week IV

Ege Ulgen

Discrete Distributions

Discrete Uniform Distribution

```
### simulating 10000 die rolls
set.seed(123)
n_rolls <- 10000
results <- sample(1:6, n_rolls, replace = TRUE)

# pmf = 1/6 = 0.16667
table(results) / n_rolls
```

```
## results
##      1      2      3      4      5      6
## 0.1720 0.1712 0.1673 0.1590 0.1656 0.1649
```

```
# true expected value = (1+6) / 2 = 3.5
mean(results)
```

```
## [1] 3.4697
```

```
# true variance = (6^2 - 1) / 12 = 2.92
var(results)
```

```
## [1] 2.9444
```

Bernoulli Distribution

```
# install.packages("Rlab") # to install the packagee "Rlab"
library(Rlab)
```

```
## Rlab 2.15.1 attached.
```

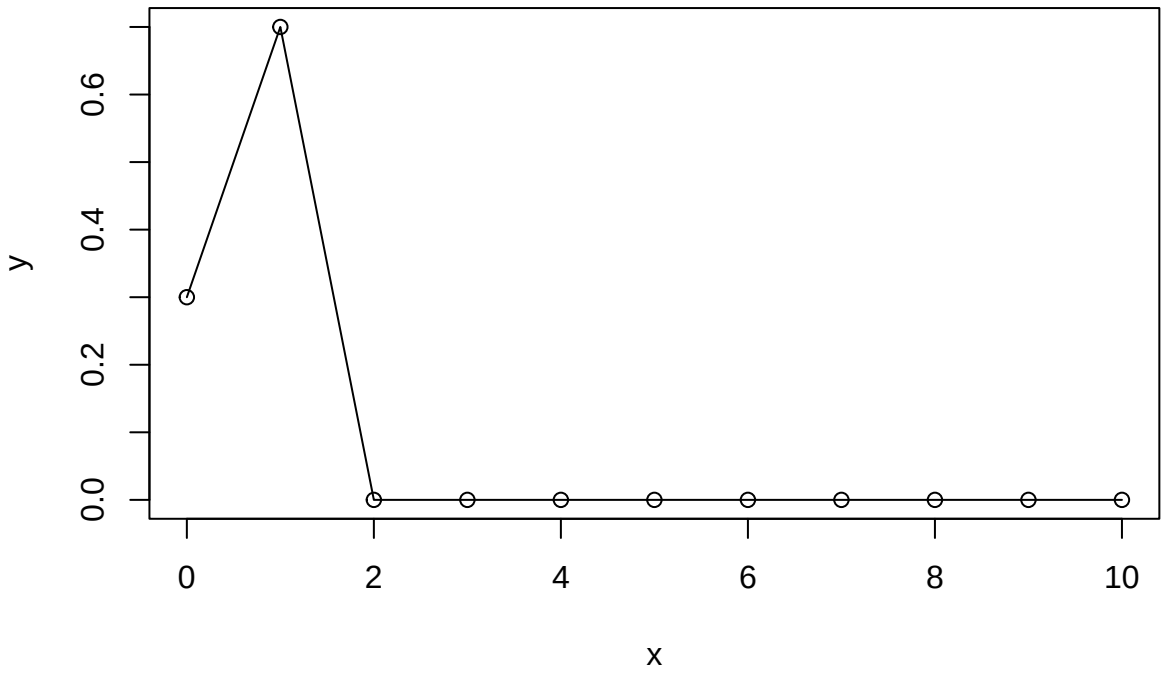
```
##
## Attaching package: 'Rlab'
```

```
## The following objects are masked from 'package:stats':
##
##     dexp, dgamma, dweibull, pexp, pgamma, pweibull, qexp, qgamma,
##     qweibull, rexp, rgamma, rweibull
```

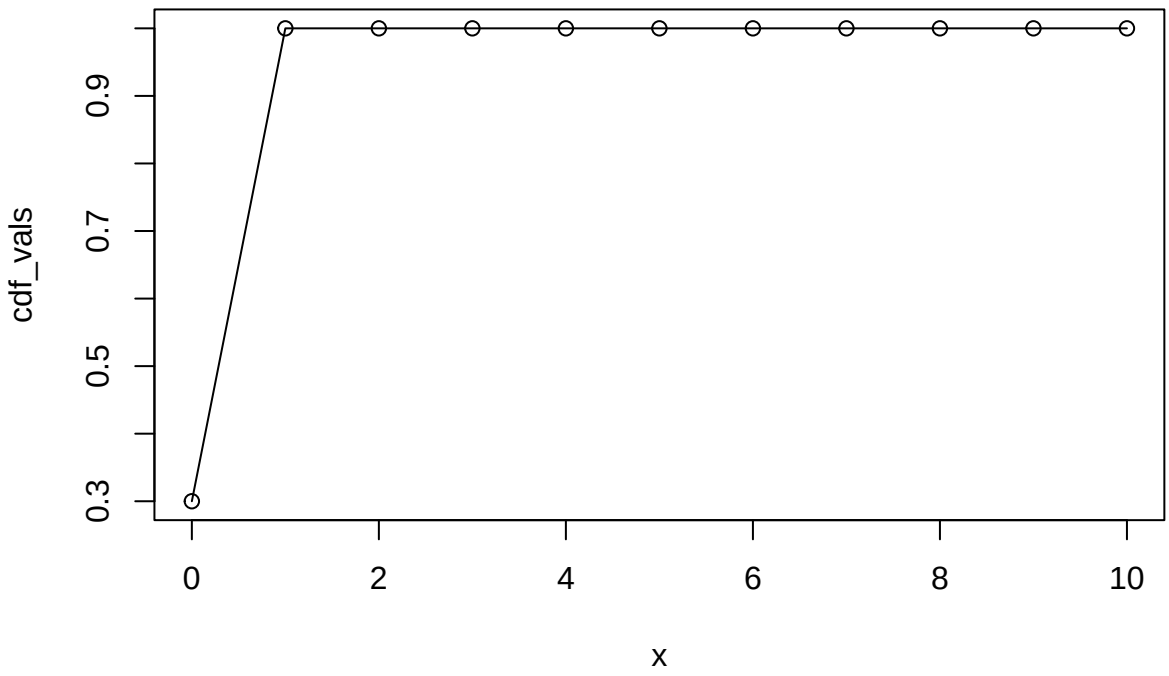
```
## The following object is masked from 'package:datasets':
##
##     precip
```

```
?dbern
```

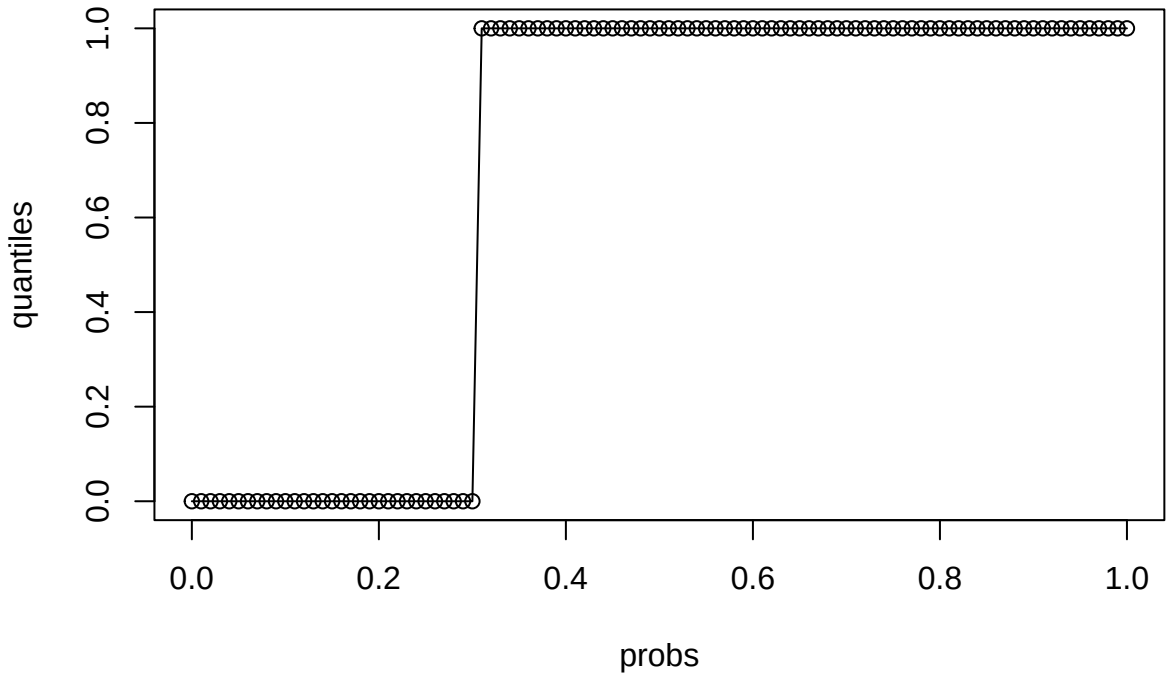
```r
# d** returns the density
x <- 0:10
y <- dbern(x, prob = 0.7)
plot(x, y, type = "o")
```



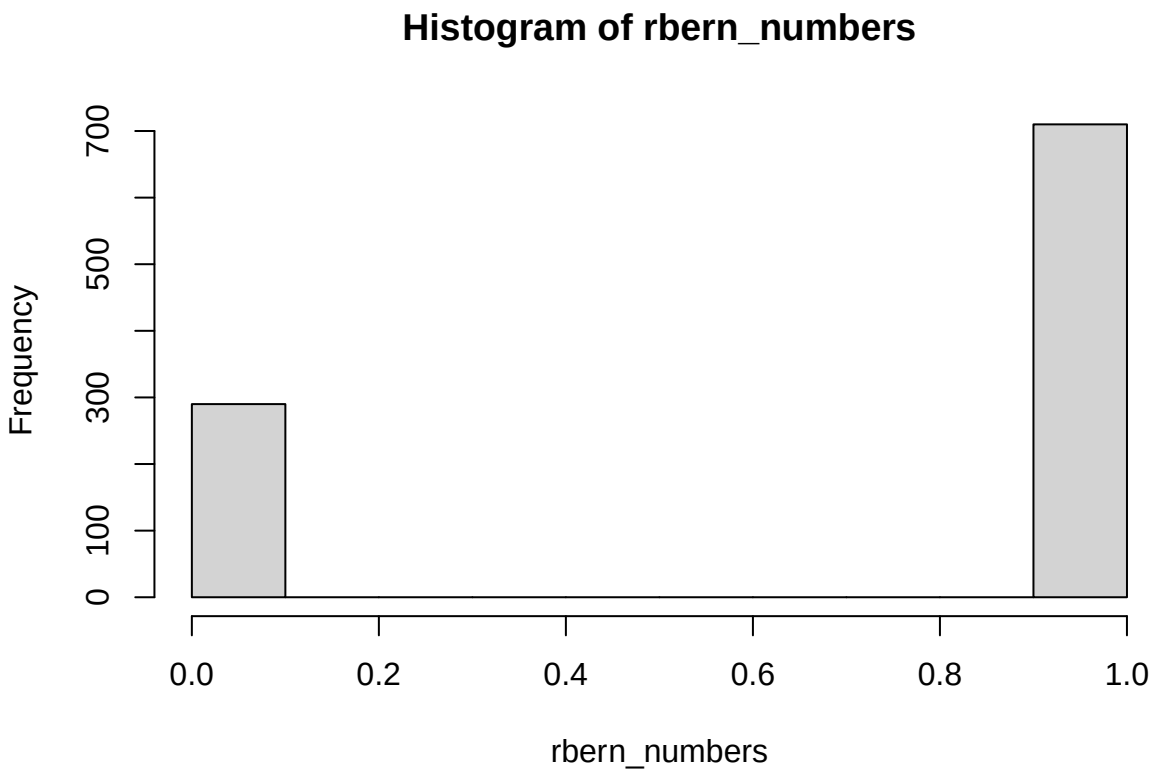

```r
# p** is the CDF
cdf_vals <- pbern(x, prob = 0.7)
plot(x, cdf_vals, type = "o")
```




```r
# q** is the quantile function
probs <- seq(0, 1, by = 0.01)
quantiles <- qbern(probs, prob = 0.7)
plot(probs, quantiles, type = "o")
```

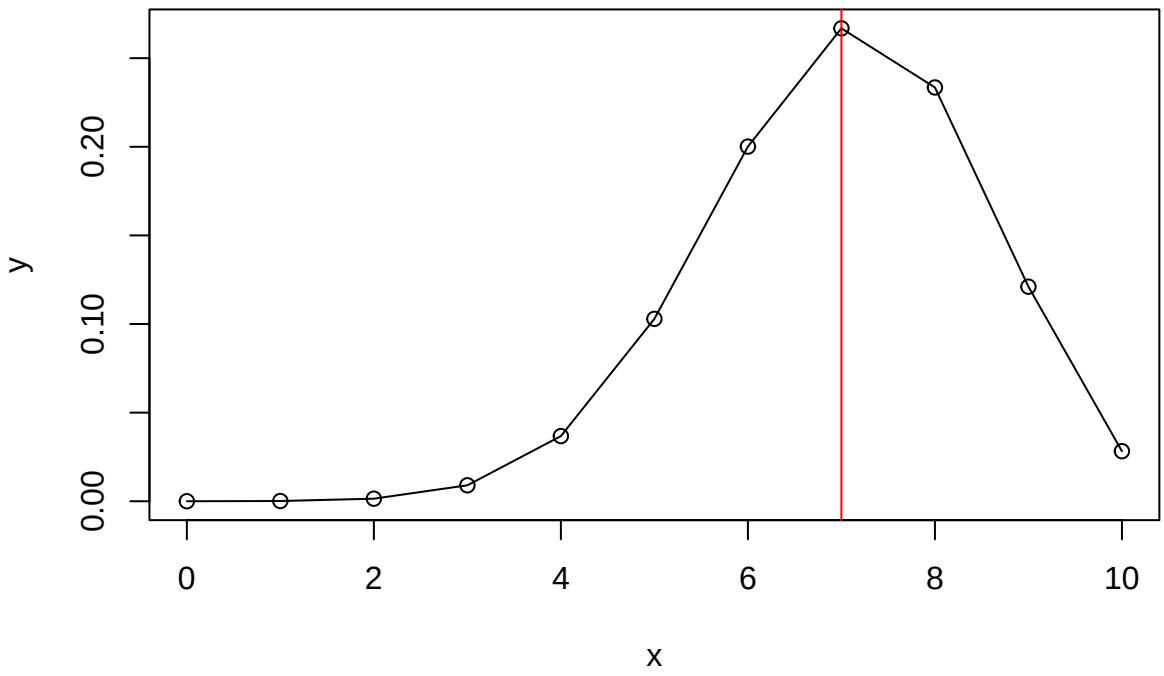



```r
# r** is the random number generator (random numbers that follow the given distribution)
rbern_numbers <- rbern(1000, prob = 0.7)
hist(rbern_numbers)
```

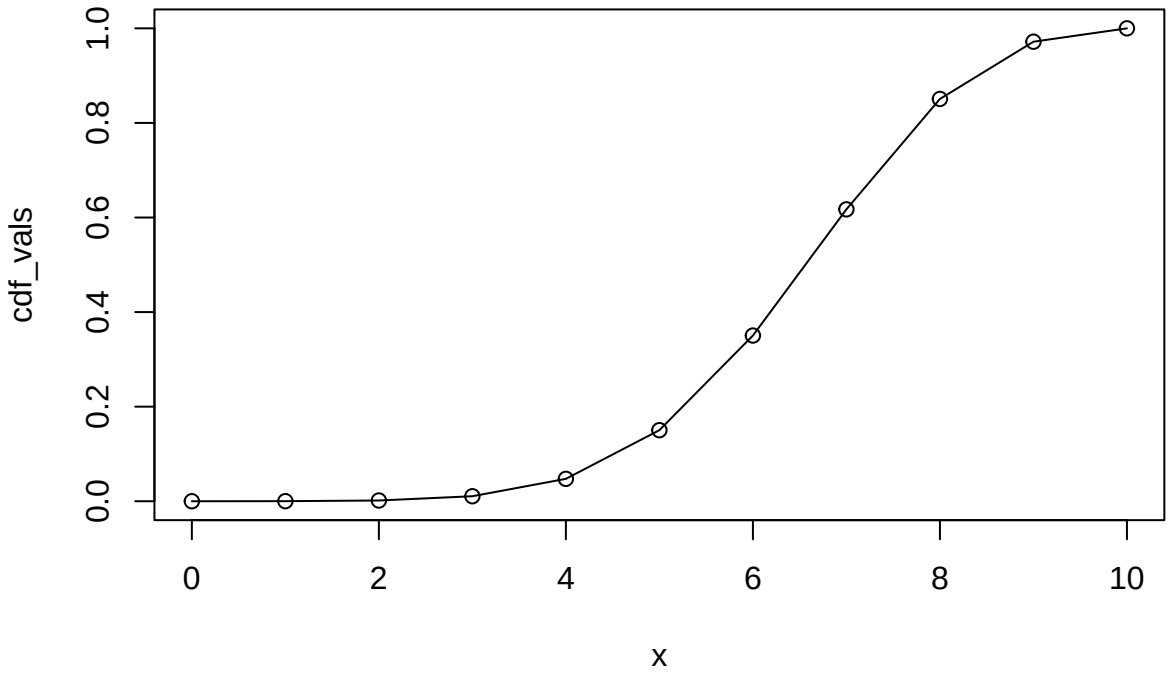


Binomial Distribution

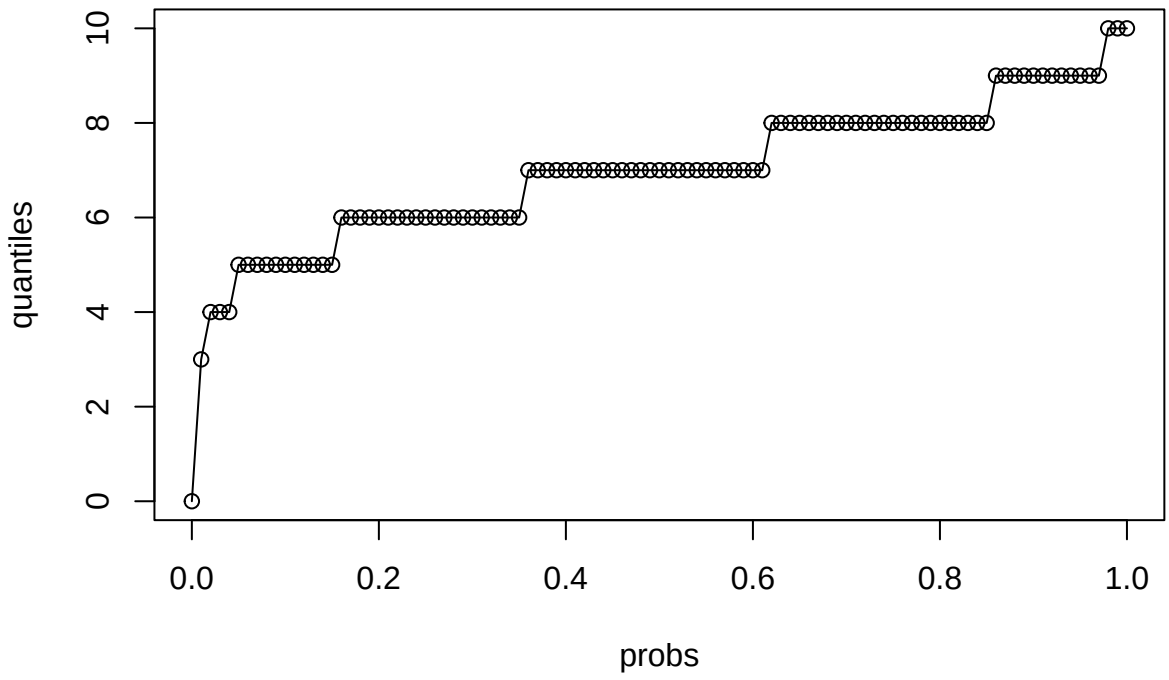
```
?dbinom

## flipping a coin 10 times

# d** returns the density
x <- 0:10
y <- dbinom(x, size = 10, prob = 0.7)
plot(x, y, type = "o")
abline(v = 7, col = "red") # expected value
```



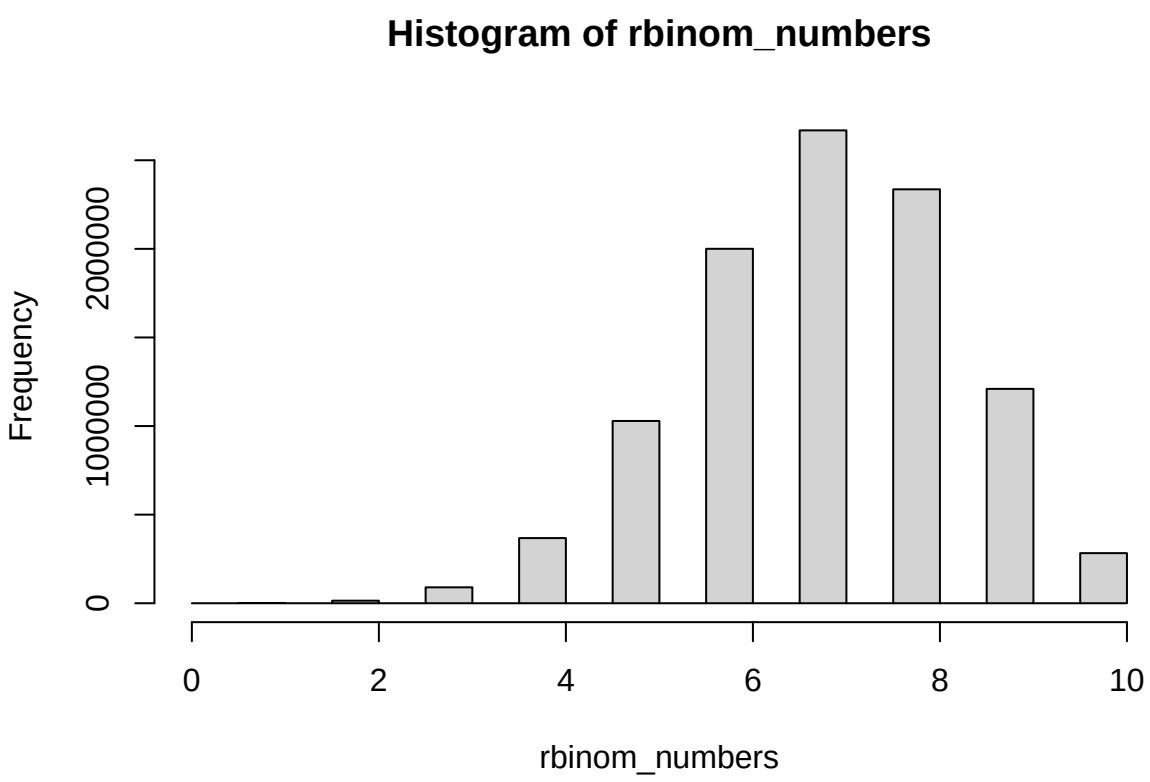

```
# p** is the CDF
cdf_vals <- pbinom(x, size = 10, prob = 0.7)
plot(x, cdf_vals, type = "o")
```




```r
# q** is the quantile function
probs <- seq(0, 1, by = 0.01)
quantiles <- qbinom(probs, size = 10, prob = 0.7)
plot(probs, quantiles, type = "o")
```




```r
# r** is the random number generator (random numbers that follow the given distribution)
rbinom_numbers <- rbinom(10000000, size = 10, prob = 0.7)
hist(rbinom_numbers)
```




```
### The success rate of a novel treatment is 20%. For 12 patients undergoing the novel
# treatment,

# a) what is the probability that the treatment is succcessfull in exactly 4?
dbinom(x = 4, size = 12, prob = 0.2)
```

```
## [1] 0.13288
```

```
# b) what is the probability that the treatment is succcessfull in at most 4?
dbinom(x = 4, size = 12, prob = 0.2) +
    dbinom(x = 3, size = 12, prob = 0.2) +
    dbinom(x = 2, size = 12, prob = 0.2) +
    dbinom(x = 1, size = 12, prob = 0.2) +
    dbinom(x = 0, size = 12, prob = 0.2)
```

```
## [1] 0.92744
```

```
# alternatively:
pbinom(q = 4, size = 12, prob = 0.2)
```

```
## [1] 0.92744
```

```
# c) what is the probability that the treatment is succcessfull in at least 4?
1 - pbinom(q = 3, size = 12, prob = 0.2)
```

```
## [1] 0.20543
```

```
pbinom(q = 4 - 1, size = 12, prob = 0.2, lower.tail = FALSE)
```

```
## [1] 0.20543
```

Geometric Distribution

Suppose the mortality rate of a novel operation technique is estimated to be 1%. What is the probability that the first mortality occurs on the 6th operation? $P(X = 6)$

```
dgeom(6, 0.01)
```

```
## [1] 0.0094148
```

What is the probability that the first mortality occurs before the 25th operation? $P(X <= 25)$

```
pgeom(25, 0.01)
```

```
## [1] 0.22996
```

Let's determine the number of operations we would expect to be performed until a mortality occurs:

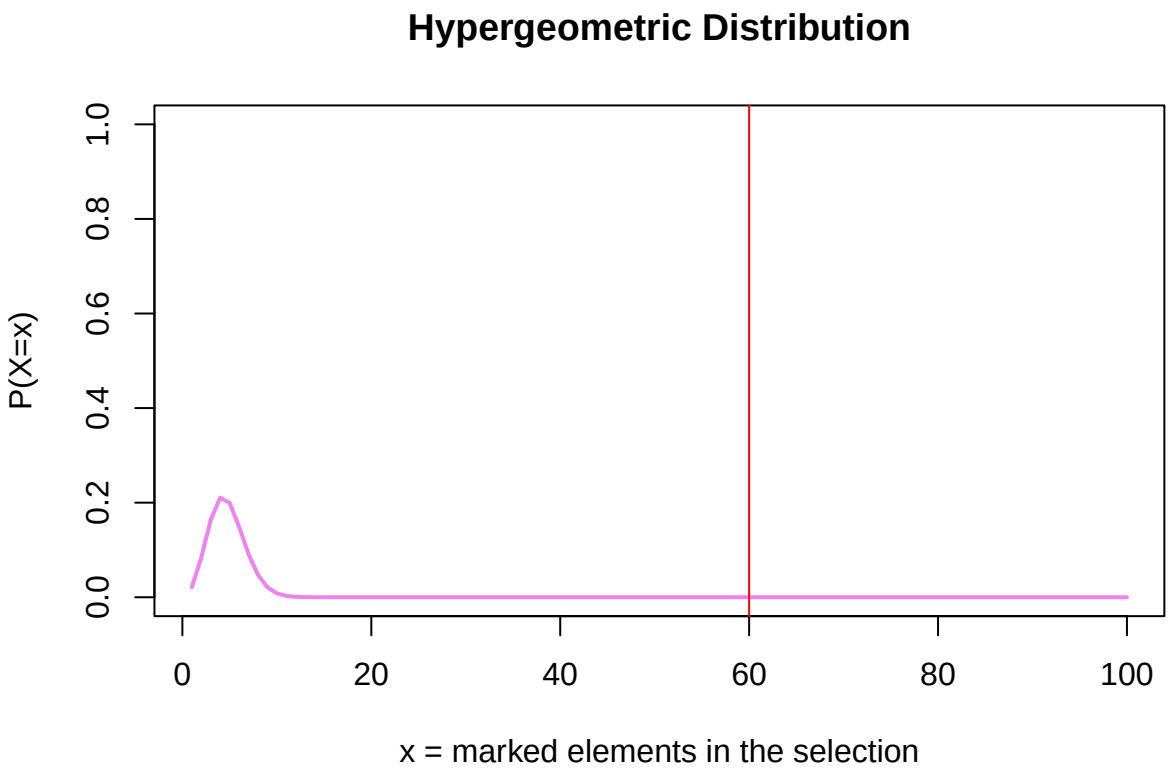
```
# E[X] = 1 / p
1 / 0.01
```

```
## [1] 100
```

Hypergeometric Distribution

Suppose we are performing an RNA array experiment that targets 20000 genes in total. We find that 75 genes are important (e.g., deferentially expressed between tumors vs. controls). We find that 60 of the important genes are in a functional gene set (e.g., Cell cycle) of 1000 genes.

```
### Over Representation Analysis
N <- 20000 ## Total number of genes targeted by array
k <- 75 ## Number of important genes (size of the selection)
m <- 1000 ## Size of the gene set, i.e. genes associated to this biological process
n <- N - m ## Number of "non-marked" elements, i.e. genes not associated to this biological process
x <- 60 ## Number of "marked" elements in the selection, i.e. genes of the group of interest that are a

## Compute the distribution (for all possible values of x).
x.range <- 1:100
hyper_geom_dens <- dhyper(x = x.range - 1, m = m, n = n, k = k)

## Plot the pdf
plot(x.range, hyper_geom_dens, type="l",
     lwd=2,
     col="violet",
     main="Hypergeometric Distribution",
     xlab="x = marked elements in the selection",
     ylab="P(X=x)", ylim=c(0, 1))
abline(v = x, col = "red")
```




```
p.value <-  phyper(q = x - 1, m = m, n = n, k = k, lower.tail = FALSE)
```

Poisson Distribution

In a city, the mean number of people dying from a rare disease is 4 in a week. In a certain week,

a) What is the probability that no one dies from the disease?

b) What is the probability that at least 2 people die from the disease?

```
# a) the probability that no one dies from the disease
dpois(x = 0, lambda = 4)
```

```
## [1] 0.018316
```

```
# b) the probability that at least 2 people die from the disease
1 - (dpois(1, 4) + dpois(0, 4))
```

```
## [1] 0.90842
```

```
ppois(2 - 1, 4, lower.tail = FALSE)
```

```
## [1] 0.90842
```