I. Distribution of feature vectors

- each feature is binary, with the same binomial distribution

\[
\begin{align*}
N & \leftarrow 10^6 & \quad \text{# number of observations} \\
\text{dim} & \leftarrow 7 & \quad \text{# number of dimensions} \\
\text{prob} & \leftarrow 1/10 & \quad \text{# probability of } W_i = 1
\end{align*}
\]

\[
\text{binom\_vector} \leftarrow \text{character}(N) \\
\text{for}(i \in 1:N) \text{ binom\_vector}[i] \leftarrow \text{paste(rbinom(dim,1,prob), collapse="")}
\]

\[
\text{expected\_values} = 2^\text{dim} \\
\text{emerged\_values} = \text{length(unique(binom\_vector))}
\]

\[
\text{print( sort(table(binom\_vector), dec=T) )}
\]

observations = 1,000,000
dimensions = 7
p(x = 1) = 0.1
number of possible different values = 128
number of emerged different values = 125
II. Distribution of distances from 0 of randomly distributed points in a unit cube

dim <- 6
n <- 10000

cube <- data.frame(
    x1 = runif(n),
    x2 = runif(n),
    x3 = runif(n),
    x4 = runif(n),
    x5 = runif(n),
    x6 = runif(n)
)

distances <- numeric(n)

for(i in 1:n) distances[i] <- sqrt(sum(cube[i,]^2))
greater_than_1 <- sum(distances > 1)

message("Most of the distances (",
        format(greater_than_1/n*100, digits=3), ",\%) are greater than 1."
)

message("Frequency of distances in intervals:")
print(table(cut(distances, breaks=seq(0, 2.5, 0.5))))

----------------------------------------------------------------

This program generates 10,000 random 6-dimensional sample points in a unit cube.
Maximum possible distance from 0 is: 2.45

Distances from 0 in the sample of 10000 points:

   Min. 1st Qu. Median    Mean 3rd Qu.    Max.
 0.324 1.214 1.407 1.390 1.578 2.183

Most of the distances (91.7\%) are greater than 1.
Frequency of distances in intervals:

(0,0.5] (0.5,1] (1,1.5] (1.5,2] (2,2.5]
   7   824  5591  3529   49
III. Distribution of mutual distances between randomly distributed points in a unit cube

```r
dim <- 6
n <- 150
d <- choose(n,2)
lim <- 0.5

message("Maximum possible distance between two points is: ",
       format(sqrt(dim), digits=3) )
message("Number of different pairs is: ", d)

cube <- data.frame(
    x1 = runif(n),
    x2 = runif(n),
    x3 = runif(n),
    x4 = runif(n),
    x5 = runif(n),
    x6 = runif(n)
  )

distances <- numeric(d)

k <- 1
for(i in 1:(n-1) ) for(j in (i+1):n ) {
  distances[k] <- sqrt( sum((cube[i,]-cube[j,])^2) ); k <- k+1
}
greater_than_lim <- sum(distances > lim)

message("Most of the distances (",
       format(greater_than_lim/d*100, digits=3), ",%\) are greater than ", lim, ",.\)

message("Frequency of distances in intervals:")
print(table(cut(distances, breaks=seq(0, 2.5, 0.25))))
```

This program generates 150 random 6-dimensional sample points in a unit cube. Maximum possible distance between two points is: 2.45 Number of different pairs is: 11,175

Mutual distances in the sample of 150 points:

```
   Min. 1st Qu.  Median     Mean 3rd Qu.    Max.  
0.1173  0.8098  0.9797  0.9732  1.1420  1.8770
```

Most of the distances (97%) are greater than 0.5. Frequency of distances in intervals:

```
(0,0.25] (0.25,0.5] (0.5,0.75] (0.75,1] (1,1.25] (1.25,1.5] (1.5,1.75] (1.75,2] (2,2.25] (2.25,2.5] 
  9   322   1704   3914   3793   1301    128    4     0     0
```