# The Nature of Collisions

*By Grandmaster Mark Aarøe Nissen*

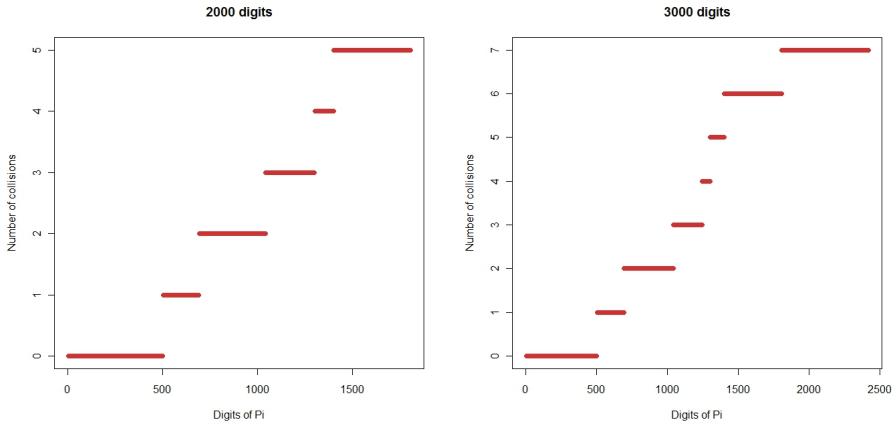
**December 2014**

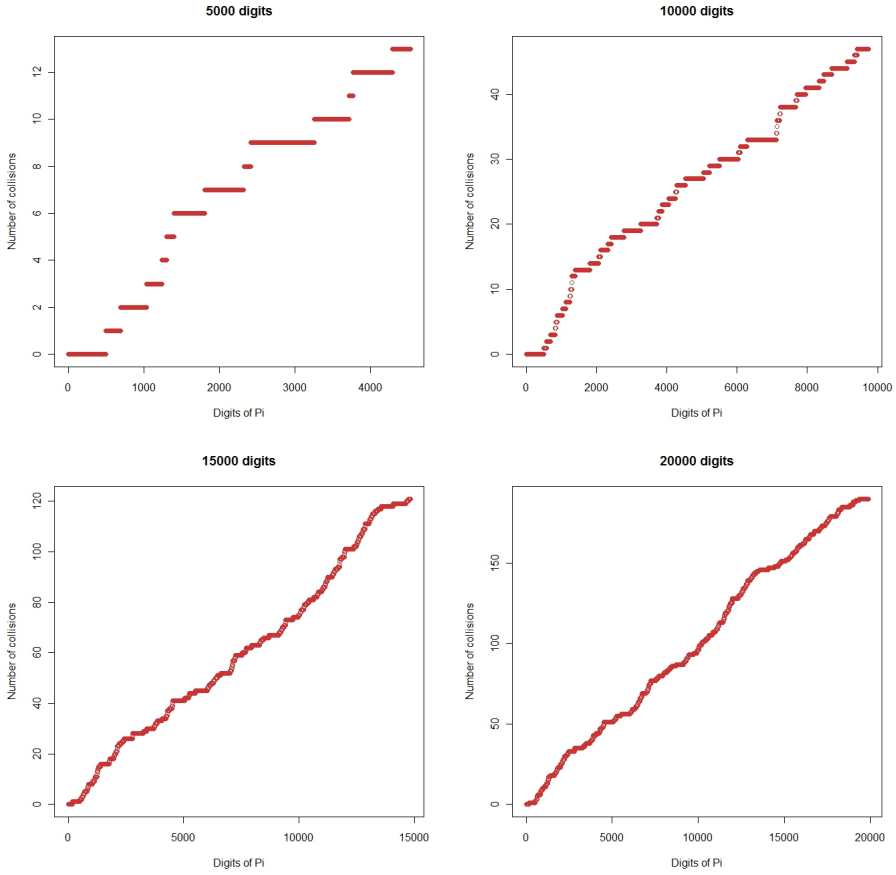
## Introduction

We will consider the digits of $\pi$ in blocks of 5 digits. A *collision* is the same block of five digits occurring two or more times in a sequence of digits. In this paper we will briefly investigate the development of collisions in $\pi$ subjected to a fixed number of digits of $\pi$. The most common choice is $\pi = 3.14159....$, however, we will briefly calculate the number of collisions for the 4 other cases, namely $\pi = 3.1, 41592... = 3.14, 15926... = 3.141, 59265... = 3.1415, 92653....$ We will make a list of all the collisions and their positions for the first 10,000 digits of $\pi$ and see how the frequency of collisions develops subjected to the number of digits of $\pi$. Finally we will calculate the probability of getting a collision after 50 random draws without replacement. All the calculations are successfully done in the program R and the code is available in the appendix.

## The increase of collisions

First we wish to plot the number of collisions against a fixed number of digits of $\pi$. The following plots shows us how the collisions increases depending on the digits of $\pi$.






## The collisions for the first 10,000 digits of $\pi$

As mentioned our aim is to find the collisions and their positions for the first 10,000 digits of $\pi$. For the order $\pi = 3.14159, ...$ we have the following collisions.

```
      [,1] [,2] [,3] [,4] [,5]
 [1,] "9"  "4"  "9"  "1"  "2"
 [2,] "2"  "9"  "3"  "1"  "7"
 [3,] "9"  "2"  "2"  "7"  "9"
 [4,] "4"  "2"  "5"  "2"  "2"
 [5,] "3"  "3"  "4"  "4"  "6"
 [6,] "3"  "2"  "0"  "8"  "3"
 [7,] "8"  "2"  "7"  "9"  "6"
 [8,] "5"  "0"  "8"  "2"  "9"
 [9,] "4"  "6"  "7"  "6"  "7"
[10,] "5"  "5"  "3"  "7"  "9"
[11,] "4"  "6"  "6"  "8"  "4"
[12,] "9"  "4"  "9"  "1"  "2"
```

```
[13,] "9"  "2"  "2"  "7"  "9"
[14,] "8"  "2"  "7"  "9"  "6"
[15,] "4"  "7"  "8"  "2"  "6"
[16,] "1"  "4"  "2"  "9"  "8"
[17,] "8"  "0"  "7"  "9"  "7"
[18,] "4"  "6"  "7"  "6"  "7"
[19,] "6"  "5"  "4"  "0"  "3"
[20,] "9"  "7"  "1"  "2"  "0"
[21,] "9"  "7"  "1"  "2"  "0"
[22,] "0"  "3"  "2"  "0"  "3"
[23,] "3"  "2"  "1"  "5"  "1"
[24,] "1"  "2"  "5"  "7"  "0"
[25,] "9"  "4"  "5"  "6"  "1"
[26,] "8"  "0"  "7"  "9"  "7"
[27,] "0"  "3"  "2"  "0"  "3"
[28,] "2"  "9"  "7"  "4"  "1"
[29,] "3"  "3"  "4"  "4"  "6"
[30,] "9"  "4"  "5"  "6"  "1"
[31,] "4"  "2"  "5"  "2"  "2"
[32,] "2"  "9"  "1"  "7"  "4"
[33,] "9"  "6"  "5"  "0"  "1"
[34,] "5"  "5"  "3"  "7"  "9"
[35,] "2"  "9"  "3"  "1"  "7"
[36,] "3"  "2"  "0"  "8"  "3"
[37,] "4"  "6"  "6"  "8"  "4"
[38,] "7"  "7"  "3"  "7"  "9"
[39,] "5"  "0"  "8"  "2"  "9"
[40,] "1"  "2"  "5"  "7"  "0"
[41,] "2"  "9"  "1"  "7"  "4"
[42,] "7"  "7"  "3"  "7"  "9"
[43,] "4"  "7"  "8"  "2"  "6"
[44,] "9"  "6"  "5"  "0"  "1"
[45,] "2"  "9"  "7"  "4"  "1"
[46,] "3"  "2"  "1"  "5"  "1"
[47,] "1"  "4"  "2"  "9"  "8"
[48,] "6"  "5"  "4"  "0"  "3"
```

The probability for a collision is therefore $p = \frac{48}{2000} \approx 2.4$ percent. The positions of the collisions are given below.

```
 [1]  500  575  690  825  835
      880 1040 1125 1245 1260
      1290 1300 1400 1805 2060
[16] 2125 2325 2420 2790 3260
      3715 3770 3860 4060 4245
      4290 4530 5045 5215 5505
[31] 6025 6090 6290 7110 7130
      7140 7215 7230 7670 7700
      7950 8340 8465 8690 9130
```

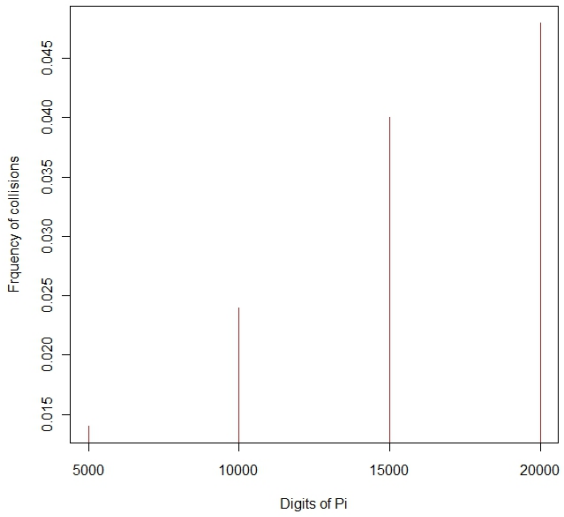
```
[46]  9345 9425 9735
```

For the other four orders of $\pi$ we have other numbers of collisions, namely 30,50,38,22.

## The frequency of collisions

Now consider the case where we choose to have a block of a different length than 5. Consider the 10,000 digits as a $1 \times 1000$ matrix we clearly have 10,000 collisions (100 percent), for a $2 \times 5000$ matrix we have 10,000 collisions (100 percent), for a $3 \times 3333$ matrix (removing the first digit) we have 3206 collisions ($\frac{32}{333} \approx 96$ percent), for a $4 \times 2500$ matrix we have 577 collisions ($\frac{577}{2500} \approx 23$ percent), for a $5 \times 2000$ matrix we have 48 collisions ($\frac{48}{2000} \approx 2.4$ percent), for a $6 \times 1666$ matrix (remove the first 4 digits) we have 2 collisions ($\frac{2}{1666} \approx 0.12$ percent), these two are 248248 and 886886, and for a $7 \times 1428$ (remove the first 4 digits) we have 0 collisions (0 percent).

A more interesting thing to investigate is how the frequency of the collisions increases subjected to a number of digits of $\pi$. We have calculated and plotted the frequency for a given number of digits of $\pi$ and we can see an exponentially growth. The frequency more than doubles from 10,000 digits to 20,000 digits. The frequency goes to 1 as the number of digits increases.




## The probability of getting a collision after 50 number of draws

In the Pi Matrix Challenge one randomly chooses 50 blocks to the competitor. We want to calculate the probability of getting at least one collision after 50 random draws without replacement. This can be considered as a hypergeometric distribution.

Notice that the discrete distribution with probability function

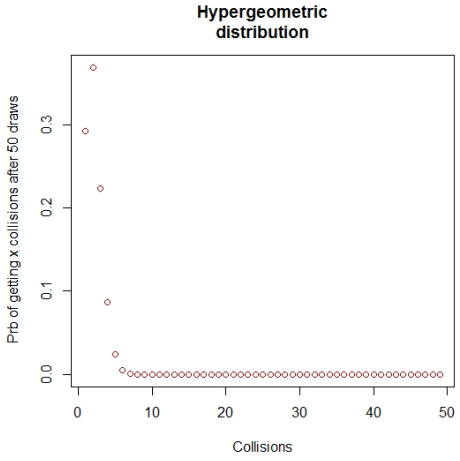
$$h(k; M, N, n) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \ , \quad k = \max\{0, n+m-N\}, ..., \min\{n, M\} \ .$$

is the *hypergeometric distribution* with parameters $M$, $N$ and $n$. The hypergeometric distribution appears in connection with drawing a sample size $n$ from a population with $N$ elements of which a subpopulation with $M$ elements is of particular interest. Then $h(k; M, N, n)$ is the probability that the sample contains precisely $k$ elements from the subpopulation of particular interest. Furthermore we have that for a fixed $\psi \in \{\max\{0, n+m-N\}, ..., \min\{n, M\}\}$

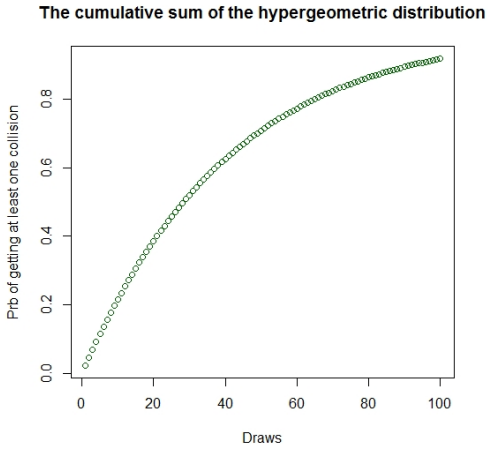
$$h(k \leq \psi, M, N, n) = \sum_{k=0}^{\psi} h(k, M, N, n) \ ,$$

$$\begin{aligned} P(\text{There exist at least one collision}) &= 1 - P(\text{There doesn't exist a collision}) \\ &= 1 - h(0, N, M, n) \approx 71 \text{ percent.} \end{aligned}$$

In other words, it's likely to get a collision during the Pi Matrix Challenge. The plot below shows us the different probabilities for a fixed draw.




The plot below shows the cumulative sum of the hypergeometric distribution. We see what the probability is about 0.71 after 50 draws.




### Conclusion

We found 48 collisions for the first 10,000 digits of $\pi$ and the positions of these. The probability of having a collision from a random sample is 2.4 percent and the probability of getting at least one collision after 50 draws without replacement is 71 percent. If the length of a block is smaller than 5 there are simply too many collisions and finally we saw that the collisions- and the frequency of the collisions rapidly increased if we have more than 10,000 digits of $\pi$.

## Appendix

### The R Code

```
# Reading data
p=read.table("C:\\Users\\Mark\\Desktop\\data10000.txt",
header=TRUE)

# Splitting the data into lines
text =
    readLines("C:\\Users\\Mark\\Desktop\\data10000.txt"
,encoding="UTF-8")

v=c(strsplit(text,"")[[1]])
# convert string v to integer v
strtoi(v, base = 10)

# Chose the number of blocks
h=2000
```

```
u=matrix(v,ncol=5,nrow=h,byrow=TRUE)

# Finding identical blocks (collisions)
ind=function(j){
e=c()
for(i in 1:h){

        if( all(u[j,] == u[i,] )==TRUE ){
                e[i]=i
        }
        else{e[i]=0}

}
e[e!=0]
}
q=c()
for(j in 1:h){
        if(length(ind(j)) > 1 ) {
        q[j]=ind(j)
        }
}

# Finding entries in q that are not NA
# Q then gives us the position of where
# a collision occurs for the first time.
Q=which(!is.na(q))
# Lenght(Q) then gives us the number of collisions

qq=matrix(,ncol=5,nrow=length(Q),byrow=TRUE)
for(k in 1:length(Q)){
        qq[k,]=u[Q[k],]
}
qq
## qq then gives is the collisions.

# Plot of the collisions
wQ=Q*5
# wQ gives the positions of the collisions

z=rep(0,wQ[1])
for(i in 2:length(Q)){
        z=c(z,rep(i-1,wQ[i]-wQ[i-1] ))
}
z

plot(z,col="brown3",xlab="Digits of Pi",ylab="Number of
   collisions",main="3000 digits")
```

```
prb=length(Q)/h
# 3000: p=0.013 - 8
# 5000: p=0.014 - 14
# 10000: p= 0.024 - 48
# 15000: p= 0.04 - 122
# 20000: p= 0.048 - 191
p=c(0.014,0.024,0.04,0.048)
# co gives the number of collisions
#co=c(14,48,122,191)
plot(c(5000,10000,15000,20000),p,type="h",
col="brown",xlab="Digits of Pi",ylab="Frquency of
    collisions")

m=48
n=2000-48
# Number of draws
k=50
# Prb of getting a collision after k draws.
1-dhyper(0,m,n,k)
# = sum(dhyper(1:k,m,n,k))
# Note that sum(dhyper(0:k,m,n,k)) = 1.

x=max(0,k-n):min(m,k)
plot(dhyper(x,m,n,k),xlab="Draw",col="red3",
ylab="Prb of getting a collision",main="Hypergeometric
    distribution")

# Plot of the cumulative sum of the hypergeometric
    distribution
plot(1 - dhyper (0 ,m ,n , k=1:100 )  , xlab =" Draws "
    , col ="darkgreen" ,
ylab =" Prb of getting at least one collision " ,
main ="The cumulative sum of the hypergeometric
    distribution")
```