# Probability of Shared Birthdays

*Patrick Kelly*

*July 26, 2015*

## Contents



# INTRODUCTION

### Question: Given 30 randomly selected singleton (no twins) people, what is the probability that at least 2 of them have the same birthday? The answer might surprise you. Salman Khan finds the likelihood of this occuring, using permutations and a virtual TI - 85 calculator.

KhanAcademy

Click here"English"

**Wikipedia also explores the problem and possible solutions.**

[Wikipedia]http://tinyurl.com/BirthdayMatches

**P(At least 2 people share a birthday) = 1 - (365! / (365-30)! / $365^{30}$) = 0.7063**

### Since R has a problem with huge factorials, this program calculates the expected mean probability and its 95% confidence interval, using another approach.

---

**Choose the numbers for people, repeats and iterations.**

```
people <- 20
repeats <- 1000
iterations <- 200
```

## You chose 20 people, 1000 repeats and 200 iterations.

**Create a vector of numbers 1 through 365 and 3 empty dataframes.**

```
days <- matrix(c(1:365))
prob1 <- data.frame()
prob.at.least.1 <- data.frame()
prob2 <- data.frame()
```

**Create a function that finds the average probability of shared birthdays for a given number of people and repeated trials.**

```
SameBirthday <- function (people, repeats) {
  for (i in 1:repeats){
    sampl <- days[sample(days, people, replace = TRUE),]
    diff.birthday <- length(unique(sampl)) < people
    prob1 <- rbind(prob1,diff.birthday)}
  trues <- sum(prob1$TRUE. == TRUE) + sum(prob1$FALSE. == TRUE)
  trials <- repeats
  options(digits=3)
  prob.at.least.1 <- trues / trials * 100
  prob.at.least.1 <<- data.frame(prob.at.least.1)
  # <<- variable is available outside of the function
  # cat("Probability of at least 1 match =",
  # prob.at.least.1,"%","(",trials,"trials )", "\n")
}
```

**Calculate the estimated average probability for 20 people in 1000 trials.**

```
SameBirthday(people,repeats)
```

## Probability of at least 1 match = 40.1% (1000 trials)

**Now create another function that iterates the first one 200 times to calculate an estimate of an overall average mean statistic of shared birthdays for 20 people, and the 95% confidence interval.**

```
OverallMean <- function(iterations) {
  for (i in 1:iterations){
    SameBirthday(people,repeats)
    prob2 <<- rbind(prob2,prob.at.least.1)}
    options(digits=3)
  mean.prob <<- mean(prob2$prob.at.least.1)
  sd.prob <- sd(prob2$prob.at.least.1)
  sem.prob <- sd.prob / sqrt(iterations)  # Standard error of the mean
  t <- qt(0.975,df = iterations-1)
  marg.error <- t * sem.prob  # Margin of error of the mean
  ci.prob <<- mean.prob + c(-marg.error, marg.error)
  cat(" Mean probability of at least 1 match = ",mean.prob,"\n")
```

```
  cat(" 95% confidence interval for the mean:","\n",
    "Lower Limit =", ci.prob[1],"      Upper Limit = ",ci.prob[2])
}
```

**Run the second function for 200 iterations.**

```
OverallMean(iterations)
```

```
##  Mean probability of at least 1 match =  41.3
##  95% confidence interval for the mean:
##  Lower Limit = 41       Upper Limit =  41.5
```

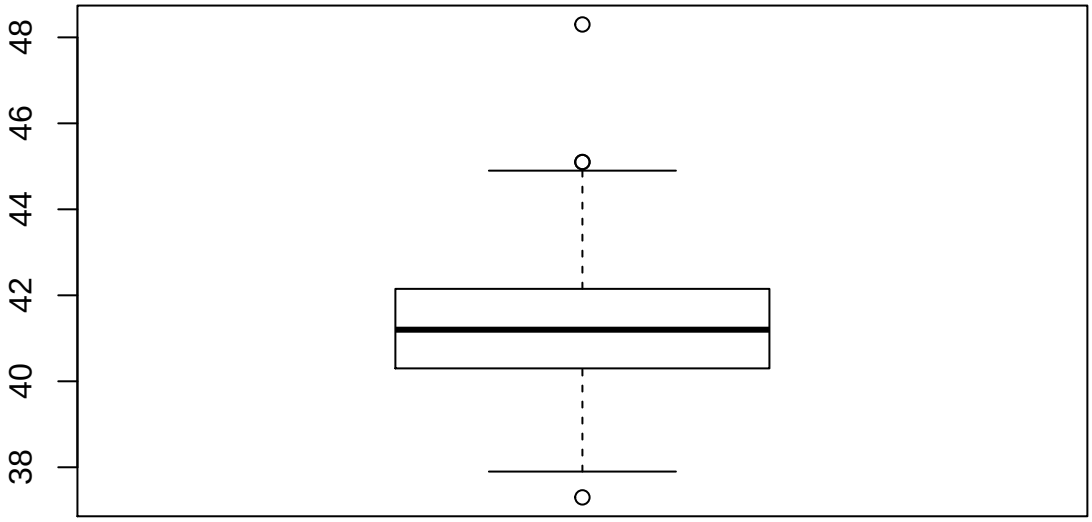
**Salman Khan's estimate for the likelihood of shared birthdays for 30 random people is 70.6%. Is this estimate included in the 95% confidence interval?**

---

## Visualization of the results

**First a boxplot, which may or may not show outliers (circles), defined by John Tukey as values more extreme than 1.5 times the Inner Quartile Range**

```
boxplot(prob2$prob.at.least.1)
```




**Finally a histogram of the frequencies with a vertical line indicating the overall mean and an overlying density plot.**

```
hist(prob2$prob.at.least.1,
  col = "grey", prob = T,
  main = "Distribution of Estimated Means",
  xlab = "% Probability of at least one match",
  sub ="Blue Line = Overall Mean")
```

```
abline(v = mean.prob, col = "blue",lwd=2)
abline(v = ci.prob[1], col = "green",lwd=2)
abline(v = ci.prob[2], col = "green",lwd=2)
lines(density(prob2$prob.at.least.1), col="red", lwd=2)
```