Estimating the Probability of the Birthday Problem Using R
Stat 341 - Fall 2008

This is the R code that looks at the probability of the birthday problem, the probability of at least 2 people in a room of n people sharing a birthday. In this handout, we will look at the problem using $n = 26$ since this was the number of people in Stat 341 on the first day we began discussion of this problem.

To begin calculating an emperical or observed probability, we need to first set up the problem. There are 365 different possible birthdays (ignoring Feb. 29th). In R, we will use the variable **days** to indicate the possible birthdays for each person.

```
days<- c(1:365)
```

In a class of $n = 26$ people, getting a random birthday for each person is like sampling from the **days** variable, with replacement. In R, this command is

```
bdays26<- sample(days, 26, replace = T)
```

Here is an example of the variable **bdays26**.

```
bdays26
214 183  52 284  77 111 116  72  73 104 342 125 184
133 240 160 321  31 263 135 149 144 284 259  66 143
```

Are any of these birthdays the same? Scanning through the list of 26 birthdays to pick matching ones is a little difficult. Since there are only 365 possible values that can be in **bdays26**, if we use a histogram with breaks on these 365 values, we can look at birthday matches by looking at the counts for the histogram. To set this up in R, the commands are

```
bdbreaks<- c(0:365) + 0.5
hist(bdays26, breaks = bdbreaks)$count
```

The counts for the histogram of the 365 days of the year for **bdays26** are

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 1 0
0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1
0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0
1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Scanning through this list, it is easy to see that each day of the year is a birthday for either 0, 1 or 2 people in the sample of 26 people. So, in this sample, we have two people who share a birthday. What if no one shared a birthday? Then the largest count for the histogram of the 365 days of the

year for **bdays26** would be 1. We can now use R to estimate the probability that the maximum count of this histogram will be two or more.

To estimate the probability of the birthday problem, we will need to use R to repeatedly select birthdays randomly for groups or samples of 26 people each. For each sample, we will look at whether or not two or more people in the sample share a birthday. The commands in R to do this loop 10,000 times are

```
maxcounts<- rep(0,10000)

for (i in 1:10000){
bdays26<- sample(days, 26, replace = T)
bdcounts<- hist(bdays26, breaks = bdbreaks, plot = F)$counts
maxcounts[i]<- max(bdcounts)}
}
```

The first line initializes the variable **maxcounts** for use in the loop. The **for** loop then repeats the pattern discussed above and saves the maximum count of the histogram of the days of the year for the **bdays26** variable.

We are interested in any value of **maxcounts** that is 2 or larger. This indicates that at least 2 people share a birthday in the sample of 26 people in the room. One way to get this information from the variable in R is to add up all the times that **maxcounts** is greater than or equal to 2. The command is

```
sum(ifelse(maxcounts >=2, 1, 0))
```

You can also have R make a table of the **maxcounts** variable. This will give you the observed values of **maxcounts** and the number of times these values occurred in the 10,000 samples. The command and the output is below.

```
table(maxcounts)
   1    2    3    4
4040 5784  175    1
```

The output of the table command is given below. You can see that in 5,960 out of the 10,000 samples $(5784 + 175 + 1)$, at least two people shared a birthday, for an emperical probability of 0.5960. This emperical probability is not exact, but thanks to the relative frequency idea of probability, this emperical probability of 0.5960 based on 10,000 samples will be close to the theoretical probability value.