Using R to Compute Probabilities of Matching Birthdays

Bruce E. Trumbo, Eric A. Suess, and Clayton W. Schupp

CSU Hayward, May 28, 2004
Statement of Birthday Problem

- 25 randomly chosen people in a room
- What is the probability two or more of them have the same birthday?
- Model simplifications for a simple combinatorial solution:
  - Ignore leap years
  - Assume all 365 days equally likely
- What error from assumptions?
Combinatorial Solution

• Let $X =$ Number of matches.
• Seek $P\{X \geq 1\} = 1 - P\{X = 0\}$

\[
P\{X = 0\} = \frac{P_{25}^{365}}{365^{25}} = \prod_{i=0}^{24} \left(1 - \frac{i}{365}\right) = 0.4313
\]
• $P\{X \geq 1\} = 1 - 0.4313 = 0.5687$
• In R: `prod(1 - (0:24)/365)` returns 0.4313003.
For $n$ people in the room

- As $n$ increases, clearly $P\{X \geq 1\}$ increases.
- If $n = 366$ (still ignoring leap years), we are sure to get at least one duplication.
- But we will see $P\{X \geq 1\}$ becomes very close to 1 for much smaller values of $n$.
- Easy to show using a loop in R.
p <- numeric(50)
for (n in 1:50) {
    q <- 1 - (0:(n - 1))/365
    p[n] <- 1 - prod(q)
}
plot(p)        # Makes Figure
p              # Makes prinout

> p
  [1] 0.0000000000 0.002739726
  ... 
[21] 0.443688335 0.475695308
[23] 0.507297234 0.538344258
[25] 0.568699704 0.598240820 
  ... 
[49] 0.965779609 0.970373580
Probabilities of Matching Birthdays in a Room
Simulating the Birthday Process

- Use R to “survey” of $m = 10000$ rooms, each with $n = 25$ randomly chosen people.

- Two convenient functions in R:
  
  ```r
  b <- sample(1:365, 25, repl=T)
  ```

  samples 25 objects from among \{1, 2, ..., 365\} with replacement,

  ```r
  length(unique(b))
  ```

  counts the number of unique objects.
n <- 25;
m <- 10000;
x <- numeric(m)
for (i in 1:m)
{
    b <- sample(1:365, n, repl=T)
    x[i] <- n - length(unique(b))
}
mean(x)
mean(x==0)
cut <- (0:(max(x) + 1)) - 0.5
hist(x, breaks=cut, freq=F, col=8)

> mean(x)
[1] 0.8081
> mean(x==0)
[1] 0.4338
Comments on Simulation Results

> mean(x==0)
[1] 0.4338

The proportion of rooms with no matches is very close to \( P\{X = 0\} = 0.4313 \).

> mean(x)
[1] 0.8081

The average number of matches per room simulates \( E(X) \), which is not easily obtained by combinatorial methods.
How Serious Are Errors From Simplifying Assumptions?

• *We hope* the 25 people in our room are randomly chosen. (For example, not a convention of twins, or Sagittarians.)

• *We know* there are Feb 29 birthdays and that other birthdays are not uniformly distributed. We hope this does not matter much.
Empirical Daily Birth Proportions: By Month Jan '97 - Dec '99
(Percent of Uniform = 1/365 Per Day)

Source: Nat'l Ctr. for Health Statistics
Simulation Easy to Generalize

Define a vector of weights:
\[
w <- \text{c(rep}(4, 200), \text{rep}(5, 165), 1)\]

More extreme variation than actual.

Sample according to these weights:
\[
sample(1:366, n, \text{repl}=T, \text{prob}=w)\]

Slight modification of the R code simulates the nonuniform problem.
n <- 25
m <- 20000
x <- numeric(m)
w <- c(rep(4,200), rep(5,165), 1)
for (i in 1:m)
{
  b <- sample(1:366, n, repl=T, prob=w)
  x[i] <- n - length(unique(b))
}
mean(x)
mean(x==0)

> mean(x)
[1] 0.819
> mean(x==0)
[1] 0.42215

Uniform Simulation

> mean(x)
[1] 0.8081
> mean(x==0)
[1] 0.4338
Concluding Remarks

• Simulation allows us to assess assumptions. Here the result was that our assumptions were fairly harmless.

• Simulating a case with a known answer builds confidence in the simulation model to be used for generalizing results.

• Using R, simulations are easy enough to use in practical applications and in the classroom.
References

[1] R software and manuals are available at
www.r-project.org


   *An Introduction to Probability Theory and Its Applications*,

[4] Birth frequencies are based on raw monthly totals available at
www.cdc.gov/nchs/products/pubd/vsus/vsus.html

[5] Thomas. S. Nunnikhoven:
   A birthday problem solution for nonuniform frequencies,