There is an experiment that I use to motivate the human tendency to dramatically underestimate the frequency of coincidences. It starts by asking how likely it is, statistically, that someone in your class had a birthday on the same day of the year as you did. It only takes 23 people to produce a sample with a better than 50% chance of having two people with the same birthday; however, enrollment is below 23 students in one Arts140 class this term; and there are *two* sets of twins in the other!

Last fall, I had three classes of 25 students -- and no twins. In each class, we checked to see if there were any people sharing a birthday. Three times, the experimental result was not a single pair.

That fact illustrates something about samples. Just as it is possible to flip a coin and come up heads ten times in a row, it's possible that three random samples of 25 Arts first-years could generate no birthdays in common. But what about the pooled sample of all three classes? A sample of 75 students (three times 25) should have at least one pair of people sharing a birthday.

Data: The raw birthday data for all students enrolled in Arts140 taught by Bloemhof came from the registrar's office in a spreadsheet. This use of data was only permitted because (1) it is for pedagogic purposes, and (2) no names or other identifiers were released (birthdays were the only information).

Method: The raw data were first truncated to remove the birth year (which is not relevant to the comparison); and then sorted from earliest in the year to latest. With so few observations (75), it is simple to count matches by looking at the sorted list. Table 1 shows when the matching birthdays are.

Table 1: Number of Birthdays by Month, Arts140 Fall 2018 (n=75, bold indicates matches present)

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
8	5	**10**	5	7	7	**8**	**6**	**5**	3	9	2

Discussion: Presumably the month with the largest number of birthdays should have a pair of people matching the date, and that does happen in this aggregated sample: there were two people celebrating their birthday on March 31. But there were no birthday matches in the second most popular month (November) and July had two instances of matching (two people on July 28th and *THREE* people on July 3rd). There were two more pairs of birthdays, August 27 and September 11, for a total of five days out of the year when two or more people in the sample share a birthday. Again, just having a high number in a particular month may not matter: in the sample, January and November have relatively high numbers of birthdays celebrated, but none of these birthdays were shared by two or more people.

Here's the proof that it takes just 23 people for a greater than 50%chance of having *some* birthday in common: (Ignore leap year, for simplicity)

(1) The theoretical number of choices of 23 birthdates is
365 x 365 x … x 365 = 365^{23} (*sampling with replacement*)

(2) The theoretical number of unique birthdates in 23 choices is
365 x 364 x 363 x … x 343 (*sampling without replacement*)

(3) Divide (2) by (1) to get p, the probability of finding 23 unique birthdates in a sample of 23.

Subtract p from 1 provides the complementary probability of at least two people with a birthday *in common*: (1 – p) = 0.507297234, or a 50.8% chance of a match in a sample of 23.