The Birthday Paradox

"The laws of probability, so true in general, so fallacious in particular."

Edward Gibbon

Introduction

To begin, I would like to begin with a question: what is the minimum number of people needed in a group for there to be more than a 50% chance of at least two people sharing the same birthday? Well, you may think that because there are 365 days in a year (not including leap days), it must be somewhere 180 people. While, on the surface level, it may seem logical, this could not be further from the truth. I will be straightforward. The answer is 23 people. It may seem rather strange and counter-intuitive at first, but when the math demonstrated, you will realize that this number is actually rather reasonable. This situation is known as the **birthday paradox**. It is not called a paradox because there is something about it that is contradictory to the natural world, it is very much naturally reasonably. The reason it is referred to as a paradox is because the conclusion of 23 people counters the human mind's innate ability at estimation and probability.

The aim of this exploration is to understand the solution behind the birthday paradox. However, because this is a very specific example, I will aim to also create an equation that will be able to approximate the probability of at least 2 people in any size group sharing the same birthday in a year of a length of any days. This equation will also approximate situations not just pertaining to people sharing birthdays. For instance, I will aim to use my approximation to predict the probability of at least two people in a table group of 4 in our math class sharing the same day of the week they were born on, as well as the probability that at least 2 people in my math class of 25 people share a birthday.

Why is the answer 23 people?

To begin, simply figuring out the probability that at least two people share the same birthday in a group is challenging, and the calculations would be tedious. This is where the **complement rule** is applied. Before explaining this, some notation needs to be explained. $P(E)$ represents the probability of an event occurring, with $P()$ denoting probability and E denoting the event. $P(E')$ is the probability of the opposite of the event occurring, or in other words, the probability of the event not occurring. The complement rule states that $P(E)+P(E')=1$ (Kernler). Translated into words, the probability of an event happening or the same event not happening must be 100%, for there is no way for something to both not occur and not **not** occur. One or the other must happen. With this in mind, we can apply this to the birthday paradox.

Let's begin by finding the probability of two people sharing the same birthday. According to the complement rule, the opposite event would be that the two people don't share a birthday. So, if $E =$ a birthday shared, then $P(E)=1-P(E')$. Starting with the first person, how many valid birth dates can the first person have while meeting the conditions? At this point, no birth dates have been taken, so the first person can have any of the 365 dates. This can be written as the fraction $\frac{365}{365}$. The second person, on the other hand, must choose any date that is not the birth date of the first person. Hence, he can choose only 364 of the 365 available dates, or in a fraction, $\frac{364}{365}$. Now that we have the possible cases of each of the two people, we can multiply them together to get the probability that the two people do not share a birthday.

$$P(E') = \frac{365}{365} \times \frac{364}{365} = \frac{132860}{133225} \rightarrow \frac{364}{365}$$
$$P(E) = 1 - \frac{364}{365} \rightarrow \frac{1}{365} \approx 0.274\%$$

Therefore, the probability of two people sharing the same birthday is about 0.274%.

If we were to calculate the probability for a group of three, the fractions for the first and second person would be the same, $\frac{365}{365}$ and $\frac{364}{365}$, respectively. The third person would have 363 out of 365 birth dates to choose, as 2 dates have already been chosen by the first and second person. The calculations, therefore, are as follows:

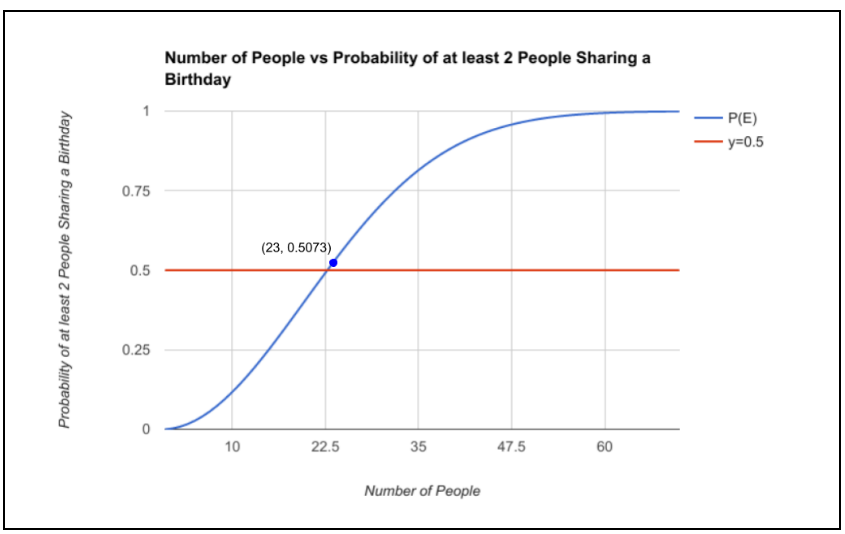
$$P(E') = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} = \frac{132132}{133225}$$
$$P(E) = 1 - \frac{132132}{133225} \approx 0.820\%$$

So, the probability of at least two people sharing the same birthday in a group of three is 0.820%.

With the two calculations of $P(E')$ shown above, we can see a pattern forming. With every new person in a group, the number of birth dates available to them decreases by $\frac{1}{365}$. Therefore, if we assign the variable p to be the number of people in a group, we can make this equation:

$$P(E') = \frac{365}{365} \times \frac{365-1}{365} \times \frac{365-2}{365} \ldots \times \frac{365-p+1}{365}$$

In this equation, the product of the numerators is the number of ways that n people in a group can all have a different birthday, while the product of the denominators is the number of total arrangements that n people in a group can have a birthday. The denominator is simply 365^n. This follows the pattern of the sample calculations shown above. Below is a graph showing the progression of $P(E)$ as the number of people, n, increases. A table is also available in appendix A.



As seen by the graph above, when there are 23 people in a group, $P(E)$ is 50.73%. When calculated this way, we can see that this problem is almost no different from any other probability problem of similar situations. It is the scope of the problem that makes our minds overestimate the solution. In fact, the number of people required for there to be at least a 99.9% chance that a birthday is shared is actually only 70 people!

This problem demonstrates something fascinating about our perception of probability. Our minds think rather linearly, and we cannot estimate things on an exponential level. This is why the everyday person is such an awful gambler. People simply cannot calculate the odds with precision, yet we still have confidence in them. The birthday paradox is a harmless example of how our minds can trick us into misjudging the natural odds because of our simple, linear thinking. However, when facing situations in which something could be lost, whether it be money or even your life, these linear estimations may be costly. The birthday paradox teaches us that we must not come to immediate assumptions in life, but to stop and think deeper and analyze a situation to truly be able to go against the odds. This is why this topic captivated me so much. It forces us to think harder at how the world works, and proves that our perception of the world is not always so accurate.

With the math behind the solution of 23 people in mind, we can create an equation that can be used to approximate any situation similar to this. The following section will discuss how this is derived.

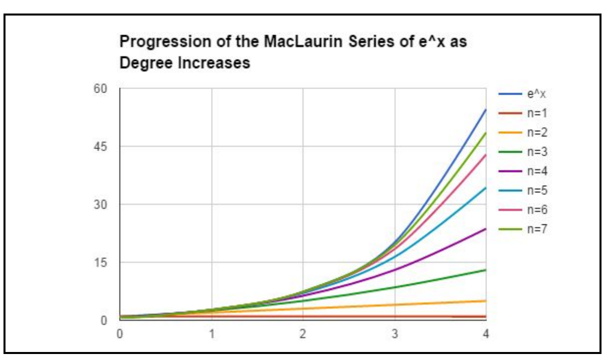
<u>An Approximation</u>

In order to understand where this approximation comes from, we must first be introduced to the **Taylor series**. A Taylor series is meant to be a representation of a function about a point $x = a$. This may be a bit confusing to understand, so I will provide the general equation (Weisstein).

$$f(x) = f(a)(x-a)^0 + \frac{f'(a)}{1!}(x-a)^1 + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \frac{f''''(a)}{4!}(x-a)^4 \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n \ldots$$

With every term added onto the series, the summation becomes closer and closer to the actual function.

When $a = 0$, we call the Taylor series for a function with this condition a **MacLaurin series**. Below is the MacLaurin Series for $f(x) = e^x$ as every nth term is added. Below is a graph of the MacLaurin Series of e^x as its degree of approximation increases. As you can, as the n increases, the degree of the series increases, as well as its accuracy at approximating the actual function.

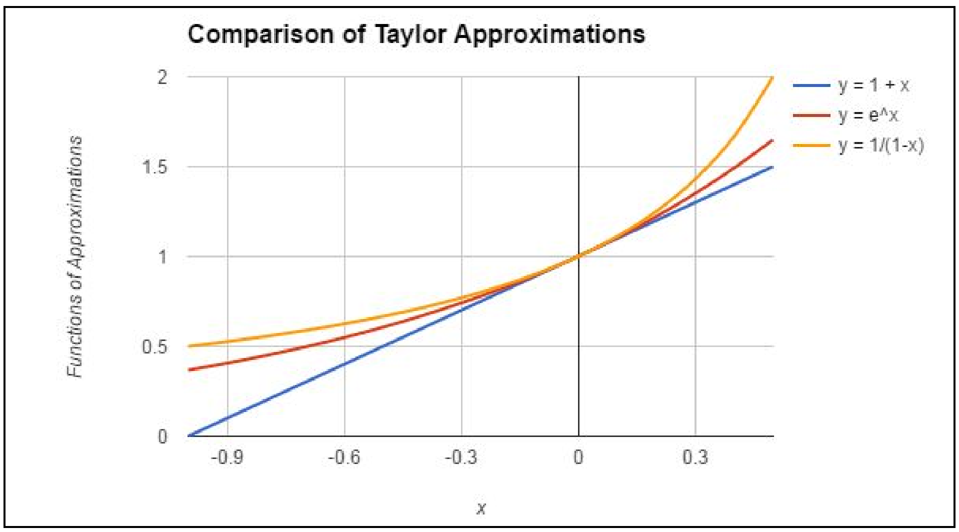


For our approximation we will be using a MacLaurin series, as the goal of our approximation is to be as simple as possible while still retaining accuracy to the 0.5%, and the MacLaurin Series usually creates the simplest form of a Taylor Series. In line with this rule of simplicity, we will keep our MacLaurin series to a **first-order approximation**, which is the sum of the first two terms of the series (Pfleuger).

There are two commonly MacLaurin series that are simple enough to be used for an approximation. They have been written out below (Girardi).

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 \ldots$$
$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} \ldots$$

When both series are turned into first order approximations, both functions are approximated down to $1 + x$. So, because both first-order approximations are identical, now we must ask: which one should be chosen? Which one is more viable? Well, we already have simplicity met as a criterion, so we need to see which one is more accurate. For this, consider the graph on the following page.



As we can see from the graph, it is very clear that $1 + x$ approximates e^x more accurately. Thus, our approximation will be:

$$e^x \approx 1 + x$$

It is important to note that past $x = -1$, this approximation will begin to become inaccurate. This will not be a worry for us, however, as we will see later on.

Recall in the previous section when we discussed how the solution of 23 people is found. The first person had a choice of 365 out of 365 days, the second person had a choice of 364 out of 365 days, all the way until the nth person has 365 - p + 1 out of 365 days. Well, the fractions $\frac{365}{365}$, $\frac{364}{365}$, and $\frac{365-p+1}{365}$ can easily be rewritten as $1 - \frac{0}{365}$, $1 - \frac{1}{365}$, and $1 - \frac{p-1}{365}$. So what is important about this? If we rewrite the fractions in this manner, it can be used in our approximation. For example, if we were to calculate percentage of days that the second person can choose out of the 365 days available, the approximation would be as follows.

$$When\ n = 2,\ \frac{365-n+1}{365} = \frac{364}{365} \rightarrow 1 - \frac{1}{365}$$
$$e^x \approx 1 + x$$

$$Therefore,\ x = -\frac{1}{365}$$
$$e^{\frac{-1}{365}} \approx 1 - \frac{1}{365}$$
$$0.997264 \approx 0.997260$$

As you can see, this approximation was very accurate, with the actual value being 99.726%, while the approximation gave 99.7264%. This was much more accurate than the aimed 0.5% accuracy.

Something also very important has also been found. Because we can approximate the percentage of birth dates each person in a group can choose from out of 365, we can approximate the $P(E')$ equation we found in the section "Why is the answer 23 people?"! We can thus generalize one person's fraction of birth dates available for them to be:

$$When\ n = a + 1,\ \frac{365 - n + 1}{365} = \frac{365 - (a + 1)\ + 1}{365} = \frac{365 - a}{365} \to 1 - \frac{a}{365}$$
$$e^x = 1 + x$$
$$Therefore,\ x = -\frac{a}{365}$$
$$e^{-\frac{a}{365}} \approx 1 - \frac{a}{365}$$

With all of this in mind, the approximation is beginning to come together. To start, let us find the approximation of the probability of 2 people sharing the same birthday in a group of 3.

$$Let\ p = 3$$
$$P(E') = (1 - \frac{0}{365})(1 - \frac{1}{365})...(1 - \frac{p-1}{365})$$
$$P(E') = (1 - \frac{0}{365})(1 - \frac{1}{365})(1 - \frac{2}{365})$$
$$P(E) = 1 - P(E')$$
$$P(E) = 0.00820417$$
$$e^{-\frac{a}{365}} \approx 1 - \frac{a}{365}$$
$$1 - (e^{-\frac{0}{365}} \times e^{-\frac{1}{365}} \times e^{-\frac{2}{365}}) \approx 1 - (1 - \frac{0}{365})(1 - \frac{1}{365})(1 - \frac{2}{365})$$
$$0.00818549 \approx 0.00820417$$

So the actual probability was about 0.8204%, and we approximated it to 0.8185%. The difference in these values is just 0.0019%! The goal of being within 0.5% of the actual probability was absolutely reached for this case. A table of actual probabilities can be found under Appendix B.

We are now able to make our general approximation. For this we will be forgoing the denominator of 365 for the variable d for number of days in the year.

$$Let\ d \neq 0$$
$$P(E') = (1 - \frac{0}{d})(1 - \frac{1}{d})(1 - \frac{2}{d})...(1 - \frac{p-1}{d})$$
$$e^{-\frac{a}{d}} \approx 1 - \frac{a}{d}$$
$$e^{-\frac{0}{d}} \times e^{-\frac{1}{d}} \times e^{-\frac{2}{d}}... \times e^{-\frac{p-1}{d}} \approx (1 - \frac{0}{d})(1 - \frac{1}{d})(1 - \frac{2}{d})... \times (1 - \frac{p-1}{d})$$
$$P(E') \approx e^{-\frac{0}{d}} \times e^{-\frac{1}{d}} \times e^{-\frac{2}{d}}... \times e^{-\frac{p-1}{d}}$$

From here, we can use one of the exponent laws to simplify the above equation down (Exponent Laws).

$$x^a x^b = x^{a+b}$$
$$Therefore,\ P(E') \approx e^{-\frac{0+1+2...+p-1}{d}}$$

The final component is to simplify our numerator. Luckily there is a way to do so. The formula for the sum of first n positive integers is:

$$\sum_{i=1}^{n} n = \frac{n(n+1)}{2}$$

Thus:

$$Let\ n = p - 1$$
$$P(E') \approx e^{-\frac{p(p-1)}{2d}}$$
$$P(E) = 1 - P(E')$$
$$P(E) = 1 - e^{-\frac{p(p-1)}{2d}}$$

And so we have derived our approximation! In order to check validity, I will be plugging in when $p = 23$ and $p = 70$, both when $d = 365$. As usual, the table of actual probabilities are in Appendix B. We should get somewhere around 0.5 and 0.999, respectively.

$$Let\ p = 23\ and\ d = 365$$
$$P(E) \approx 1 - e^{-\frac{23\times22}{730}}$$
$$P(E) \approx -.5000017\ or\ 50.00017\%$$
$$Let\ p = 70\ and\ d = 365$$
$$P(E) \approx 0.99866\ or\ 99.866\%$$

When p = 23, the approximation was off by about 0.73%, while when p = 70, it was off by only 0.02%. So while approximation does not always hit within the 0.5% margin of error that we were aiming for, it definitely is still able to create some very accurate probabilities without having to take any powers of 365, which would be ridiculous. Overall, do I believe that this approximation produces reasonable results? Given its simplicity, I would say that it definitely creates accurate results, and that it would be an equation that is appropriate to use if needed to approximate any situation similar to the birthday paradox.

Applying the Approximation to our Math Class

In this section, I will be using this approximation to find two statistics. First, the likelihood that at least two people in my math class, which as 25 people. Second, the likelihood that at least two of the four people in my table group in math class were born on the same day of the week. For the first one, we simply plug in $p = 25$ and $d = 365$.

$$P(E) \approx 1 - e^{-\frac{25\times24}{730}} \approx 0.5604$$

So it seems like the odds are in our favor that someone shares the same birthday. However, that is not the case for our class.

For the second statistic, $p = 4$, but instead of 365 days, we only have 7 days, so $d = 7$.

$$P(E) \approx 1 - e^{-\frac{4\times3}{14}} \approx 0.5756$$

Again, the odds seem to be slightly in our favor. So let's see anybody does share a birthday. The four birthdays at our table are August 16, 2000, August 26, 2000, October 28, 2000, and November 3, 2000. The day of the week these days land on are Wednesday, Saturday, Saturday, and Friday. Well, two people, in fact, do share a same "birth day of the week"! While this does not prove that the probability is about 57.6%, it does demonstrate that there is a bit of validity to this. In order to fully verify this approximation, however, it would require a very large sample size that would be very difficult to acquire.

Limitations

Of course, this approximation and the calculations made have some drawbacks. The main problem with this is that it assumes that all birthdays are equal. This, of course, is not the case. According to the New York Times, the most common birthday is September 16th, and the least common birthday is, obviously, February 29th, otherwise known as Leap Day. In order to properly calculate the probability of there being a shared birthday, the distribution of birth dates must be taken into account, which will significantly change how the probability is calculated. In addition, it does not consider Leap Day. This can be solved, if we disregard uneven birth date distribution, with changing d to 366 when using situations involving days in a year.

Real-World Applications

Surprisingly, this paradox actually has some applications in the real world. One of the main areas that it has significance in is cryptography, which is the study of writing and solving codes. It is used in something called a collision attack, which determines whether two inputs on a table containing code known as a cryptographic hash result in the same value (Batista). This is important for solving codes which involve hashes. Some other ways it is used is to evaluate the randomness of a PRNG (pseudo random number generator), and to optimize communicating nodes' power consumption in a wireless network. The latter application ensures that there are no nodes.

Conclusion

In conclusion, our aim of understanding how the birthday paradox is solved, as well as finding an accurate approximation for any situation similar to it was achieved. In terms of global perspective, I believe that this is not something that is of any historical perspective, but instead a lesson in our human intuition. We like to believe that as the most intelligent species on Earth that our assumptions are mostly correct. This birthday paradox completely shatters this expectation, and teaches us that we cannot simply rely on our minds to realize the reality of our surroundings. Despite how complex our minds are and how far we have come, we still fail to answer a question that seems as simple as the birthday paradox. We see this exploitation of our human instinct everywhere, at casinos, restaurants, even while driving. I think that the main thing that this paradox has done was humble me and how I think. It taught me to be more careful when analyzing a situation, and not to jump to conclusions so quickly. Even when you think you know the answer, the birthday paradox has taught me that it is best if you stop and think before you make a decision. Sometimes, even the human mind can falter.

Appendix A

Works Cited

Batista, Erika. "The Birthday Paradox." Expertise Economique & Juridique Des Systèmes D'Information, Dec. 2012. Web. 15 May 2017.

"Exponent Laws." *Exponent Laws -- from Wolfram MathWorld*. Wolfram Alpha, n.d. Web. 15 May 2017.

Gibbon, Edward. *Memoirs of My Life and Writings*. Place of Publication Not Identified: Tredition Gmbh, 2011. Print.

Girardi, Maria. "Commonly Used Taylor Series." University of South Carolina, n.d. Web. 14 May 2017.

Kernler, Dan. "Section 5.2: The Addition Rule and Complements." *Chapter 5*. N.p., n.d. Web. 13 May 2017.

"How Common Is Your Birthday?" *The New York Times*. The New York Times, 19 Dec. 2006. Web. 15 May 2017.

Pfleuger, Nathan. *Taylor Approximation*. Bletchley: Open UP, 1972. *Taylor Approximation*. Brown University, 7 Nov. 2014. Web. 14 May 2017.

Weisstein, Eric W. "Taylor Series." From *MathWorld*--A Wolfram Web Resource.

Appendix B

Table 1: Number of People vs Probability of at least 2 People Sharing a Birthday

Number of People	Probability of at least 2 People Sharing a Birthday
1	0
2	0.00274
3	0.0082
4	0.01636
5	0.02713
6	0.04046
7	0.05624
8	0.07434
9	0.09462
10	0.11695
11	0.14114
12	0.16703
13	0.19441
14	0.2231
15	0.2529
16	0.2836
17	0.31501
18	0.34691
19	0.37912
20	0.41144
21	0.44369
22	0.4757
23	0.5073
24	0.53834
25	0.5687
26	0.59824
27	0.62686
28	0.65446
29	0.68097
30	0.70632
31	0.73046
32	0.75335
33	0.77497
34	0.79532
35	0.81438
36	0.83218
37	0.84873
38	0.86407
39	0.87822
40	0.89123
41	0.90315
42	0.91403
43	0.92392
44	0.93289

45	0.94098
46	0.94825
47	0.95477
48	0.9606
49	0.96578
50	0.97037
51	0.97452
52	0.97808
53	0.9812
54	0.98393
55	0.98631
56	0.98837
57	0.99016
58	0.99169
59	0.99301
60	0.99414
61	0.99511
62	0.99592
63	0.99662
64	0.9972
65	0.9976
66	0.9981
67	0.99845
68	0.99873
69	0.99897
70	0.99916