The Birthday Paradox

This document contains my personal notes about the so-called "Birthday Paradox". When I first stumbled across this problem, I found it very interesting but also difficult to understand and explain to others! Moreover, there is similar problem that seems to be equivalent but in fact it isn't. Therefore, I will also concentrate on this difference and will try to show it using a simple example.

The Birthday Paradox is the following question:

What is the probability that in a given a set of n randomly chosen people, at least two of them have the same birthday?

A more specific version (but in fact the same problem) is:

How many people must a group of randomly chosen people contain, such that there is more than 50% probability that at least two of them have the same birthday?

At first, I am going to explain the Birthday Paradox, logically and mathematically. Afterwards, we will talk about the related problem in a same way. Additionally, we will look at a small example.

1. The solution to the Birthday Paradox

The will solve it step by step before deriving the general formula. Let's say the number of people in the group is denoted by **n**. We also assume that a year has *365* days, thus ignoring leap years.

n = 1:
There is only one person in the group. The person has definitely birthday on one day in the year, so we can say the probability $p_1 = 1 = \frac{365}{365}$. Well, this basic case is a bit senseless, so let's head to the next case.

n = 2:
The group consists of two persons. Person 1 has birthday at one day in a year.
Then for the second person applies:
The probability that person 2 has birthday on another day than person 1 is $\frac{364}{365}$ (as there are 364 other days in the year). So the probability that both persons have the same birthday is $p = 1 - \frac{364}{365} = \frac{1}{365}$.

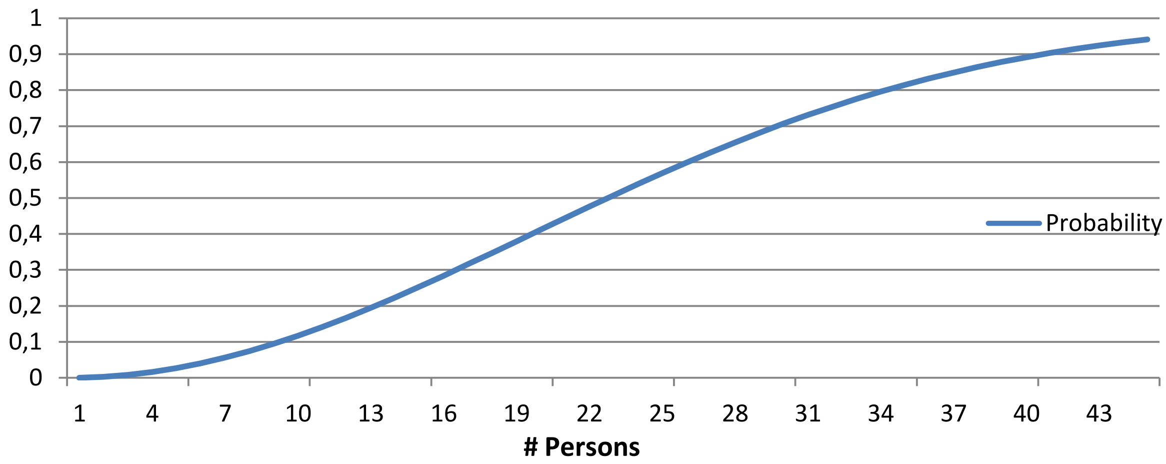
n = 3:
Here we extend the case n = 2. Again, person 1 has birthday at one day, person 2 has birthday on a different with $p = \frac{364}{365}$. So it follows that person 3 has birthday on a different day than person 1 and also on a different day than person 2 with a probability of $\frac{363}{365}$. So the probability that all three people have varying birthdays is
$p' = 1 * \frac{364}{365} * \frac{363}{365} \approx$ 99,1796%.

This implies that the probability that at least two of them have the same birthday is
p = 1 – p' = 1 – (1 * $\frac{364}{365} * \frac{363}{365}$) ≈ 0,0082%.

Arbitrary n:
So we can derive the generalization for n persons (I will also do some recasting to get a bunch of equal formulas - you can choose the most appropriate one for you) :

$$
\begin{aligned}
p_n &= 1 - \left(\frac{364}{365} * \frac{364-1}{365} * \ldots * \frac{364-n}{365}\right) \\
&= 1 - \prod_{i=1}^{n} \frac{365-i+1}{365} \\
&= 1 - \left(\frac{365-1+1}{365} * \frac{365-2+1}{365} * \ldots * \frac{365-n+1}{365}\right) \\
&= 1 - \left(\frac{365!}{365^n * (365-n)!}\right) \\
&= 1 - \left(\frac{1}{365^n} * \frac{365!}{(365-n)!}\right) \\
&= 1 - \left(\frac{n!}{365^n} * \binom{365}{n}\right)
\end{aligned}
$$

Here the diagram of the equation (The Y-axis is the probability while on the X-axis you see the number of people denoted by n):



1	0	**26**	0,59824082
2	0,00273973	**27**	0,62685928
3	0,00820417	**28**	0,65446147
4	0,01635591	**29**	0,68096854
5	0,02713557	**30**	0,70631624
6	0,04046248	**31**	0,73045463
7	0,05623570	**32**	0,75334753
8	0,07433529	**33**	0,77497185
9	0,09462383	**34**	0,79531686
10	0,11694818	**35**	0,81438324
11	0,14114138	**36**	0,83218211
12	0,16702479	**37**	0,84873401
13	0,19441028	**38**	0,86406782
14	0,22310251	**39**	0,87821966
15	0,25290132	**40**	0,89123181
16	0,28360401	**41**	0,90315161
17	0,31500767	**42**	0,91403047
18	0,34691142	**43**	0,92392286

19	0,37911853	**44**	0,93288537
20	0,41143838	**45**	0,94097590
21	0,44368834	**46**	0,94825284
22	0,47569531	**47**	0,95477440
23	0,50729723	**48**	0,96059797
24	0,53834426	**49**	0,96577961
25	0,56869970	**50**	0,97037358

So what can we learn from this result? For n = 23, the probability 50,7% - that means in a room with 23 people, the probability that two or more persons have the same birthday is bigger than 50%! In a room with 41 people, it's even bigger than 90%!

2. The intuitive trap

For most of us (including me :-) this result seems to be surprisingly low. I would have estimated a lot more people to exceed the 50% border. Other estimated even 365/2 = 182. The most frequently reason is that the actual problem is misinterpreted. Many people mix the birthday paradox with the following question which is *NOT the same*:

What is the probability that in a given a set of n randomly chosen people, at least one of them has the same birthday as a specific person, like e.g. me (or: at least one of them has birthday on a special date, e.g. 12th February)?

Another (similar) version is:

How many people must a group of randomly chosen people contain, such that there is more than 50% probability that at least one of them has birthday at the same day as person x (e.g. me)?

Compare those versions with the correct birthday paradox in order to understand the small, but far-reaching difference!

So let's solve this one. We want to derive the probability for n people that at least two of them have birthday on the same special, predefined date; called 'special date'.

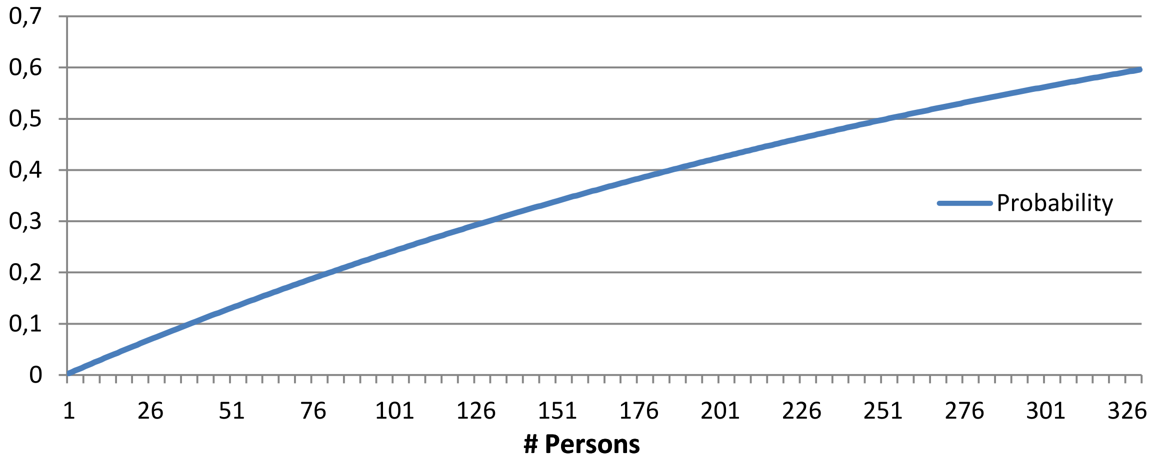
n = 1:
The first person has birthday at one date. The probability that this person has *NOT* birthday on that 'special date' is p' = $\frac{364}{365}$. The other way, the probability that he/she has birthday on that given 'special date', is p = 1 – p' = 1 - $\frac{364}{365} = \frac{1}{365}$.

n = 2:
The first person has NOT birthday on that 'special date' with p'' = $\frac{364}{365}$; the same applies for the second person. This results that both persons have not birthday on a given 'special date' with probability p' = p'' * p'' = $\left(\frac{364}{365}\right)^2$. So the wanted probability that at least one of them has birthday on that special date is therefore p = 1 – p' = 1 - $\left(\frac{364}{365}\right)^2$.

Using that, we can state the generalized formula for **n**:

$$p_n = 1 - \left(\frac{364}{365}\right)^n$$

So for what value of n does this probability exceeds 50%? In fact that is for n >= 253! So there must be at least 253 persons in a room that the probability exceeds 50% that at least two of them has birthday in a predefined special date.



10	0,02706194	**160**	0,35529198
20	0,05339153	**170**	0,37273903
30	0,0790086	**180**	0,38971393
40	0,10393241	**190**	0,40622946
50	0,12818174	**200**	0,42229804
60	0,15177484	**210**	0,43793178
70	0,17472946	**220**	0,45314244
80	0,19706288	**230**	0,46794147
90	0,21879192	**240**	0,48234000
100	0,23993293	**250**	0,49634889
110	0,26050182	**260**	0,50997866
120	0,28051407	**270**	0,52323959
130	0,29998476	**280**	0,53614166
140	0,31892853	**290**	0,54869456
150	0,33735965	**300**	0,56090776

3. Emphasizing the difference of both cases

Even at this point it may seem not clear why there is such a different result for the two quite similar problems. For this reason, we can transform both to a dice game – as there are less combinations possible, the difference can be explained easier and well-defined. We assume that each person in a group of people throw a dice exactly once – thus can throw a number between 1 and 6. Then we compare both questions for a group of 3 people.

a) The birthday paradox transformed to a dice game

What is the probability that in a given a set of n randomly chosen people, at least two of them throw the same number?

n = 1: There is only one person, and this person throws any number with p = 1.

n = 2: The second person throws with p' = $\frac{5}{6}$ another number, thus with p = 1 – p' = 1 - $\frac{5}{6}$ = $\frac{1}{6}$ ≈ 16,67% the same number as person 1.

n = 3: The third person throws with p' = $\frac{4}{6}$ another number than person 1 and person 2. (The other way round, the third person throws with probability $\frac{1}{6}$ the same number as person 1, and with probability $\frac{1}{6}$ the same number as person 2).
It follows that the probability that all 3 persons throw distinct numbers is p' = 1 * $\frac{5}{6} * \frac{4}{6} = \frac{20}{36}$ ≈ 55,55%. So the wanted probability that at least two people toss the same number is therefore p = 1 – p' = 1 - $\frac{20}{36} = \frac{16}{36}$ ≈ 44,45%.

This last probability can also be calculated in another way:

Probability that at least two people toss the same number
= probability all three persons throw the same number
+ probability two persons throw the same number
= $\frac{6}{216} + \frac{90}{216} = \frac{96}{216}$ ≈ 44,45%.

From where the numbers 90 and 6 come from you ask? Well, as stated above, the number of possible combinations is manageable; let's look at these combinations:

111	**121**	**131**	**141**	**151**	**161**
112	**122**	132	142	152	162
113	123	**133**	143	153	163
114	124	134	**144**	154	164
115	125	135	145	**155**	165
116	126	136	146	156	**166**
211	**221**	231	241	251	261
212	**222**	**232**	**242**	**252**	**262**
213	**223**	**233**	243	253	263
214	**224**	234	**244**	254	264
215	**225**	235	245	**255**	265
216	**226**	236	246	256	**266**
311	321	**331**	341	351	361
312	**322**	**332**	**342**	**352**	**362**
313	323	**333**	343	353	363
314	324	**334**	**344**	354	364
315	325	**335**	345	**355**	365
316	326	**336**	346	356	**366**
411	421	431	**441**	451	461
412	**422**	432	**442**	452	462
413	423	**433**	**443**	453	463
414	**424**	**434**	**444**	**454**	**464**
415	425	435	**445**	**455**	465
416	426	436	**446**	456	**466**
511	521	531	541	**551**	561
512	**522**	532	542	**552**	562
513	523	**533**	543	**553**	563
514	524	534	**544**	**554**	564

515	**525**	**535**	**545**	**555**	**565**
516	526	536	546	**556**	**566**
611	621	631	641	651	**661**
612	**622**	632	642	652	**662**
613	623	**633**	643	653	**663**
614	624	634	**644**	654	**664**
615	625	635	645	**655**	**665**
616	**626**	**636**	**646**	**656**	**666**

These are all combinations for three people, each throwing a dice once. E.g. 315 means that person 1 throws a '3', person 2 a '1' and person 3 toss number '5'. In total, there are 216 different combinations as seen above. There are 6 combinations where all three people throw the same number and 90 combinations where two people out of the three throw a same number, so in total 96 as also our calculation above state.

b) The 'intuitive approach' transformed to a dice game

Now we consider the similar question I called the 'intuitive approach' from chapter 2 in relation to the dice game.

What is the probability that in a given a set of n randomly chosen people, a person throws the same number as the first person?

n = 1: There is only one person, and this person throws any number with p = 1.

n = 2: The second person throws with p' = $\frac{5}{6}$ another number, thus with p = 1 – p' = 1 - $\frac{5}{6}$ = $\frac{1}{6}$ ≈ 16,67% the same number as person 1.

n = 3: Here comes the difference.
Person 3 throws with p = $\frac{1}{6}$ the same number as person 1; with p'' = $\frac{5}{6}$ another number.
Thus the probability that neither person 2 nor person 3 throws the same number as person 1 is p' = 1 * p'' * p'' = 1 * $\frac{5}{6}$ * $\frac{5}{6}$ = $\frac{25}{36}$ ≈ 69,44%. The other way round, the probability that either person 2 or person 3 do toss the same number as person 1 is then p = 1 – p' = $\frac{11}{36}$ ≈ 30,56%.

This last probability can also be calculated in another way:

Probability that someone throws the same number as Person 1
= Person 2 throws the same number as person 1 (*)
+ Person 3 throws the same number as person 1 (**)
- Person 2 throws the same number as person 3 (***)
= $\frac{1}{6}$ + $\frac{1}{6}$ - $\frac{1}{36}$ = $\frac{11}{36}$ ≈ 30,56%.

Again, we can have a look at all combinations. The green combinations show the cases where either person 2 or person 3 throws the same number as person 1. The blue cases are somehow special as all three persons toss the same number. In this case, this event is considered twice in the above formula – in (*) and (**) – although it is just one event. That's why we subtract this case again in (***).

111	**121**	**131**	**141**	**151**	**161**
112	122	132	142	152	162
113	123	133	143	153	163
114	124	134	144	154	164
115	125	135	145	155	165
116	126	136	146	156	166
211	**221**	231	241	251	261
212	**222**	**232**	**242**	**252**	**262**
213	**223**	233	243	253	263
214	**224**	234	244	254	264
215	**225**	235	245	255	265
216	**226**	236	246	256	266
311	321	**331**	341	351	361
312	322	**332**	342	352	362
313	**323**	**333**	**343**	**353**	**363**
314	324	**334**	344	354	364
315	325	**335**	345	355	365
316	326	**336**	346	356	366
411	421	431	**441**	451	461
412	422	432	**442**	452	462
413	423	433	**443**	453	463
414	**424**	**434**	**444**	**454**	**464**
415	425	435	**445**	455	465
416	426	436	**446**	456	466
511	521	531	541	**551**	561
512	522	532	542	**552**	562
513	523	533	543	**553**	563
514	524	534	544	**554**	564
515	**525**	**535**	**545**	**555**	**565**
516	526	536	546	**556**	566
611	621	631	641	651	**661**
612	622	632	642	652	**662**
613	623	633	643	653	**663**
614	624	634	644	654	**664**
615	625	635	645	655	**665**
616	**626**	**636**	**646**	**656**	**666**

This dice example is by far not perfect, but maybe you can see from my explanations why even for n = 3, the probability for the ‘intuitive trap’ does increase slower than in case a). Therefore you need also more people to exceed the 50% border compared to case a) - the equivalent of the birthday paradox – in which the probability increases faster.

Summary:

I hope I could help some of you on their way to understand the Birthday Paradox and why it seems that strange at the beginning. I tried my best, but I am unsure if I did not make any mistakes – if so, please correct me!

Sunshine, April 2010

www.sunshine2k.de
www.bastian-molkenthin.de

Mail: webmaster_(@)_sunshine2k.de