birthday problem - expected number of collisions

There are many descriptions of the "birthday problem" on this site — the problem of finding the probability that in a group of n people there will be any (= at least 2) sharing a birthday.

I am wondering how to find instead the expected number of people sharing a birthday in a group of n people. I remember that expectation means the weighted sum of the probabilities of each outcome:

$$E[X] = \sum_{i=0}^{n-1} x_i p_i$$

And here x must mean the number of collisions involving $i+1$ people, which is $\binom{n}{i}$. All n people born on different days means no collisions, $i = 0$; two people born on the same day means n collisions, $i = 1$; all n people born on the same day means n collisions, $i = n-1$.

Since the probabilities of three or more people with the same birthday are vanishingly small compared to two people with the same birthday, and decreases faster than x increases, is it correct to say that this expectation can be approximated by

$$E[X] \approx \binom{n}{0} p_{no\ collisions} + \binom{n}{1} p_{one\ collision}$$

This doesn't look right to me and I'd appreciate some guidance.

Sorry - edited to change $\binom{n}{1}$ to $\binom{n}{0}$ in second equation. Sloppy of me.

(probability) (calendar-computations)

edited Apr 29 '11 at 7:49

asked Apr 29 '11 at 7:32

brannerchinese
350 3 12

Just because there is a collision of five people does not mean that there is not also a collision of three other people, do you count this with 8? with 5?. Also, how do you avoid counting the collision of four people among the five people a second time? In other words, define p_i, explain what you actually want to count and then try to justify your formula for the expectation. – Phira Apr 29 '11 at 7:44

@user9325: I would say a collision with 5 people should mean with exactly 5 people; a collision with 3 people would have a different probability and be counted as a different term. – brannerchinese Apr 29 '11 at 7:49

4 Again, you have 3 people who have birthday on May 1st, 5 people who have birthday on September 20, and 1 other person. What is the value of X in this case? 3,5,8, 30 ? Note that the term 30 comes from counting *all* "collisions" number of 2-collisions, 3-collisions, etc. So you should not tell me that something "contributes another term", you should first tell me what you *want* to count. – Phira Apr 29 '11 at 9:03

@user9325:Now I see your point. I actually want to know the number of people involved in any collision, and I see that what I have written will not do. – brannerchinese Apr 29 '11 at 13:07

3 Answers

The probability person B shares person A's birthday is $1/N$, where N is the number of equally possible birthdays,

so the probability B does not share person A's birthday is $1 - 1/N$,

so the probability $n-1$ other people do not share A's birthday is $(1 - 1/N)^{n-1}$,

so the expected number of people who do not have others sharing their birthday is $n(1 - 1/N)^{n-1}$, so the expected number of people who share birthdays with somebody is $n\left(1-(1-1/N)^{n-1}\right)$.

answered Apr 29 '11 at 9:09

Henry
71.8k 3 41 106

Beautiful in its clarity. Thank you. – brannerchinese Apr 29 '11 at 13:15

I wrote a simulation and ran several million trials using various N and n; the results are within .001n of what your formula predicts. Thanks again. – brannerchinese May 5 '11 at 18:05

I would like to be able to cite your help in the paper I am writing (about philology, not birthdays). Would you mind to look me up at brannerchinese.com and contact me off-list? There is no regular private-messaging function on the SE site (meta.math.stackexchange.com/q/632/9263) and I can see no other non-public means to ask you for a name by which I can acknowledge your help. I understand if you prefer to remain anonymous or "Henry". – brannerchinese May 6 '11 at 11:38

I will try to get control of the most standard interpretation of our question by using (at first) very informal language. Let us call someone *unhappy* if one or more people share his/her "birthday." We want to find the "expected number" of unhappy people.

Define the random variable X by saying that X is the number of unhappy people. We want to find $\mathrm{E}(X)$. Let p_i be the probability that $X=i$. Then

$$\mathrm{E}(X)=\sum_{i=0}^{n} i\,p_i$$

That is roughly the approach that you took. That approach is correct, and a very reasonable thing to try. Indeed have been *trained* to use this approach, since that's exactly how you solved the exercises that followed the definition of expectation.

Unfortunately, in this problem, finding the p_i is very difficult. One could, as you did, decide that for a good approximation, only the first few p_i really matter. That is sometimes true, but depends quite a bit on the values N of "days in the year" and the number n of people.

Fortunately, in this problem, and many others like it, there is an alternative *very* effective approach. It involves a bit of theory, but the payoff is considerable.

Line the people up in a row. Define the random variables $U_1, U_2, U_3, \dots, U_n$ by saying that $U_k=1$ if the k-th person is unhappy, and $U_k=0$ if the k-th person is not unhappy. The crucial observation is that

$$X=U_1+U_2+U_3+\cdots+U_n$$

One way to interpret this is that you, the observer, go down the line of people, making a tick mark on your tally sheet if the person is unhappy, and making no mark if the person is not unhappy. The number of tick marks is X, the number of unhappy people. It is also the sum of the U_k.

We next use the following very important theorem: **The expectation of a sum is the sum of the expectations**. This theorem holds "always." The random variables you are summing *need not be independent*. In our situation, the U_k are not independent, but, for expectation of a sum, that does not matter. So we have

$$\mathrm{E}(X)=\mathrm{E}(U_1)+\mathrm{E}(U_2)+\mathrm{E}(U_3)+\cdots+\mathrm{E}(U_n)$$

Finally, note that the probability that $U_k=1$ is, as carefully explained by @Henry, equal to p, where

$$p=1-(1-1/N)^{n-1}$$

It follows that $\mathrm{E}(U_k)=p$ for any k, and therefore $\mathrm{E}(X)=np$.

edited Apr 29 '11 at 18:40

answered Apr 29 '11 at 17:08

André Nicolas
419k 32 358 701

Thanks very much. – brannerchinese Apr 30 '11 at 0:16

@user6312, any pointers on finding the probability that k people share the same birthday? – user4143 Apr 30 '11 at 1:58

@user6312: I'm grateful for this patient contextualization of @Henry's answer. – brannerchinese May 5 '11 at 18:06

This the approach the professor will expect to see in the test. Thanks. – vincent mathew Sep 25 '14 at 19:03

The following approximation may be useful.

If there are k people and N possible birthdays (or in case of a hash table, k items being hashed into N buckets), then the expected number of people/items that collide with at least one of the others is exactly (see Henry's answer or André Nicolas's answer)

$$\begin{aligned} k\left(1-\left(1-\frac{1}{N}\right)^{k-1}\right) \\ = \frac{k(k-1)}{N} - \frac{k(k-1)(k-2)}{2N^2} + O\left(\frac{1}{N^3}\right) \\ \approx \frac{k^2}{N}. \end{aligned}$$

The above is one possible definition of "expected number of collisions". If there are r birthdays/buckets each with two people/items in them, the above expression gives count $2r$, as it counts each member of each pair. If instead you want to count the number of buckets/birthdays that have multiple people in them, then the answer is approximately

$$\approx \frac{k^2}{2N}.$$

This result can be derived either

- from the previous analysis, by noting that to the first order the most common type of collision is to have 2 in a bucket (3-way and higher collisions will be statistically rare), so you just halve the count;
- or, by doing a similar analysis focusing on birthdays/buckets: the probability that either 0 or 1 of the k people have that particular birthday is

  $$\left(1-\frac{1}{N}\right)^k + k\frac{1}{N}\left(1-\frac{1}{N}\right)^{k-1}$$

  So the expected number of buckets with multiple values in them is

  $$\begin{aligned} N\left(1-\left(1-\frac{1}{N}\right)^k - k\frac{1}{N}\left(1-\frac{1}{N}\right)^{k-1}\right) \\ = \frac{k(k-1)}{2N} - \frac{k(k-1)(k-2)}{3N^2} + O\left(\frac{1}{N^3}\right) \\ \approx \frac{k^2}{2N}. \end{aligned}$$

edited Dec 9 '13 at 8:27

answered Dec 6 '13 at 9:55

ShreevatsaR
29.2k 4 58 85