Случайные совпадения

Если вы уже изучали теорию вероятностей, то скорее всего знакомы с парадоксом дней рождения. Но могу заверить, что вы все равно найдете что-то новое для себя в этой статье.

Представьте, что вы находитесь в комнате с 23 людьми. Как вы считаете, какова вероятность того, что хотя бы у двух из присутствующих совпадут дни рождения? Интуитивно кажется, что она не будет столь высока, но она же составляет чуть более 50%.

Давайте сначала строго сформулируем поставленную задачу:

Задача: В комнате находится n человек. Какова вероятность того, что хотя бы у двух из них совпадут дни рождения? *Примечание: в году 365 дней и каждый день рождения равновероятен для каждого человека.*

Введем пространство элементарных исходов Ω. Каждый исход пускай будет представлять упорядоченный набор дней рождения всех n человек:

$$\Omega = \{(d_1, d_2, \dots, d_n) \mid d_i \in \{1, 2, \dots, 365\}\}$$

Прекрасно! По нашей формулировке каждый исход в Ω равновероятен.

Введем искомое событие A: у хотя бы двух человек совпадут дни рождения:

$$A = \{(d_1, d_2, \dots, d_n) \in \Omega \mid \exists i \neq j \text{ такие что } d_i = d_j\}$$

Здесь может возникнуть проблема, так как не совсем понятно, как посчитать $P(A)$. Поэтому давайте посчитаем вероятность противоположного события $\bar{A}$ - у всех людей разные дни рождения:

$$\bar{A} = \{(d_1, d_2, \dots, d_n) \in \Omega \mid \forall i \neq j \; d_i \neq d_j\}$$

Хорошо. Если ни у кого не совпадают дни рождения, то первый человек может родиться в любой из 365 дней, второй из уже оставшихся 364 и т.д.:

$$P(\bar{A}) = \frac{|\bar{A}|}{|\Omega|} = \frac{365 \cdot 364 \cdot \dots \cdot (365 - n + 1)}{365^n}$$

Следовательно, вероятность того, что хотя бы у двух человек совпадут дни рождения:

$$P(A) = 1 - P(\bar{A}) = 1 - \frac{365 \cdot 364 \cdot \dots \cdot (365 - n + 1)}{365^n}$$

Тем самым мы получили точную формулу. Давайте попробуем в нее подставить $n = 23$:

$$P(A) = 1 - \frac{365 \cdot 364 \cdot \dots \cdot (365 - 23 + 1)}{365^{23}} \approx 0.5073$$

Как вы могли заметить, руками не столь приятно вычислять такие произведения.

Давайте обобщим задачу:

Пусть у нас есть n объектов и N равновероятных значений. Какова вероятность того, что хотя бы два объекта попадут в один и тот же исход?

Решим аналогично. Введем событие B: хотя бы два объекта попадут в один и тот же исход. Теперь введем противоположное событие $\bar{B}$: все объекты попали в разные исходы.

Тогда:

$$P(\bar{B}) = \frac{N \cdot (N-1) \cdot \dots \cdot (N-n+1)}{N^n}$$

Следовательно:

$$P(B) = 1 - P(\bar{B}) = 1 - \frac{N \cdot (N-1) \cdot \dots \cdot (N-n+1)}{N^n}$$

Прекрасно. Получили то же самое ахах. Теперь разделим каждый множитель числителя на свой N из знаменателя:

$$P(\bar{B}) = \frac{N}{N} \cdot \frac{N-1}{N} \cdot \dots \cdot \frac{N-n+1}{N} = \prod_{k=0}^{n-1} \left(1 - \frac{k}{N}\right)$$

Но по пути докажем простую лемму:

- $1 - x \le e^{-x} \quad \forall x \in \mathbb{R}$ - ?
- $1 - x \ge e^{-x-x^2} \quad \forall x \in [0, \frac{1}{2}]$ - ?

1. Пусть $f(x) = e^{-x} - (1-x)$

   Тогда $f(0) = 0$ и $f'(x) = -e^{-x} + 1$

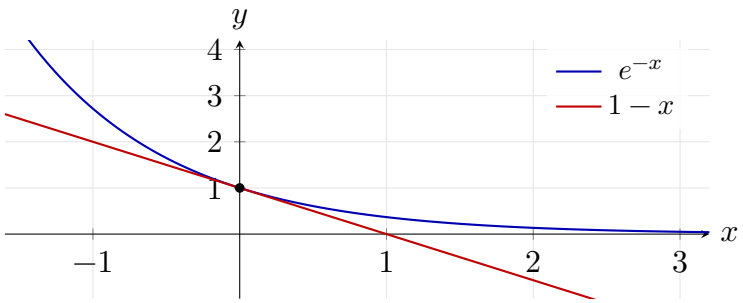
   - Если $x > 0$, то $e^{-x} < 1$, потому $f'(x) > 0$, т.е. ф-ия возрастает
   - Если $x < 0$, то $e^{-x} > 1$, потому $f'(x) < 0$, т.е. ф-ия убывает.

   Слева от нуля ф-ия убывает, справа растет. $\implies$ в нуле минимум

   $\implies f(x) \ge f(0) = 0 \implies e^{-x} - (1-x) \ge 0 \implies 1 - x \le e^{-x}$

   ч.т.д.

   (вообще это просто касательная)



2. Хотим доказать $1 - x \geq e^{-x-x^2}$

   Возьмем лонарифм:

$$\ln(1-x) \geq -x - x^2 \iff -\ln(1-x) \leq x + x^2$$

   Теперь вспомним ряд Тейлора:

$$-\ln(1-x) = x + \frac{x^2}{2} + \frac{x^3}{3} + ...$$

   Очев:

$$x + \frac{x^2}{2} + \frac{x^3}{3} + ... \leq x + \frac{x^2}{2} + \frac{x^3}{2} + ...$$

   Где вынесем за скобку $\frac{x^2}{2}$:

$$= x + \frac{x^2}{2}\left(1 + x + x^2 + x^3 + ...\right)$$

   В скобках получилась беск. геом. прогрессия. Ее сумма равна $\frac{1}{1-x}$

$$= x + \frac{x^2}{2} \cdot \frac{1}{1-x}$$

   Теперь вспомним, что $x \leq \frac{1}{2}$ по условию, тогда $(1-x) \geq \frac{1}{2}$, следовательно:

$$\frac{x^2}{2(1-x)} \leq \frac{x^2}{2 \cdot \frac{1}{2}} = x^2$$

   Итого:

$$-\ln(1-x) \leq x + \frac{x^2}{2(1-x)} \leq x + x^2$$

   (ряд Тейлора здесь корректно использовать, так как $\ln(1+x)$ *сходится на* $(-1, 1]$*)*

   ч.т.д.

Вернемся к нашей изначальной задаче! Напомню:

$$P(\bar{B}) = \prod_{k=0}^{n-1} \left(1 - \frac{k}{N}\right)$$

Обозначим $x = \frac{k}{N}$, тогда по доказанной лемме:

$$e^{-\frac{k}{N} - \frac{k^2}{N^2}} \leq 1 - \frac{k}{N} \leq e^{-\frac{k}{N}}$$

Подставим в произведение:

- Левая граница:

$$\prod_{k=0}^{n-1} e^{-\frac{k}{N}-\frac{k^2}{N^2}} = \prod_{k=0}^{n-1} e^{-\frac{k}{N}} \cdot \prod_{k=0}^{n-1} e^{-\frac{k^2}{N^2}} = e^{-\sum_{k=0}^{n-1}\frac{k}{N}} \cdot e^{-\sum_{k=0}^{n-1}\frac{k^2}{N^2}} = e^{-\frac{n(n-1)}{2N}} \cdot e^{-\sum_{k=0}^{n-1}\frac{k^2}{N^2}}$$

Сумма квадратов: $\sum_{k=0}^{n-1} k^2 = \frac{(n-1)n(2n-1)}{6} \le \frac{n^3}{3}$, потому:

$$\ge e^{-\frac{n(n-1)}{2N}} \cdot e^{-\frac{n^3}{3N^2}}$$

- Правая граница:

$$\prod_{k=0}^{n-1} e^{-\frac{k}{N}} = \exp\left(-\sum_{k=0}^{n-1}\frac{k}{N}\right) = e^{-\frac{n(n-1)}{2N}}$$

Итого:

$$e^{-\frac{n(n-1)}{2N}} \cdot \mathbf{e}^{-\frac{\mathbf{n^3}}{\mathbf{3N^2}}} \le P(\bar{B}) \le e^{-\frac{n(n-1)}{2N}}$$

Так мы и вывели с вами приблизительную формулу для вычисления наших вероятностей.

Можно еще заметить, что нижняя граница отличается от верхней только одним множителем. В целом, его можно назвать штрафным коэффициентом.

(Стоит отметить, в лемме мы одно неравенство доказали только $\forall x \in [0, \frac{1}{2}]$, поэтому данная оценка корректна только при $n \le \frac{N}{2} + 1$)

$\sqrt{N}$

Теперь найдем при каком n вероятность совпадения станет равна $\frac{1}{2}$. Тут уже воспользуемся выведенным приближением:

$$e^{-\frac{n^2}{2N}} = \frac{1}{2} \quad \Longrightarrow \quad n \approx \sqrt{2\ln 2 \cdot N} \approx 1.18\sqrt{N}$$

$e^x - ?$

Сейчас мы повыводили разные формулы. Но что за ними вообще стоит кроме алгебраических преобразований?

Пока взглянем на число, которое стоит в показателе экспоненты:

$$\frac{n(n-1)}{2N}$$

Его мы получили, сложив $\frac{k}{N}$ по всем k от 0 до $n-1$.

Ненадолго вернемся к произведению и нашей изначальной задаче:

$$P(\bar{B}) = \prod_{k=0}^{n-1} \left(1 - \frac{k}{N}\right)$$

Теперь представим, что люди входят в комнату по одному:

- Вошел первый человек. В комнате никого нет, поэтому его день рождения ни с кем совпасть не может, потому множитель равен $1 - \frac{0}{N} = 1$
- Входит второй человек. В комнате уже есть один присутствующий и второму нельзя родиться с ним в один день $\implies$ свободных дней остается $N - 1$ $\implies$ множитель равен $1 - \frac{1}{N}$
- и тд. надеюсь, логика понятна.

То бишь, если в комнате уже находится k человек с k различными днями рождения, новый не должен совпасть ни с одним из имеющихся $\implies$ множитель равен $1 - \frac{k}{N}$

$\implies$ каждая пара людей проверяется ровно один раз, когда приходит новенький.

Теперь посмотрим еще раз на сумму в показателе:

$$\frac{0 + 1 + 2 + ... + (n-1)}{N}$$

Здесь числитель по сути просто складывает пары людей, которые не должны совпасть друг с другом.

Потому мы можем переписать показатель так: :)

$$\frac{n(n-1)}{2N} = C_n^2 \cdot \frac{1}{N}$$

Первый множитель C_n^2, как мы догадались, это всевозможные пары людей. А вот $\frac{1}{N}$ - вероятность того, что у одной конкретной пары людей совпадут дни рождения.

Чтобы это доказать, возьмем двух людей с номерами i и j и посчитаем исходы $(d_1, ..., d_n) \in \Omega$, в которых $d_i = d_j$. Общий день для этих двоих можно выбрать N способами, а дни рождения остальных $n - 2$ человек любые, это ещё N^{n-2} способов. Поэтому

$$P(d_i = d_j) = \frac{N \cdot N^{n-2}}{N^n} = \frac{1}{N}$$

Небольшой итог

$$\frac{n(n-1)}{2N} = \underbrace{C_n^2}_{\text{число пар}} \cdot \underbrace{\frac{1}{N}}_{\text{вероятность совпадения пары}}$$

Здесь и кроется разгадка парадокса, как мне кажется. Ибо когда мы впервые сталкиваемся с этой задачей, мы интуитивно можем думать о вероятности совпадения, например, своего дня рождения с кем-то. И правда, такая вероятность не столь и высока. Если в комнате 23 человека, то всевозможных пар именно с тобой - 22.

Но совпасть же могут совершенно **любые двое!** А всего пар среди 23 людей 253:

$$253 \cdot \frac{1}{365} \approx 0.693 \approx \ln 2$$

Число людей растет линейно, когда же число пар растет квадратично. Вот так вот.

Среднее число совпадений

Окей, показатель экспоненты - это сумма по всем парам на вероятность совпадения.

Давайте определим случайную величину. Напомню, что мы работаем на:

$$\Omega = \{(d_1, \dots, d_n) \mid d_i \in \{1, \dots, N\}\}$$

где $P(\omega) = \frac{1}{N^n} \quad \forall \omega \in \Omega$

Введем:

$$X(d_1, \dots, d_n) = |\{(i,j) \mid i < j,\ d_i = d_j\}|$$

— число пар людей с общим днём рождения.

Ну и событие, которое мы так долго мучили, можно будет обозначить теперь так:

$$\bar{B} = \{X = 0\}$$

Теперь для каждой пары $i < j$ введём индикатор её совпадения:

$$I_{ij} = \mathbb{1}\{d_i = d_j\}$$

Тогда число совпавших пар раскладывается в сумму индикаторов:

$$X = \sum_{i<j} I_{ij}$$

По линейности:

$$\mathbb{E}X = \sum_{i<j} \mathbb{E}I_{ij} = \sum_{i<j} P(d_i = d_j) = C_n^2 \cdot \frac{1}{N} = \frac{n(n-1)}{2N}$$

Тот же самый показатель экспоненты btw)))

Тогда наша оценка сверху, которую мы вывели, приобретает достаточно красивый вид!

$$P(\text{все значения различны}) \leq e^{-\mathbb{E}X}$$

Я считаю, це изумительно.

Событие же "есть совпадение" это $B = \{X \geq 1\}$. Тогда

$$\mathbb{E}X = \sum_{k \geq 0} k \cdot P(X = k) \geq \sum_{k \geq 1} 1 \cdot P(X = k) = P(X \geq 1) = P(B)$$

выбросили слагаемое с $k = 0$ и заменили каждое $k \geq 1$ единицей

Объединяя это с нашей первой оценкой, получаем, что для **любых** n и N, обозначив $\lambda = \mathbb{E}X = \frac{n(n-1)}{2N}$,

$$1 - e^{-\lambda} \leq P(B) \leq \lambda$$

Почему вообще у нас возникла зависимость от экспоненты и от среднего?

Распределение Пуассона

Чтобы понять, откуда берётся $e^{-\mathbb{E}X}$, ненадолго забудем про дни рождения и рассмотрим модель попроще.

Независимые испытания

Напомню определение, которое нам понадобится:

Определение. События A и B называются *независимыми*, если

$$P(A \cap B) = P(A)\, P(B)$$

Аналогично если есть несколько событий:

$$P(A_{i_1} \cap \ldots \cap A_{i_r}) = P(A_{i_1}) \cdot \ldots \cdot P(A_{i_r})$$

Перейдем к самой модели. Проведём m испытаний, каждое из которых заканчивается успехом 1 или неудачей 0. Тогда:

$$\Omega = \{(\omega_1, \ldots, \omega_m) \mid \omega_i \in \{0, 1\}\}$$

Пусть каждое испытание успешно с вероятностью p и испытания независимы в совокупности.

Тогда вероятность исхода, в котором $s = \omega_1 + \ldots + \omega_m$ единиц, равна:

$$P(\omega) = p^s (1 - p)^{m - s}$$

Эта модель называется **схемой Бернулли**. Думаю, с ней вы уже знакомы.

Биномиальное распределение

Пусть $S = \omega_1 + ... + \omega_m$ - число успехов. Найдём $P(S = k)$.

Каждый исход с k успехами имеет вероятность $p^k(1-p)^{m-k}$, как мы уже выяснили, а таких строк столько, сколькими способами можно выбрать места для k единиц среди m позиций, то есть C_m^k. Поэтому

$$P(S = k) = C_m^k\, p^k(1-p)^{m-k}, \quad k = 0, 1, ..., m$$

Это уже **биномиальное распределение**.

В частности, вероятность того, что успехов не будет вовсе, равна

$$P(S = 0) = (1-p)^m$$

Ну а среднее:

$$\mathbb{E}S = \sum_{i=1}^{m} \mathbb{E}\,\omega_i = \sum_{i=1}^{m} P(\omega_i = 1) = mp$$

Редкие события

Представим, что у нас испытаний очень много и каждое маловероятно. Положим:

$$p = \frac{\lambda}{m}$$

и устремим m к бесконечности, оставляя λ фиксированным.

Теорема Для любых фиксированных $\lambda > 0$ и $k \geq 0$

$$C_m^k \left(\frac{\lambda}{m}\right)^k \left(1 - \frac{\lambda}{m}\right)^{m-k} \xrightarrow[m\to\infty]{} e^{-\lambda}\frac{\lambda^k}{k!}$$

Доказательство. Разбиваем выражение на множители

$$\frac{\lambda^k}{k!} \cdot \underbrace{\frac{m(m-1)...(m-k+1)}{m^k}}_{\text{(I)}} \cdot \underbrace{\left(1 - \frac{\lambda}{m}\right)^m}_{\text{(II)}} \cdot \underbrace{\left(1 - \frac{\lambda}{m}\right)^{-k}}_{\text{(III)}}$$

1. I это произведение k дробей $\frac{m}{m}, \frac{m-1}{m}, ..., \frac{m-k+1}{m}$, каждая из которых стремится к 1 $\implies$ I $\to 1$

(здесь у нас k фиксировано)

2. III - это k одинаковых множителей, каждый из которых стремится к 1 $\implies$ III $\to 1$

3. В II множителей уже m и их число растёт. Но у нас есть лемма! При $m \geq 2\lambda$ $x = \frac{\lambda}{m}$ лежит в $[0, \frac{1}{2}]$:

$$-x - x^2 \leq \ln(1-x) \leq -x$$

Подставим $x = \frac{\lambda}{m}$ и умножим на m:

$$-\lambda - \frac{\lambda^2}{m} \leq m \ln\left(1 - \frac{\lambda}{m}\right) \leq -\lambda$$

Левая часть стремится к $-\lambda$, правая равна $-\lambda$, поэтому по теореме о двух милиционерах $\text{II} \to e^{-\lambda}$.

Перемножая пределы, получаем $e^{-\lambda}\frac{\lambda^k}{k!}$.
ч.т.д.

Определение. Говорят, что случайная величина Y имеет *распределение Пуассона с параметром* $\lambda > 0$, если она принимает значения $0, 1, 2, ...$ и

$$P(Y = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Среднее:

$$\mathbb{E}Y = \sum_{k=1}^{\infty} k \cdot e^{-\lambda}\frac{\lambda^k}{k!} = \lambda \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!} = \lambda$$

(сократили k с $k!$ и ряд превратился в сумму всех вероятностей)

Итак, число успехов в большом количестве независимых редких испытаний приближённо распределено по Пуассону.

В частности, при $k = 0$:

$$P(\text{ни одного успеха}) \approx e^{-\lambda} = e^{-\text{среднее число успехов}}$$

Вот и возникла та самая экспонента :)

Дни рождения

Теперь попробуем сопоставить две задачи:

Схема Бернулли	Дни рождения
Испытание	Пара людей
Число испытаний m	Число пар C_n^2
Успех	Пара совпала, $I_{ij} = 1$
Вероятность успеха p	$P(d_i = d_j) = \frac{1}{N}$
Число успехов S	Число совпавших пар X

Схема Бернулли	Дни рождения
Среднее mp	$\mathbb{E}X = \frac{C_n^2}{N} = \lambda$

Иии у нас пар много, каждая совпадает редко, а среднее число совпавших пар λ. Вот та самая теорема Пуассона.

Если бы события "пара совпала" были **независимы В СОВОКУПНОСТИ**, то X имело бы биномиальное распределение и

$$P(X = 0) = \left(1 - \frac{1}{N}\right)^{C_n^2} \approx e^{-\lambda}$$

А так ли это?

Проверим независимость.

Тот факт, что любые 2 пары независимы, проверяется несложно. Оставлю это как упражнение читателю :)

А вот три пары, уже нет:

$$P(I_{12} = 1,\ I_{13} = 1,\ I_{23} = 0) = 0 \neq \frac{1}{N} \cdot \frac{1}{N} \cdot \left(1 - \frac{1}{N}\right)$$

Потому у нас события попарно независимы, но не независимы в совокупности.

Но в любом случае, зависимость здесь слабая, потому можно сказать, что X распределена почти по Пуассону. *(кстати, это можно доказать при помощи метода Чена-Стейна)*

Хэши

Начнем с определения.

Хеш-функцией называется функция

$$h : U \to \{0, 1, \dots, 2^b - 1\}$$

т.е. любому объекту из некого множества сопоставляет b-битное число.

Объектов в U может быть больше, чем возможных значений, потому разные объекты могут получать одинаковый хеш, что называется **коллизией**

Может появиться вопрос. Причем тут вообще вероятности и дни рождения. Ибо вообще хеш-функция детерминирована.

Но, как правило, хеш-функцию проектируют так, чтобы её значения на разных объектах выглядели как независимые и равномерно распределённые.

Опять же сравним задачи :)

Дни рождения	Хеши
Человек	Объект
День рождения d_i	Хеш $h(x_i)$
Число дней N	Число значений 2^b
Совпадение дней рождения	Коллизия

Буквально то же самое, только с $N = 2^b$!

Тогда среднее число пар объектов с одинаковым хешем:

$$\lambda = \frac{n(n-1)}{2 \cdot 2^b}$$

Вероятность коллизии

Вспомним нашу двустороннюю оценку:

$$1 - e^{-\lambda} \leq P(\text{коллизия}) \leq \lambda$$

Нас так-то интересует, чтобы вероятность коллизии была как можно меньше, то есть λ мало. А при малом λ выполнено $1 - e^{-\lambda} \approx \lambda$, и обе границы практически совпадают:

$$P(\text{коллизия}) \approx \lambda = \frac{n(n-1)}{2 \cdot 2^b}$$

(перейдем к упрощениям, да)

Для вас я посчитал вероятности по выведенной нами формуле:

Объектов n	Хеш	λ	P(коллизия)
10^3	32 бита	$1.2 \cdot 10^{-4}$	$\approx 1.2 \cdot 10^{-4}$
10^5	32 бита	1.16	от 0.69 до 1
10^6	32 бита	116	практически 1
10^6	64 бита	$2.7 \cdot 10^{-8}$	$\approx 2.7 \cdot 10^{-8}$
10^9	64 бита	0.027	≈ 0.027

Выходит, при 32 битах, вероятность коллизии становится очень значительной даже при всего ста тысячах объектов.

Сколько бит нужно

Решим обратную задачу.

Пусть у нас n объектов и мы хотим, чтобы вероятность коллизии не превышала заданного ε.

Из верхней оценки:

$$P(\text{коллизия}) \leq \lambda \leq \frac{n^2}{2 \cdot 2^b}$$

Требуем, чтобы $\frac{n^2}{2 \cdot 2^b} \le \varepsilon$, то есть

$$b \ge 2 \log_2 n + \log_2 \frac{1}{\varepsilon} - 1$$

Например, для миллиарда объектов и $\varepsilon = 10^{-6}$ нам нужно как минимум:

$$b \ge 60 + 20 - 1 = 79 \text{ бит}$$

В общем, правило достаточно простое. Если у вас есть уже N ячеек в той же хеш-таблице, уже примерно при $\sqrt{N}$ ключах у вас возникнет коллизия.

Однако, все это верно только если хеш-функция равномерна. А что если какие-то значения выпадают чаще? Да и те же дни рождения распределены по году неравномерно. Как вы думаете, в таких случаях совпадений станет больше или меньше?

Неравномерность

До сих пор мы исходили из предположения, что все N значений равновероятны. На практике это далеко не всегда так.

Давайте разбираться!

Пусть значение $v \in \{1, \dots, N\}$ выпадает с вероятностью p_v, где $p_v \ge 0$ и $\sum_v p_v = 1$. Объекты по-прежнему независимы, поэтому вероятность исхода равна

$$P(d_1, \dots, d_n) = p_{d_1} \cdot p_{d_2} \cdot \dots \cdot p_{d_n}$$

Возьмём пару $i \ne j$. Событие $d_i = d_j$ тогда:

$$P(d_i = d_j) = \sum_{v=1}^{N} P(d_i = v)\, P(d_j = v) = \sum_{v=1}^{N} p_v^2$$

Обозначим $q = \sum_v p_v^2$. В равномерном случае $q = N \cdot \frac{1}{N^2} = \frac{1}{N}$, как и раньше.

Распишем сумму квадратов отклонений от равномерного распределения:

$$\sum_{v=1}^{N} \left(p_v - \frac{1}{N} \right)^2 = \sum_v p_v^2 - \frac{2}{N} \sum_v p_v + N \cdot \frac{1}{N^2} = q - \frac{2}{N} + \frac{1}{N} = q - \frac{1}{N}$$

Получаем:

$$q = \frac{1}{N} + \sum_{v=1}^{N} \left(p_v - \frac{1}{N} \right)^2$$

Выходит, что вероятность совпадения пары равна равномерной вероятности **плюс суммарный квадрат отклонения от равномерности**.

Значит, $q \ge \frac{1}{N}$!

Равенство достигается только при равномерном распределении

(То же неравенство btw можно получить из неравенства Коши-Буняковского: $1 = \left(\sum p_v \cdot 1\right)^2 \le \sum p_v^2 \cdot \sum 1^2 = Nq$*)*

Среднее

$$\mathbb{E}X = \sum_{i<j} P(d_i = d_j) = C_n^2 \cdot q \geq \frac{C_n^2}{N}$$

Среднее число совпавших пар при неравномерности только растёт.

Оценка $P(B) \leq \mathbb{E}X$ тоже остаётся верной *(при ее выводе пользовались только предположением о неотрицательности X)*

Почему равномерность

По пути можно доказать оч прикольную и классную теоремку, которая объясняет, почему пытаются придумать именно равномерную хеш функцию:

Теорема. При $n \leq N$ вероятность того, что все n значений различны, максимальна при равномерном распределении и только при нём.

Найдем эту вероятность.

ля тут чета сложно доказывать. сошлюсь на упражнение читателю. *(если коротко: вероятность, что все значения различны, равна $n! \cdot e_n(p_1, \dots, p_N)$, где e_n это элементарный симметрический многочлен, а он Шур-вогнутый, поэтому максимум достигается в точке $p_v = \frac{1}{N}$)*

Заключение

Если вкратце обо всем и ни о чем:

- парадокс объясняется тем, что пар очень много: людей n, а пар порядка n^2, потому совпадения начинаются уже при $n \sim \sqrt{N}$
- экспонента появляется, так как совпадения пар - редкие и почти независимые события, а число таких событий распределено по Пуассону, вероятность нуля у которого равна $e^{-\lambda}$
- при неравномерности вероятность совпадения увеличивается