2020

Решение тестового задания

НА ПОЗИЦИЮ СТАЖЁРА АНАЛИТИКА ДАННЫХ

АЛЕКСАНДР КАСЬЯН

Оглавление

Задание №1

Условие:

Рекламные кампании на площадках Rambler Group используют самые запоминающиеся и цепляющие баннеры. Аналитики имеют доступ к базам, в которых есть информация про каждый показ креатива и клик по нему.

Таблица **Shows_table** содержит:

- *show_id* - идентификатор показа
- *day* - день показа

show_id	*day*
12367	2018-10-04
28736	2019-02-22
19862	2019-01-31

Таблица Clicks_table содержит:

- *click_id* - идентификатор показа, по которому совершили клик
- *bounce* - отказ пользователя от лендинга рекламодателя после клика (0 - пользователь после перехода на сайт изучает информацию о продукте, 1 - пользователь при переходе на сайт рекламодателя сразу покинул страницу)

click_id	*bounce*
12367	1
15627	0
28735	0

Необходимо выбрать всех пользователей, которые кликнули по баннеру в феврале 2020 и не отказались от сайта рекламодателя.

Результат: sql-запрос в любой СУБД

Ответ:

```
SELECT DISTINCT show_id FROM Shows_table
INNER JOIN Clicks_table ON
show_id=click_id AND bounce='0' AND day BETWEEN '2020-02-01' AND '2020-02-29'
```

Пояснения:

Для решения поставленной задачи я использовал SQL Fiddle и PostgreSQL 9.6.

Я создал и заполнил две таблицы с помощью DDL:

```
CREATE TABLE Shows_table (
  show_id INT,
  day DATE
);

CREATE TABLE Clicks_table (
  click_id INT,
  bounce BOOLEAN
 );

insert into Shows_table (show_id, day) values (12367, '2018-10-04');
insert into Shows_table (show_id, day) values (28736, '2019-02-22');
insert into Shows_table (show_id, day) values (19862, '2019-01-31');
insert into Shows_table (show_id, day) values (11111, '2020-02-04');
insert into Shows_table (show_id, day) values (22222, '2020-02-01');
insert into Shows_table (show_id, day) values (33333, '2020-02-29');
insert into Clicks_table (click_id, bounce) values (12367, '1');
insert into Clicks_table (click_id, bounce) values (28736, '0');
insert into Clicks_table (click_id, bounce) values (19862, '0');
insert into Clicks_table (click_id, bounce) values (11111, '1');
insert into Clicks_table (click_id, bounce) values (22222, '0');
insert into Clicks_table (click_id, bounce) values (33333, '0');
```

Тем самым создал две таблицы:

show_id	*day*
12367	2018-10-04
28736	2019-02-22
19862	2019-01-31
11111	2020-02-04
22222	2020-02-01
33333	2020-02-29

click_id	*bounce*
12367	1
28736	0
19862	0
11111	1
22222	0
33333	0

Запрос, который я составил:

```
SELECT DISTINCT show_id FROM Shows_table
INNER JOIN Clicks_table ON
show_id=click_id AND bounce='0' AND day BETWEEN '2020-02-01' AND '2020-02-29'
```

Запрос выводит неодинаковые значения *show_id*/*click_id* (т.к. я не особо понял, что значит: “Необходимо выбрать всех пользователей, которые кликнули по баннеру…”, я предположил, что требуемыми значениями являются значения столбца *click_id*), у которых значение *bounce* равно 0 (т.е. пользователь заинтересовался продуктом и остался на сайте) и значение *day* находится в пределах от 2020-02-01 до 2020-02-29 (т.е. февраль 2020 года), включая границы. Результат для вышеприведённого примера: 22222 и 33333.

Задача №2

Условие:

В дружном коллективе Рамблера любят играть в настольный футбол: по четным дням мы играем в футбол после обеда, а по нечетным - до. Все разбиваются на N команд и каждая команда играет с каждой. Так как деление на команды случайно, то результаты игр (победа или поражение, **без ничьих**) абсолютно случайны.

1. Какова вероятность, что какая-то из команд закончит турнир без поражений.

2. Сколько раз нужно провести чемпионат, чтобы с вероятностью 98% хотя бы раз такое случилось?

Ответ:

1. $P(\text{какая} - \text{то команда закончит турнир без поражений}) = \frac{N}{2^{N-1}}$
2. Для того, чтобы хотя бы один раз какая-нибудь команда закончила турнир без поражений с вероятностью 98% потребуется провести $\dfrac{\log 0.02}{\log(1-N\times(2)^{1-N})}$ турниров.

Решение:

Дано:
N – общее количество команд $P(\text{победы}) = \dfrac{1}{2}$ $P(\text{поражения}) = \dfrac{1}{2}$
Найти:
1) P(какая-то из команд закончит турнир без поражений) = ? 2) Сколько раз нужно провести чемпионат, чтобы с вероятностью 98% хотя бы раз такое случилось?

Решение

1. (Первая часть)

Способ I (комбинаторика):

a) Рассчитаем, сколько всего игр было сыграно между N команд (количество сочетаний N по 2, т.к. в одной игре участвуют 2 команды, и каждая команда играет с каждой):
$$\binom{N}{2} = \frac{N!}{2! \times (N-2)!} = \frac{\cancel{1\times2\times3\times\ldots\times(N-2)}\times(N-1)\times N}{2\times\cancel{(N-2)!}} = \frac{(N-1)\times N}{2}$$

b) Т.к. каждая команда может либо победить, либо проиграть, то у каждой команды есть 2 исхода, следовательно $2^{\frac{(N-1)\times N}{2}}$ – количество всевозможных исходов за весь турнир.

c) Предположим, что команда А обыграла все команды в турнире, т.е. из сыгранных $(N-1)$ игр все игры – победные. Следовательно, количество игр с неопределённым исходом уменьшится на $(N-1)$ игр. В связи с этим уменьшится количество неопределённых исходов. Теперь количество неопределённых исходов равно:

$$2^{\frac{(N-1)\times N}{2}-(N-1)}$$

Но, т.к. нигде не сказано, что именно команда А обыграет все остальные (это может быть B, C, D, …), тогда:

$$N \times 2^{\frac{(N-1)\times N}{2}-(N-1)}$$

Мы получили всевозможное количество исходов, при условии, что какая-то из команд одержит верх над остальными.

d) Чтобы получить вероятность того, что какая-то команда закончит турнир без проигрышей мне нужно поделить количество исходов, где какая-то из команд закончила турнир без единого поражения (пункт ‘c’) на общее всевозможное количество исходов (пункт ‘b’):

$$P(\text{какая} - \text{то команда закончит турнир без поражений}) = \frac{N \times 2^{\frac{(N-1)\times N}{2}-(N-1)}}{2^{\frac{(N-1)\times N}{2}}} = N \times 2^{\frac{(N-1)\times N}{2} - (N-1) - \frac{(N-1)\times N}{2}} = \frac{N}{2^{N-1}}$$

Ответ: $\boldsymbol{\frac{N}{2^{N-1}}}$**;**

Способ II (теория вероятности):

a) Только одна команда может выиграть весь матч (обыграть каждую команду), пусть это будет команда A, тогда вероятность того, что команда А выиграет каждую игру $\left(\frac{1}{2}\right)^{N-1}$, т.к. они играют $(N-1)$ игр.

b) Но вероятность, что **какая-то** из команд выиграет турнир, будет равна $P(\text{команда А выиграет турнир}) + P(\text{команда } B \text{ выиграет турнир}) + P(\text{команда } C \text{ выиграет турнир}) + \ldots$

*** Это взаимоисключающие события, потому что только одна команда может выиграть все матчи.*

c) Тем самым:

$$P(\text{какая} - \text{то команда закончит турнир без поражений}) = N \times \left(\frac{1}{2}\right)^{N-1} = \frac{N}{2^{N-1}}$$

Ответ: $\boldsymbol{\frac{N}{2^{N-1}}}$**;**

2. (Вторая часть)

a) Возьмём количество турниров, которое требуется провести для того, чтобы какая-то команда с 98% - ой вероятностью одержала победу над остальными командами за $\boldsymbol{x}$.

b) Тогда:

$P(\text{это произойдёт хотя бы 1 раз}) = 1 - P(\text{это никогда не произойдёт}) =$
$1 - (1 - N \times 2^{1-N})^x$

т.е. когда мы из 1 отнимаем вероятность того, что это произойдёт хотя бы 1 раз (которую мы получили в первой части задачи) – мы получаем вероятность, которая показывает, что данное событие никогда не произойдёт.

Возведение в степень x показывает, что данное событие никогда не произойдет ни в один из дней (от текущего до x):

$$\underbrace{(1 - N \times 2^{1-N})}_{\text{Ни сегодня}} \times \underbrace{(1 - N \times 2^{1-N})}_{\text{Ни завтра}} \times \dots \times \underbrace{(1 - N \times 2^{1-N})}_{\text{Ни через } x \text{ дней}}$$

c) Нам требуется: $1 - (1 - N \times 2^{1-N})^x = 0.98 => (1 - N \times 2^{1-N})^x = 0.02$

d) Прологарифмируем:

$$\log(1 - N \times 2^{1-N})^x = \log 0.02$$

По правилу log-power вынесем степень из-под знака логарифма:

$$x \times \log(1 - N \times 2^{1-N}) = \log 0.02$$

Выразим x (количество турниров):

$$x = \frac{\log 0.02}{\log(1 - N \times (2)^{1-N})}$$

Ответ: $\boldsymbol{\frac{\log 0.02}{\log\left(1-N\times(2)^{1-N}\right)}}$**;**