Введение в АД

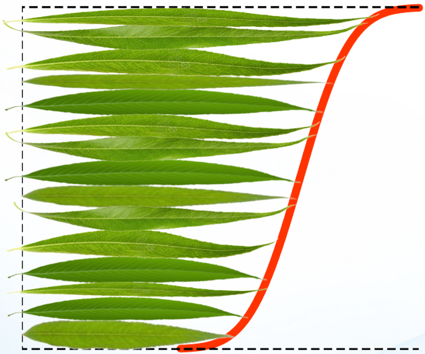
Екатерина Юшина

Лекция 3

Распределения и их свойства

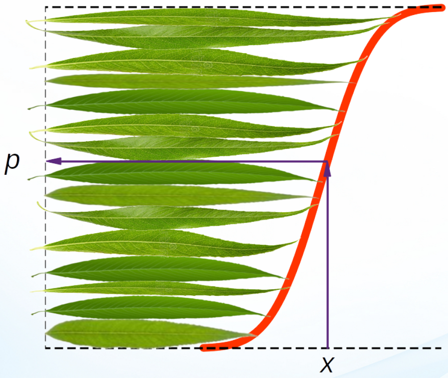
Соберем несколько листьев

Посмотрим на кончики



Приближенно получили функцию распределения нормального распределения.

Функция распределения



Функция распределения в точке x равна

доле листьев с длиной листа *не больше* x.

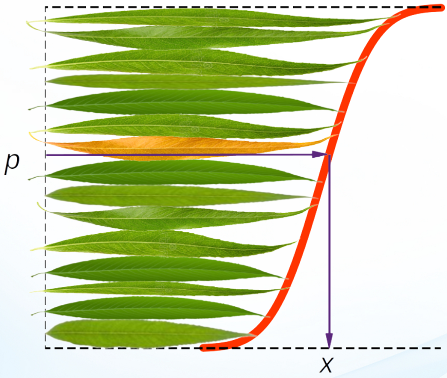
Что такое функция распределения

$F_\xi(x) = \mathrm{P}(\xi \leqslant x)$ — функция распределения случайной величины ξ.

Свойства из теории вероятностей:

1. Не убывает;
2. Непрерывная справа, может иметь разрывы;
3. $F(-\infty) = 0, F(+\infty) = 1$;
4. Однозначно характеризует распределение.

Возьмем значение p. Какой лист ему соответствует?

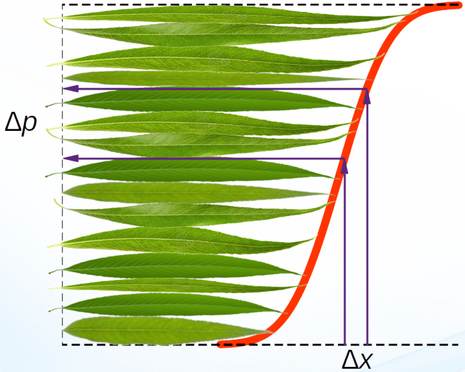


p-**квантиль** равна наименьшей длине листа, т.ч. есть не менее

$p \cdot 100\%$ листьев с длиной листа не больше данного листа.

Формально: $u_p = \min\{x \mid F(x) \geqslant p\}$

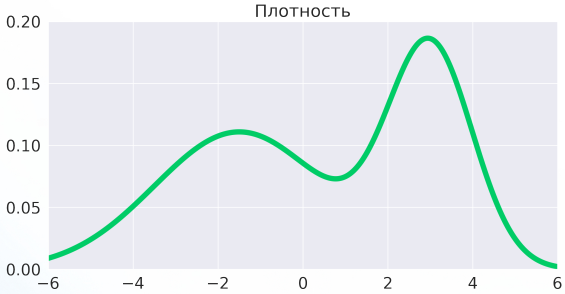
Плотность



Плотность в точке x равна $\Delta p/\Delta x$,

т.е. доле листьев с длиной листа в окрестности x.

Что такое плотность

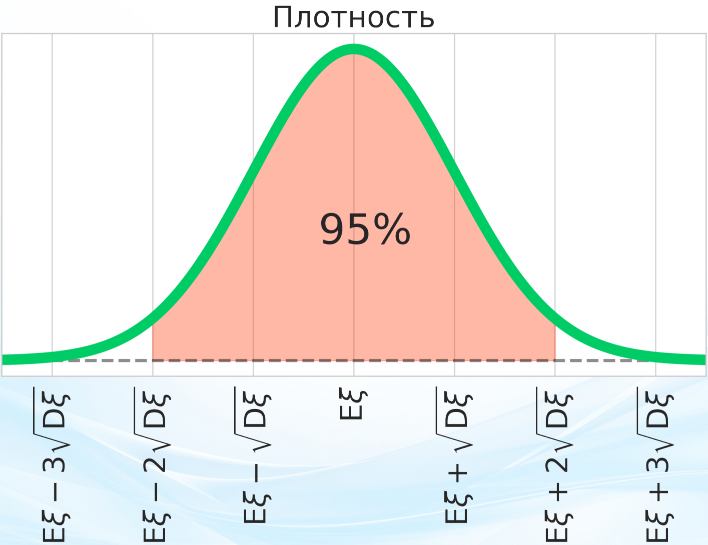


Свойства:

- лежит не ниже горизонтальной оси;
- площадь под кривой равна 1;
- неограничена сверху;
- вероятности события $\{a \le \xi \le b\}$ соответствует площадь под кривой между точками a и b;
- равна производной функции распределения.

Формальные определения и свойства см. теорию вероятностей.

Что такое мат. ожидание и дисперсия?



Формальные определения и свойства см. теорию вероятностей.

Нормальное распределение

Обозначение: $\mathcal{N}(a, \sigma^2)$

Параметры: $a \in \mathbb{R}, \sigma \in \mathbb{R}_+$

Носитель: $\mathbb{R}$, абс. непрерывное

Плотность: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-a)^2}{2\sigma^2}\right]$

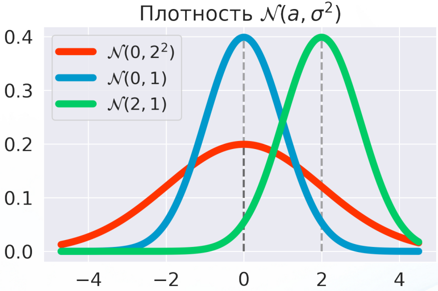
Математическое ожидание: a

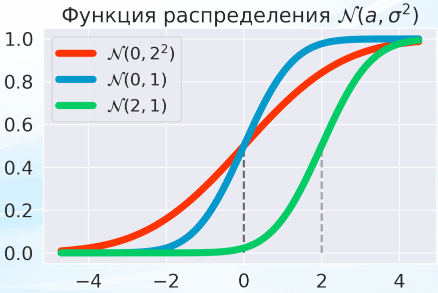
Дисперсия: σ^2

Интерпретация:

a — среднее значение

σ — разброс значений





Равномерное распределение

Обозначение: $U(a, b)$

Параметры: $a, b \in \mathbb{R}, a < b$

Носитель: $[a, b]$, абс. непрерывное

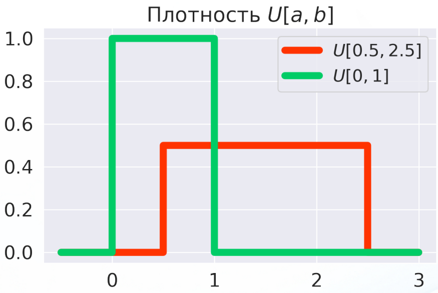
Плотность: $p(x) = \frac{1}{b-a} I\{x \in [a, b]\}$

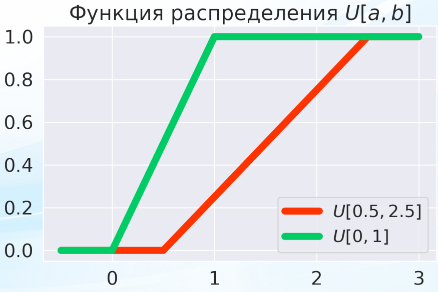
Математическое ожидание: $\frac{a+b}{2}$

Дисперсия: $\frac{(b-a)^2}{12}$

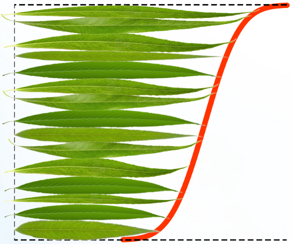
Интерпретация:

a и b — концы отрезка-носителя





Вернемся к листьям



Почему нельзя собрать все листья?

- Из слишком много.

Почему нельзя случайно равномерно выбирать листья?

- Длина листа может зависеть от региона.
- Нужно много человеческих ресурсов.
- Самые вкусные листья сожрал кот.

Почему на практике в точности не получится нормальное распред.?

- Мы собрали не все листья.
- Природа не идеальна.
- Не бывает листьев отриц. длины.

При работе с реальными данными имеют место подобные ситуации.

Бросок монетки и связанные распределения

Бернулли распределение

Обозначение: $Bern(p)$

Параметры: $p \in (0, 1)$

Носитель: $\{0, 1\}$, дискретное

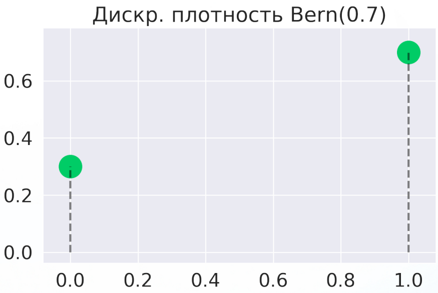
Вероятность: $P(\{1\}) = p$

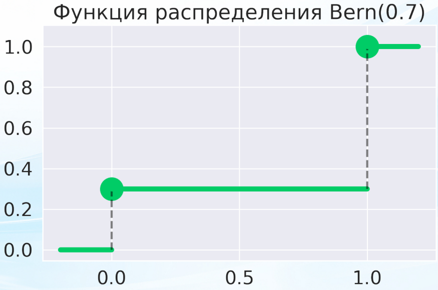
Математическое ожидание: p

Дисперсия: $p(1 - p)$

Интерпретация:

p — вероятность выпадения орла у монетки





Биномиальное распределение

Обозначение: $Bin(n, p)$

Параметры: $n \in \mathbb{N}, p \in (0, 1)$

Носитель: $\{0, 1, ... n\}$, дискретное

Вероятность: $P(\{k\}) = C_n^k p^k (1-p)^{n-k}$

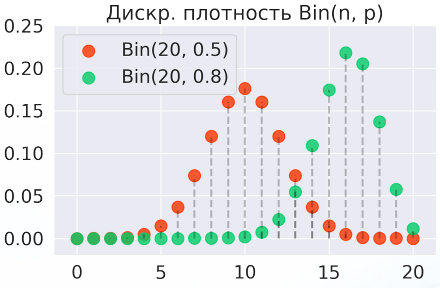
Математическое ожидание: np

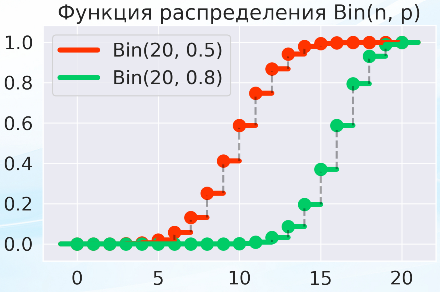
Дисперсия: $np(1-p)$

Интерпретация:

p — вероятность выпадения орла у монетки,

n — количество подбрасываний монетки





Бета-распределение

Обозначение: $\mathcal{B}(a, b)$

Параметры: $a, b > 0$

Носитель: $[0, 1]$, абс. непрерывное

Плотность: $p(x) = \frac{1}{\mathcal{B}(a,b)} x^{a-1}(1-x)^{b-1}$

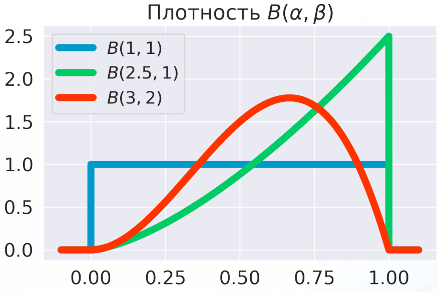
Математическое ожидание: $\frac{a}{a+b}$

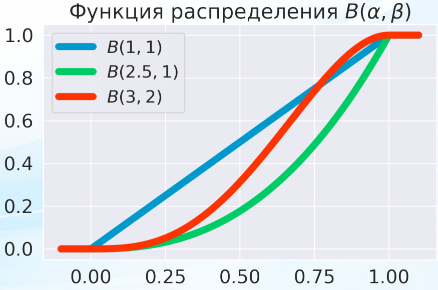
Дисперсия: $\frac{ab}{(a+b)^2(a+b+1)}$

Интерпретация:

a — "голоса" в пользу "орла"

b — "голоса" в пользу "решки"





Зачем изучать броски монет?

Бросок монеты — только простая интерпретация.

В реальности это могут быть

- корректность ответа ML-классификатора;
- клик пользователя по ссылке;
- ответы в соцопросах;
- наличие антител после вакцинации;
- и другие интерпретации.

Поток событий
и связанные распределения

Пуассоновское распределение

Обозначение: $Pois(\lambda)$

Параметры: $\lambda > 0$

Носитель: $\mathbb{Z}_+$, дискретное

Вероятность: $P(\{k\}) = \frac{\lambda^k}{k!} \exp(-\lambda)$

Математическое ожидание: λ

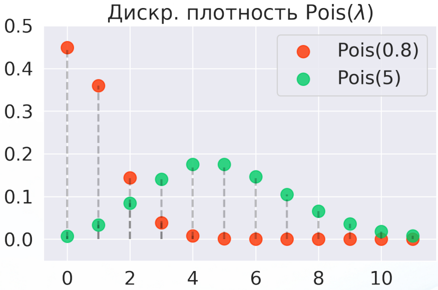
Дисперсия: λ

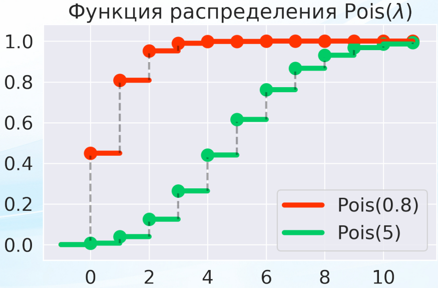
Интерпретация:

λ — интенсивность процесса

$=$ среднее количество событий,

произошедших за фикс. время





Экспоненциальное распределение

Обозначение: $Exp(\lambda)$

Параметры: $\lambda > 0$

Носитель: $\mathbb{R}_+$, абс. непрерывное

Плотность: $p(x) = \lambda e^{-\lambda x} I\{x > 0\}$

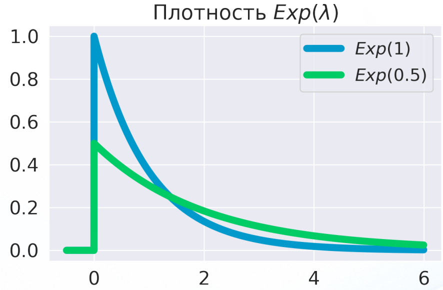
Математическое ожидание: $1/\lambda$

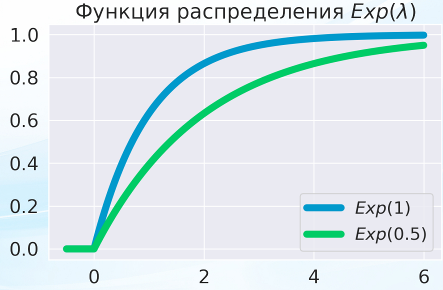
Дисперсия: $1/\lambda^2$

Интерпретация:

λ — интенсивность процесса

$=$ среднее количество событий, произошедших за фикс. время





Связь пуассоновского и экспоненц. распределений

Пусть события происходят независимо друг от друга
с одинаковой частотой λ шт. в минуту.

Примеры:

- момент прихода клиента в магазин;
- момент получения заказа от пользователя;
- момент прихода запроса на сервер;
- момент приезда автобуса на остановку.

Тогда часто предполагается, что

1. $Pois(\lambda t)$ — распр. кол-ва событий за время t.
   их среднее число λt.

2. $Exp(\lambda)$ — время между двумя соседними событиями
   среднее время $1/\lambda$ на событие.

Связь пуассоновского и экспоненц. распределений

Почему так?

Экспон. распределение обладает свойством **отсутствия памяти**:

Если $\xi \sim Exp(\lambda)$, то для любых $s, t > 0$:

$$\mathsf{P}(\xi > t + s \mid \xi > s) = \mathsf{P}(\xi > t)$$

Смысл:
Если следующего клиента ожидаем уже s минут после прихода предыдущего, то в среднем ждать осталось столько же, как будто бы предыдущий только что пришел.

Для абс. непрер. распр. экспоненциальное — единственное со свойством отсутствия памяти. Для дискретных — геометрическое.

Связь с пуассоновским будет доказана в случайных процессах.

Гамма-распределение

Обозначение: $\Gamma(\alpha, \beta)$

Параметры: $\alpha, \beta > 0$

Носитель: $\mathbb{R}_+$, абс. непрерывное

Плотность: $p(x) = \frac{\alpha^\beta}{\Gamma(\beta)} \cdot x^{\beta-1} e^{-\alpha x}$

Математическое ожидание: β/α

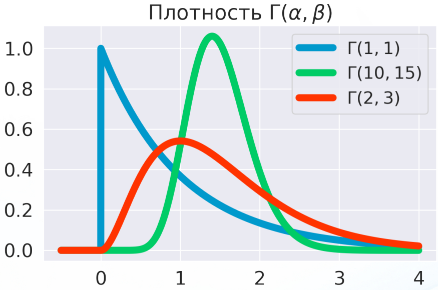
Дисперсия: β/α^2

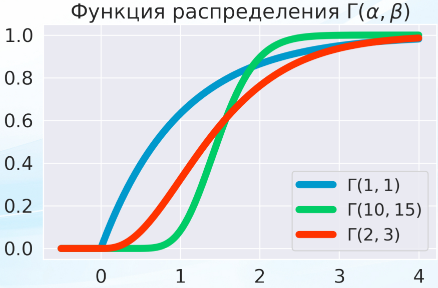
Возможная интерпретация:

α — интенсивность процесса

= среднее количество событий, произошедших за фикс. время

β — количество событий





Генерация случайных величин

Генерация случайных величин

Задача: сгенерировать $\psi \sim Bern(p)$, имея $\xi \sim U(0, 1)$

Решение: $\psi = I\{\xi \leqslant p\}$

Задача: сгенерировать $\psi \sim Bin(n, p)$, имея $\xi \sim U(0, 1)$

Решение: $\psi = \sum_{i=1}^{n} \xi_i,$

где $\xi_i \sim Bern(p)$ — независимые случайные величины.

Задача: сгенерировать $\xi \sim U(0, 1)$, имея $\psi \sim Bern(1/2)$

Решение: запишем ξ в двоичной системе счисления: $\xi = 0, \xi_1\xi_2...\xi_n$,

где $\xi_i \sim Bern(1/2)$ — независимые случайные величины.

Генерация случайных величин

Задача: сгенерировать $\xi \sim \mathcal{N}(a, \sigma^2)$, имея $\psi \sim U(0, 1)$

Решение: используем преобразования Бокса-Мюллера.

Пусть $\psi_1, \psi_2 \sim U(0, 1)$ — независимые случайные величины.

Тогда $\xi_1 = \cos(2\pi\psi_1)\sqrt{-2\ln\psi_2}$, $\xi_2 = \sin(2\pi\psi_1)\sqrt{-2\ln\psi_2}$

являются независимыми случайными величинами из $\mathcal{N}(0, 1)$

Почему так? Задачка из теории вероятностей.

Задача: сгенерировать $\xi \sim \mathcal{N}(a, \sigma^2)$, имея $\psi \sim \mathcal{N}(0, 1)$

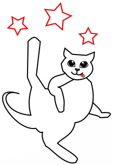
Решение: $\xi = a + \sigma\psi \sim \mathcal{N}(a, \sigma^2)$

Оценки в анализе данных

Примечание. Сегодня не погружаемся в формальности,

рассматриваем практ. точку зрения. Формальности будут, но позже.

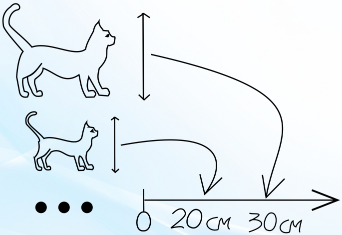
Данные



Пусть ξ — рост котика. Ничего про него не известно.

Что нужно сделать, чтобы оценить $\mathsf{E}\xi$?

Взять несколько котиков и измерить их рост!



$X_1, ..., X_n$ — **независимые одинаково распредел.** случ. величины, равные росту измеренных котиков. Называем их **выборкой**.

Независимость

Пусть $(\Omega, \mathscr{F}, \mathsf{P})$ — вероятностное пространство.

События $\{A_i\}_{i \in I} \subset \mathscr{F}$ **попарно независимы**, если

$$\forall i \neq j: \quad \mathsf{P}(A_i \cap A_j) = \mathsf{P}(A_i)\,\mathsf{P}(A_j).$$

События $\{A_i\}_{i \in I} \subset \mathscr{F}$ **независимы в совокупности**, если

$$\forall N \in \mathbb{N} \;\; \forall \{A_k\}_{k=1}^{N}: \quad \mathsf{P}(A_{i_1} \cap A_{i_2} \ldots \cap A_{i_N}) = \mathsf{P}(A_{i_1})\,\mathsf{P}(A_{i_2}) \ldots \mathsf{P}(A_{i_N}).$$

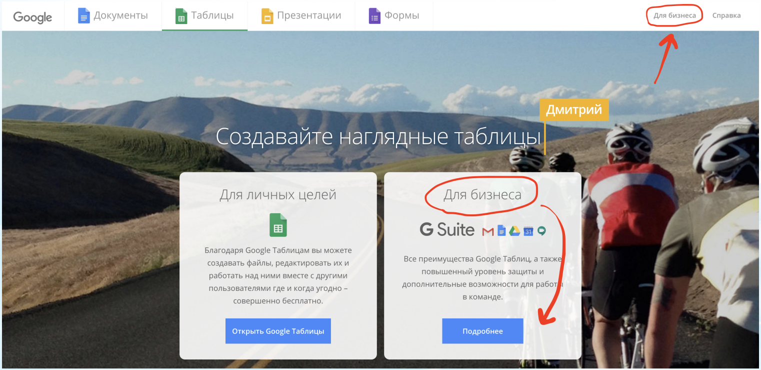
Случайные величины $\xi_1, \ldots, \xi_n$ **независимы в совокупности**, если $\forall \{A_k\}_{k=1}^{n} \subset \mathscr{F}$ события $\{\xi_1 \in A_1\}, \ldots, \{\xi_n \in A_n\}$ независимы в сов..

Независимость — важное понятие для анализа данных.

По умолчанию подразумевается независимость в совокупности.

Независимость: на практике

Вопрос: какой кнопке пользователь отдаст предпочтение?



Независимость: на практике

Собираем логи, смотрим:

timestamp	user_id	action	from
2020-02-02 23:03:04.930 UTC	123	click business button	header
2020-02-03 22:03:04.782 UTC	123	click business button	block
2020-02-03 23:03:19.837 UTC	123	click business button	header
2020-02-03 23:12:19.837 UTC	456	click business button	block
2020-02-03 23:13:01.394 UTC	456	click business button	block
2020-02-03 23:15:30.183 UTC	456	click business button	block
2020-02-03 23:18:25.938 UTC	789	click business button	header
2020-02-03 23:26:30.836 UTC	789	click business button	block
...	...	...	...

Расшифровка:

1. **timestamp** — отметка времени;
2. **user_id** — уникальный идентификатор пользователя;
3. **action** — выполненное действие (по нему отфильтровали);
4. **from** — какая из кнопок была нажата.

Независимость: на практике

Решение:

Пусть 1 — нажатие на кнопку в хэдере,

0 — нажатие на кнопку в блоке.

Метрика — вероятность выбора кнопки в header, а не в block.

Как посчитать метрику? Попробуем так:

$metric = \frac{H}{H+B}$,

где H — количество кликов на кнопку в header,

B — количество кликов на кнопку в block.

Недостатки:

- клики не являются независимыми событиями;
- один пользователь может кликнуть 1000 раз на одну кнопку;
- пользователь имеет привычки.

Независимость: на практике

Меняем решение:

Пусть 1 — нажатие на кнопку в хэдере,

0 — нажатие на кнопку в блоке.

Метрика — вероятность выбора кнопки в header, а не в block.

Посчитаем пред. метрику для каждого пользователя

и усредним по пользователям:

$metric = \frac{1}{n} \sum_{i=1}^{n} \frac{H_i}{H_i + B_i}$,

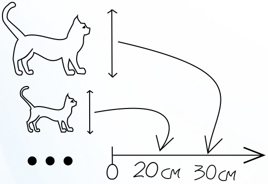
где n — количество уникальных пользователей,

H_i — кол-во кликов на кнопку в header для i-го пользователя,

B_i — кол-во кликов на кнопку в block для i-го пользователя.

Клики зависимы, а пользователи независимы.

Сходимость в теории



$X_1, ..., X_n$ — **независимые одинаково распределенные** случайные величины, равные росту измеренных котиков.

Теорема. Усиленный закон больших чисел (ЗБЧ, УЗБЧ).
Пусть $\{\xi_n\}_{n=1}^{+\infty}$ — независимые одинаково распределенные случайные величины, причем $\mathsf{E}\xi_1$ конечно. Тогда $\frac{1}{n}\sum_{i=1}^{n}\xi_i \longrightarrow \mathsf{E}\xi_1$.

Формально $\frac{1}{n}\sum_{i=1}^{n}\xi_i \xrightarrow{\text{п.н.}} \mathsf{E}\xi_1$, то есть $\mathsf{P}(\{\omega\in\Omega \mid \frac{1}{n}\sum_{i=1}^{n}\xi_i(\omega)\longrightarrow \mathsf{E}\xi_1\}) = 1$.
Подробно и с доказательством в теории вероятностей.


ВСЁ!