CS Space

# Вероятностное моделирование

## лекция 2

**Артур Игнатьев**
CS Space, МКН СПбГУ, ИТМО

**Владимир Евменов**
CS Space, Huawei

# Оценка разброса

- А если мы хотим оценивать характеристику не только одним числом?
- Среднее число клиентов в день находится в каком-то интервале с большой вероятностью
- Бизнес вопрос. Сколько проект заработает за месяц?
  - Можно дать оценку суммарного дохода за месяц
  - А можно сказать, что с большой вероятностью доход будет от $a$ до $b$

# Монте-Карло

- Уже умеем оценивать характеристики
- Если можем сэмплировать из $X$

1. Генерируем $N$ выборок $x_{[n],1}, \ldots, x_{[n],N}$ размера $n$.
2. Для каждой из них считаем нужную нам характеристику $\phi_i^\star = \phi^\star(x_{[n],i})$.
3. У полученной выборки $\phi^*_{[N]}$ считаем дисперсию, доверительный интервал.

- Выборки $x_n^\star$ и $\phi_N^b$ называются бутстраповскими.

# Бутстрап

- Всегда ли мы можем сэмплировать из $X$?
- Если можем проводить эксперимент много раз, то круто
- А если с.в. это данные, например, из бизнеса
- Бутстрап (bootstrap) — это метод Монте-Карло, примененный к какой-либо аппроксимации $\mathscr{P}_X$ (например, к $\mathscr{P}^\star_{[n]}$).
- Усредненный рецепт такой:

1. Генерируем $N$ выборок $x^\star_{[n],1}, \ldots, x^\star_{[n],N}$ размера $n$ из $X^\star$. Для генерации одной выборки $x^\star_{[n],i}$ нужно взять $n$ случайных элементов исходной выборки $x_{[n]}$.
2. Для каждой из них считаем нужную нам характеристику $\phi^b_i = \phi^\star(x^\star_{[n],i})$.
3. У полученной выборки $\phi^b_{[N]}$ считаем дисперсию, доверительный интервал.

- Выборки $x^\star_n$ и $\phi^b_N$ называются бутстраповскими.

# Ошибка бутстрапа

Ошибка бутстрапа складывается из двух слагаемых:

- устранимая ошибка — из-за того, что мы взяли $N$ выборок, а не $\infty$,
- неустранимая ошибка — из-за того, что мы взяли $\mathscr{P}^{\star}_{[n]}$ вместо $\mathscr{P}_X$.

# Интервальные оценки

Вместо того, чтобы приблизить значение $\phi$ точечно, мы хотим оценить область, в которой лежит истинное значение.

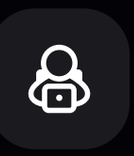
# Доверительный интервал

- Пара статистик $(\phi_L^\star, \phi_R^\star)$ называется доверительным интервалом для $\phi(\mathscr{P}_X)$ с уровнем доверия $\gamma$ если

$$P(\phi_L^\star(X_{[n]}) \leq \phi(\mathscr{P}_X) \leq \phi_R^\star(X_{[n]})) = \gamma$$

- Если сгенерировать $N$ выборок из распределения $\mathscr{P}_X$ и для каждой из них построить доверительный интервал, то примерно $\gamma N$ из них накроют истинное значение $\phi(\mathscr{P}_X)$.

- Конкретный доверительный интервал $(\phi_L^\star(x_{[n]}), \phi_R^\star(x_{[n]}))$ либо накрывает истинное значение, либо нет, никакой вероятности у этого события нет: нельзя говорить «Этот интервал с вероятностью 90% содержит истину».

# Эфронов доверительный интервал  × 

- Пусть $\phi_L^\star$ — $\alpha_1$-квантиль бутстраповской выборки $\phi_{[N]}^\star$, а $\phi_R^\star$ — $1-\alpha_2$-квантиль. Тогда $(\phi_L^\star, \phi_R^\star)$ это асимптотический доверительный интервал с уровнем доверия $1-\alpha_1-\alpha_2$.

# Пример

| Контрольная группа | A | Б |
|---|---|---|
| Размер группппы | 893 | 923 |
| Число купивших | 34 | 28 |
| Конверсия | 3.81% | 3.03% |

- Ого, выигрыш в 26%, время идти просить премию, или нет?
- Аналитический можно показать, что разница мат. ожиданий лежит в интервале (-0.93%, 2.48%) с вероятностью 0.95
- Давайте предположим, что у нас нету возможности дать ответ аналитический. Воспользуемся бутстрапом.
- 10 000 раз сэмплируем из обоих групп, считаем разницу средних, далее строим доверительный интервал по распределению разницы.
- Получаем (-0.89%, 2.4%)