Bootstrap

1.1 Бутстрапирвание совокупности бернуллиевских случайных величин (КР-2020, №4)

Условие задачи

Исходная выборка y — вектор из n независимых случайных величин, равновероятно принимающих значения 0 и 1 (коммент: можно и для любого распределения, у которого есть матожидание и дисперсия). Пусть y^* — одна из бутстрэп-выборок.

- Найдите $\mathbb{E}\left[y_i\right], Var\left(y_i\right), \mathbb{E}\left[\overline{y}\right], Var\left(\overline{y}\right)$
- $\mathbb{E}\left[y_i^*\right], Var\left(y_i^*\right), \mathbb{E}\left[\overline{y}^*\right], Var\left(\overline{y}^*\right)$
- $Cov\left(y_i, y_i^*\right), Cov\left(\overline{y}, \overline{y}^*\right)$

Решение

- Прежде всего заметим, что по условию $y_i \sim Bern(p)$ с $p = 1/2$. Отсюда сразу получаем, что $\mathbb{E}\left[y_i\right] = p = 1/2$ и $Var\left(y_i\right) = p(1-p) = 1/4$. Более того,
$$\mathbb{E}\left[\overline{y}\right] = \frac{1}{n}\mathbb{E}\left[y_1 + \ldots + y_n\right] = \frac{np}{n} = p = 1/2.$$
Если же принять во внимание независимость случайных величин $y_1, .., y_n$, получим
$$Var\left(\overline{y}\right) = Var\left(y_1 + \ldots + y_n\right)\frac{1}{n^2} = \frac{n}{n^2}Var\left(y_1\right) = \frac{1}{4n}.$$
- Ключевое соображение состоит в том, чтобы прибегнуть к представлению случайной величины y_i^* в виде линейной комбинации величин $y_1, .., y_n$ со случайными коэффициентами $d_1, .., d_n$, каждый из которых принимает значения либо 0, либо 1. Более того, $d_j = 1$ тогда и только тогда, когда $y_i^* = y_j$. Обозначим через d вектор из случайных величин $d_1, .., d_n$.

  Про это нужно думать следующим образом: генерация конкретной бутстрэп-выборки равносильна генерации квадратной матрицы D размера $n \times n$, каждый столбец которой содержит ровно одну единицу, а все остальные её элементы равны нулю. При этом позиция, содержащая единицу, определяется случайно и равновероятно для каждого столбца. Тогда
$$(y_1^*, .., y_n^*) = (y_1, .., y_n) \cdot D.$$
  Теперь несложно заметить, что величина y_i^* при условии коэффициентов d имеет распределение $Bern(p)$ независимо от значения d. Действительно, при фиксированном d мы сразу можем сказать, с какой из случайных величин $y_1, .., y_n$ совпала y_i^*. Но это означает, что условное распределение для y_i^* совпадает с безусловным. Таким образом, $y_i^* \sim Bern(p)$. Отсюда сразу получаем, что $\mathbb{E}\left[y_i^*\right] = 1/2$, а $Var\left(y_i^*\right) = 1/4$. Более того,
$$\mathbb{E}\left[\overline{y}^*\right] = \frac{1}{n}\mathbb{E}\left[y_1^* + .. + y_n^*\right] = 1/2.$$
  Теперь приступим к поиску $Var\left(\overline{y}^*\right)$, предварительно отметив, что $\mathbb{E}\left[d_j\right] = 1/n$, а случайные величины y_j и d_j являются независимыми для всех j (это простое наблюдение, очевидно следующее из самой процедуры бутстрэпа).

Итак,

$$Var\left(\overline{y}^*\right)=\frac{1}{n^2}Var\left(y_1^*+..+y_n^*\right)=\frac{1}{n}Var\left(y_i^*\right)+\frac{(n-1)}{n}Cov\left(y_1^*,y_2^*\right).$$

Знаем, что $Var\left(y_i^*\right)=1/4$.

- Таким образом, остаётся вычислить $Cov\left(y_1^*,y_2^*\right)$, для чего мы и воспользуемся линейным разложением:

  $$Cov\left(y_1^*,y_2^*\right)=Cov\left(d_1y_1+..d_ny_n,d_1'y_1+..d_n'y_n\right)=\sum_{i=1}^{n}Cov\left(d_iy_i,d_i'y_i\right)+\sum_{i\neq j}Cov\left(d_iy_i,d_j'y_j\right)=\sum_{i=1}^{n}Cov\left(d_iy_i,d_i'y_i\right).$$

  Вычислим одно из слагаемых получившейся суммы:

  $$Cov\left(d_iy_i,d_i'y_i\right)=\mathbb{E}\left[d_id_i'y_i^2\right]-\mathbb{E}\left[d_iy_i\right]\mathbb{E}\left[d_i'y_i\right]=\mathbb{E}\left[d_i\right]^2\mathbb{E}\left[y_i^2\right]-\mathbb{E}\left[d_i\right]^2\mathbb{E}\left[y_i\right]^2=\frac{1}{n^2}Var\left(y_i\right)=\frac{1}{4n^2}.$$

  В итоге имеем

  $$Cov\left(y_1^*,y_2^*\right)=\frac{1}{4n}.$$

  Подставим теперь ответ в выражение для $Var\left(\overline{y}^*\right)$:

  $$Var\left(\overline{y}^*\right)=\frac{1}{4n}+\frac{(n-1)}{4n^2}.$$

- Остаётся вычислить

  $$Cov\left(\overline{y},\overline{y}^*\right)=\frac{1}{n^2}\left(\sum_{i=1}^{n}Cov\left(y_i,y_i^*\right)+\sum_{i\neq j}Cov\left(y_i,y_j^*\right)\right).$$

  Рассмотрим вначале слагаемое первого типа.

  $$Cov\left(y_i,y_i^*\right)=\mathbb{E}\left[y_i(d_1y_1+..d_ny_n)\right]-\mathbb{E}\left[y_i\right]\mathbb{E}\left[y_i^*\right]=\frac{n-1+2}{4n}-\frac{1}{4}=\frac{1}{4n}.$$

  Второе слагаемое на самом деле равно первому, так как мы снова сравниваем y_i с величиной, которая совпадёт с ней с вероятностью $1/n$; таким образом,

  $$Cov\left(y_i,y_j^*\right)=Cov\left(y_i,y_i^*\right)=\frac{1}{4n}.$$

  Подставив полученные выражения в изначальное, получаем ответ:

  $$Cov\left(\overline{y},\overline{y}^*\right)=\frac{1}{n^2}\left(\frac{1}{4}+\frac{(n-1)}{4}\right)=\frac{1}{4n}.$$

1.2 Бутстрапирование мёда (КР-2021, №4)

Условие задачи

У Винни-Пуха 1 000 000 наблюдений в минуту — потоковые данные по мёду от пчёл. А общее количество наблюдений n необозримо велико.

- К какому распределению стремится количество копий i-го исходного наблюдения в бутстрэп-выборке с ростом n?

- (позже...)Вместо честного наивного бутстрэпа Винни-Пух использует аппроксимацию. В реальном времени каждое поступающее наблюдение y_i он заменяет на его k_i копий, где количество k_i выбирает случайно, независимо от y и предыдущих k_j , согласно распределению найденному в предыдущем пункте. Например, если $k_i=0$, то наблюдение y_i Винни-Пух не запоминает, а если $k_i=2$, то наблюдение y_i учитывается дважды. Во сколько раз отличается дисперсия обычного среднего выборочного $\overline{y_n}$ и среднего выборочного сделанных копий, которое посчитает Винни-Пух для n исходных наблюдений?

Решение