Введение

Опишем процедуру построения теории, на которой держится математическая статистика.

В первую очередь в курсе теории вероятностей вводится понятие вероятностного пространства. Это понятие обобщает под одним термином три сущности, позволяющие в дальнейшем исследовать различные теоретические вопросы.

Итак, вероятностным пространством называют тройку

$$(\Omega, \mathcal{F}, P),$$

где Ω — это пространство элементарных исходов (все возможные результаты эксперимента), $\mathcal{F}$ — семейство подмножеств Ω, которые называются событиями, P — функция на событиях (вероятность).

Задание вероятностного пространства означает задание модели физического эксперимента.

С точки зрения теории меры, вероятность P называют еще вероятностной мерой.

В экспериментах мы хотим уметь рассчитывать вероятности событий, которые можно описать числами. Например, температуру в комнате, количество аварий, количество голов в матче и т.д. Поэтому естественно вводится понятие случайной величины.

$$X : \Omega \to \mathbb{R}$$

Случайная величина X, заданная на пространстве $(\Omega, \mathcal{F}, \mathcal{P})$, задает автоматически свою вероятностную меру, но уже на подмножествах B значений X (по-умному можно так записать $B \subseteq \operatorname{Im} X \subseteq \mathbb{R}$):

$$P_X(B) := P(\omega : X(\omega) \in B) \equiv P(X \in B) \equiv P(X^{-1}(B))$$

То есть каждая случайная величина, как говорят, индуцирует свое вероятностное пространство

$$(\mathcal{X}, \mathcal{F}_X, P_X)$$

Задание P_X называется заданием распределения X.

В задаче математической статистики мы имеем результат некоторого эксперимента. В терминах вероятностного пространства у нас есть исход $\omega \in \Omega$, на котором посчитана некоторая числовая характеристика X. В нашей мат. модели X — это случайная величина (обычно, случайный вектор), про распределение которой мы точно не знаем, но имеем уверенность, что это распределение лежит в некотором семействе распределений $\mathcal{P}$.

Если это семейство запараметризовано, т.е. каждому распределению семейства соответствует некоторый числовой параметр $\theta \in \Theta \subseteq \mathbb{R}^k$, то такое семейство называется параметризованным, и тогда задача узнать распределение сводится к задаче узнать параметр распределения.

Таким образом, статистическая модель (статистическое пространство) выглядит следующим образом

$$(\mathcal{X}, \mathcal{F}, \mathcal{P})$$

где $\mathcal{X}$ — множество всех возможных значений X, $\mathcal{F}$ — семейство событий на $\mathcal{X}$, $\mathcal{P}$ — семейство распределений X.

Оценивание (Estimation)

Выборкой назвается набор из n случайных величин $X_1, \ldots, X_n$. Реализацией выборки называется набор чисел $X_1(\omega), \ldots, X_n(\omega)$.

Итак, задача статистики заключается в том, чтобы получить какую-либо информацию о распределении по данной выборке $X_1, \ldots, X_n$.

Узнать что-либо о распределении это то же самое, что узнать что-то о его параметре, некую величину $h(\theta)$, где θ — параметр распределения. То есть нам надо, чтобы была функция $\hat{\theta}(x_1, \ldots, x_n)$, такая что случайная величина $\hat{\theta}(X_1, \ldots, X_n)$ как-то характеризовала величину $h(\theta)$. Тогда $\hat{\theta}(X_1, \ldots, X_n)$ называют оценкой $h(\theta)$.

Несмещенность

Встает вопрос, как определить, насколько оценка хороша. Естественно потребовать, чтобы оценка в среднем была равна оцениваемому параметру, то есть, чтобы

$$E_\theta \hat{\theta}(X_1, \ldots, X_n) = h(\theta)$$

где E_θ — математическое ожидание по мере P_θ (иначе говоря, для распределения P_θ).

Если такое условие выполнено, то оценка $\hat{\theta}$ называется несмещенной для $h(\theta)$.

Квадратичное отклонение

Как для случайных величин важно понятие дисперсии (среднеквадратичного отклонения от среднего), так и для оценок важно понятие квадратичного смещения.

Квадратичным смещением называется

$$E_\theta(\hat{\theta}(X_1, \ldots, X_n) - h(\theta))^2$$

Для несмещенных оценок квадратичное смещение равно $D_\theta \hat{\theta}(X_1, \ldots, X_n)$.

Задачи

Задача 1. Пусть $F_\theta(x) = 1 - \theta^{x+1}$, $x \in \{0, 1, \ldots\}$, $\Theta = (0, 1)$. Построить несмещенную оценку для θ от одного наблюдения.

Решение. Нужно построить оценку $\hat{\theta}(X_1)$, чтобы $E_\theta \hat{\theta}(X_1) = \theta$.

Посчитаем математическое ожидание такой оценки в общем виде, а затем попытаемся понять, какой должна быть сама оценка.

$$E_\theta \hat{\theta}(X_1) = \sum_{k=0}^{\infty} \hat{\theta}(k) P_\theta(X_1 = k) = (*)$$

Найдем $P_\theta(X_1 = k)$ из нашей функции распределения.

Так как $F_\theta(x) = P_\theta(X \le x)$, то

$$P_\theta(X_1 = k) = F_\theta(k) - F_\theta(k-1) = 1 - \theta^{k+1} - 1 + \theta^k = \theta^k(1 - \theta)$$

Теперь вернемся к подсчету мат. ожидания:

$$(*) = \sum_{k=0}^{\infty} \hat{\theta}(k)\theta^k(1-\theta) = (1-\theta)\sum_{k=0}^{\infty}\hat{\theta}(k)\theta^k$$

(воспользовались суммой геом. прогрессии $\sum_{k=0}^{\infty}\theta = \frac{1}{1-\theta}$)

Итак, чтобы оценка $\sum_{k=0}^{\infty}\hat{\theta}(k)\theta^k$ была несмещенной, нужно, чтобы

$$(1-\theta)\sum_{k=0}^{\infty}\hat{\theta}(k)\theta^k = \theta \Leftrightarrow \sum_{k=0}^{\infty}\hat{\theta}(k)\theta^k = \frac{\theta}{1-\theta} = \theta+\theta^2+\theta^3+\ldots = \sum_{k=1}^{\infty}\theta^k$$

Таким образом, мы хотим, чтобы $\hat{\theta}(0)=0$, а $\hat{\theta}(k)=1$ для всех $k=1,2,\ldots$. Отсюда

$$\hat{\theta}(X_1) = I_{X_1>0}.$$

Задача 2. а) Найти несмещенную оценку для DX_1 вида cS^2, где $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$, X_i — независимые одинаково распределенные (далее н.о.р.).

б) Найти ее квадратичное смещение в случае $X_i \sim \mathcal{N}(0,\theta)$.

□ Пункт а).

$$\begin{aligned}
ES^2 &= \frac{1}{n}\sum_{i=1}^{n}E(X_i-\overline{X})^2 = \frac{1}{n}\sum(EX_i^2 - 2EX_i\overline{X} + E\overline{X}^2) = \\
&= \frac{1}{n}\sum EX_i^2 - \frac{2}{n}\sum EX_i\overline{X} + \frac{1}{n}\sum E\overline{X}^2 = \\
&= EX_1^2 - 2E\left(\frac{1}{n}\sum X_i\overline{X}\right) + \frac{1}{n}nE\overline{X}^2 = \\
&= EX_1^2 - 2E\overline{X}^2 + E\overline{X}^2 = EX_1^2 - E\overline{X}^2
\end{aligned}$$

Теперь посчитаем $E\overline{X}^2$:

$$\begin{aligned}
E\overline{X}^2 &= E\left(\frac{X_1+\ldots+X_n}{n}\right)^2 = \frac{1}{n^2}E(X_1^2+\ldots+X_n^2+2\sum_{i<j}X_iX_j) = \\
&= \frac{1}{n^2}\left(nEX_1^2 + 2C_n^2(EX_1)^2\right) = \frac{1}{n}EX_1^2 + \frac{n-1}{n}(EX_1)^2
\end{aligned}$$

Подставляя в предыдущую формулу, получаем, что

$$ES^2 = EX_1^2 - \frac{EX_1^2}{n} - \frac{n-1}{n}(EX_1)^2 = \frac{n-1}{n}(EX_1^2-(EX_1)^2) = \frac{n-1}{n}EX_1$$

Таким образом, оценка S^2 — смещенная, но оценка $S_0 = \frac{n}{n-1}S^2$ будет несмещенной.

Теперь пункт б). Теперь мы знаем, что $X_i \sim \mathcal{N}(0,\theta)$, то есть $EX_i = 0$, $DX_i = \theta$.

Так как наша оценка несмещенная, то квадратичным смещением будет просто дисперсия DS_0.

$$DS_0 = ES_0^4 - (ES_0^2)^2 = ES_0^4 - (EX_1)^2 = ES_0^4 = E\left(\frac{n}{n-1}S^2\right)^2 = \frac{n^2}{(n-1)^2}ES^4$$

Для поиска ES^4 воспользуемся тем, что $S^2 = \overline{X^2} - \overline{X}^2$. тогда

$$\begin{aligned}
ES^4 &= E\frac{1}{n^2}(\overline{X^2}^2 - \overline{X}^2)^2 = \frac{1}{n^2}E(\overline{X^2}^2 - 2\overline{X^2} + \overline{X}^4) = \\
&= \frac{1}{n^2}\left(E\left(\right)\right)
\end{aligned}$$

■