Домашнее задание №2
по курсу «Анализ панельных данных и данных о длительности состояний»

1) Создайте выборку из 1000 случайных чисел $t_1^*, \dots, t_{1000}^*$, распределённых по закону Вейбулла $S(t) = e^{-\lambda t^p}$ с параметрами $\lambda = 0.07, p = 0.7$. Это будут изучаемые длительности.
2) Теперь сгенерируйте 1000 значений $t_1^c, \dots, t_{1000}^c$, равномерно распределённых на отрезке [0; 80] — это будут моменты цензурирования.
3) Рассчитайте величины $t_i = \min\{t_i^*, t_i^c\}, i = 1, \dots, 1000$. Это будут наблюдаемые длительности. Если $t_i^c \le t_i^*$, то состояние i наблюдалось только в течение t_i^c единиц времени и наблюдение за ним оказалось цензурированным. Если же $t_i^c > t_i^*$, то наблюдаемая длительность совпадает с полной, наблюдение не цензурировано.
4) Создайте переменную $\delta_i = \begin{cases} 1, & t_i^c \ge t_i^*, \\ 0, & t_i^c \le t_i^*. \end{cases}$

То есть $\delta_i = 1$ для не цензурированных наблюдений, $\delta_i = 0$ для цензурированных.

5) По данным о наблюдаемых длительностях $t_1, \dots, t_{1000}$ и индикаторе цензурирования/завершения состояний $\delta_1, \dots, \delta_{1000}$ рассчитайте оценку Каплана–Майера для функции дожития и оценку Нельсона–Аалена для интегральной функции риска.
6) Изобразите на одном графике:
   - настоящую функцию дожития для распределения Вейбулла с параметрами $\lambda = 0.07, p = 0.7$;
   - оценку Каплана–Майера.

То же самое проделайте и для интегральной функции риска.
Если вы можете как-нибудь прокомментировать получившиеся графики, сделайте это, пожалуйста.
7) Сравните выборочную медиану с истинной медианой распределения Вейбулла с заданными параметрами.
8) Теперь рассмотрите случай, когда данные подвержены усечению справа вместо цензурирования, так что все наблюдения, в которых $t_i^c \le t_i^*$, исключаются из выборки. По полученной выборке оцените функцию дожития и интегральную функцию риска теми же способами, рассчитайте выборочную медиану. Сравните полученные оценки с истинным распределением.

Напишите отчёт на естественном языке и вышлите на адреса furmach@inbox.ru и evrumyantseva.2006@yandex.ru не позже 12 декабря.

Методические указания

Если вы выполняете задание в пакете Stata, то вам могут пригодиться команды:

generate – для генерации случайных величин (функция, которая возвращает случайное число, равномерно распределённое на отрезке [0;1), в Stata называется **uniform()**).

stset – предварительная команда для анализа данных типа длительности состояний, указывает на переменные, содержащие наблюдаемые длительности и индикатор прекращения/цензурирования, для последующих команд.

sts gen – для расчёта оценок Каплана-Майера и Нельсона-Аалена (графики этих оценок строятся по команде **sts graph**).

stsum – для нахождения медианы.