Объявление: в следующую пятницу Антона не будет, будет две практики. Начнем с контрольной работы по теорверу + проверочная по этому ДЗ на одну задачу + разбор ДЗ, после продолжим говорить о методе Монте-Карло и оценивании вообще.

1 Обязательное домашнее задание

В этот раз прогерских задач будет две, похожих. Нужно будет сравнить в среднеквадратичном различные оценки для параметров некоторых распределений с помощью моделирования.

Действовать лучше по следующей схеме: фиксируем параметры распределения (любые разумные) некоторую длину выборки N (100–1000) и число серий M. Моделируем M выборок по N случайных величин, для каждой выборки считаем все статистики[1], получаем выборку длины M для для каждой статистики. Для каждой статистики находим среднеквадратичное отклонение от известного значения параметра (т.е. $RMSE_m = \sqrt{\sum_{j=1}^{M}(t_j^{(m)} - \theta)^2/M}$, где $t_j^{(m)}$ — оценка по выборке j, полученная m-м методом), красиво выводим. M следует выбирать так, чтобы считалось разумное время, но при этом результаты с.к.о. были устойчивы и устойчиво сравнивались. Далее все это стоит обернуть в цикл по N, чтобы показать состоятельность оценок.

1.1 Прогерские задачи

Задание 1.1. Рассмотрим распределение $U[0, \theta]$. Пусть $X_1, \ldots, X_N$ — выборка из этого распределения. Проверить состоятельность следующих оценок параметра θ и сравнить в смысле среднеквадратичного отклонения с помощью моделирования:

1. $\widehat{\theta} = 2\overline{X}$
2. $\widehat{\theta} = \max_i X_i$
3. $\widehat{\theta} = \frac{N+1}{N} \max_i X_i$

Как Вы думаете, какой смысл имеет множитель в последнем случае? А можете это показать?..

Задание 1.2. Рассмотрим распределение Коши со сдвигом, т.е. распределение с плотностью

$$f(x) = \frac{1}{\pi\,(1 + (x-a)^2)}$$

Пусть мы имеем выборку $X_1, \ldots, X_N$ из этого распределения и хотим оценить параметр *сдвига* a.

Для хороших распределений параметр сдвига оценивается с помощью выборочного среднего, но в этом случае это не пройдет (ЗБЧ работает только в том случае, если существует матожидание, а здесь его нет) Убедитесь в этом, для начала.

А после проверьте являются ли следующие оценки состоятельными и сравните их в смысле среднеквадратичного отклонения от истинного значения сдвига:

1. Выборочная медиана

[1] На самом деле, можно считать разные оценки по разным выборкам, если это окажется проще закодить, в данном (!) случае это непринципиально

2. Среднее, посчитанное после отбрасывания наибольшего и наименьшего значения выборки

3. Среднее, посчитанное после отбрасывания 5% самых больших и 5% самых маленьких значений выборки (такая оценка известна как "trimmed mean" — подрезанное среднее)

Распределение Коши (без сдвига) можно промоделировать как отношение двух независимых стандартных нормальных случайных величин или воспользоваться готовой функцией. *Ниже есть расширенный вариант этой задачи. В качестве интересного теоретического упражнения предлагается проверить состоятельность оценок формально*

Во второй части есть еще несколько задач на моделирование.

1.2 Бумажные задачи

Задание 1.3. Построить оценки параметров по методу моментов и по методу максимального правдоподобия для геометрического распределения и распределения Пуассона.

Задание 1.4. Рассмотрим распределение с плотностью $f(x) = \alpha x^{\alpha-1}$, заданное на $[0,\, 1]$, так называемое *степенное* (это частный случай Бета-распределения, $\mathcal{B}(\alpha, 1)$). Построить оценки для параметра α по методу моментов и по методу максимума правдоподобия.

Задание 1.5. Пусть $X_1, \ldots, X_N$ — выборка из известного непрерывного распределения (с функцией распределения F и плотностью f). В терминах функции распределения и плотности выразить функцию распределения и плотность максимума и минимума выборки ($\xi = \max_i X_i, \eta = \min_i X_i$) *Ниже есть более общий вариант этой задачи*

Задание 1.6. Дана выборка $X_1, \ldots X_N$ из равномерного распределения $U[0,\, \theta]$ с параметром $\theta > 0$ Проверить состоятельность и несмещённость оценок:

1. $\widehat{\theta} = \max_i X_i$

2. $\widehat{\theta} = \max_i X_i + \min_i X_i$

2 Дополнительные задачи

Задание 2.1. Рассмотрим обобщенное распределение Коши, т.е. распределение с плотностью

$$f(x) = \frac{1}{\pi s \left(1 + \frac{(x-a)^2}{s^2}\right)}$$

Пусть мы имеем выборку $X_1, \ldots, X_N$ из этого распределения и хотим оценить параметры *сдвига* и *масштаба*.

Для хороших распределений оценка сдвига и масштаба делается тривиально по первым двум моментам, но у распределения Коши моменты не существуют.

Убедитесь, что выборочные среднее и дисперсия несостоятельны как оценки в этом случае и затем сравните следующие оценки по с.к.о. (предварительно показав их состоятельность):

1. $\widehat{a} = \mathrm{median}(X_1, \ldots, X_M)$, $\widehat{s} = \mathrm{mad}(X_1, \ldots, X_N)$, где $\mathrm{mad}(X_1, \ldots, X_N) = \mathrm{median}(|X_1 - \mathrm{median}(X_1, \ldots, X_N)|, \ldots, |X_N - \mathrm{median}(X_1, \ldots, X_N)|)$ — median absolute deviation, медиана расстояния от медианы, медианный аналог standard deviation (с.к.о). *будьте внимательны, в R по умолчанию `mad()` считает MAD, домноженный на константу, так, чтобы у нормального распределения получалась* 1.
2. $\widehat{a} = \mathrm{median}(X_1, \ldots, X_N)$, $\widehat{s} = e^{\overline{\log|X-a|}} = e^{\sum_{i=1}^{N} \log|X_i - a|/N}$ (проще говоря, среднее геометрическое модулей центрированных значений выборки)

Задание 2.2. Рассмотрим две оценки дисперсии: выборочную дисперсию $\widehat{\sigma_n^2}$ и исправленную выборочную дисперсию $\widehat{\sigma^2} = \frac{n}{n-1}\widehat{\sigma_n^2}$. Исправленная выборочная дисперсия является несмещенной оценкой дисперсии (не только для нормального распределения), а обычная выборочная дисперсия является только асимптотически несмещенной. Известно, что несмещенность является хорошим, но необязательным свойством для оценки. Хочется все-таки понять, какая из этих оценок лучше. Предлагается провести (для разных распределений: для нормального, равномерного непрерывного и для бросания монетки $U\{0,\,1\}$) сравнение по среднеквадратичному отклонению (aka MSE) с помощью моделирования. *В качестве дополнительного упражнения можно провести теоретический анализ для произвольного распределения*

Задание 2.3. Вычислить асимметрию и экцесс для выборок различных распределений (равномерное, экспоненциальное, Гамма для разных k, нормальное). Проверить свойства эксцесса (асимметрия — мера несимметричности распределения, положительный — значит, уходит направо; большой эксцесс означает острый пик плотности в центре распределения и тяжелые хвосты). Убедиться, что асимметрия и эксцесс может использоватся для предварительной проверки выборки на нормальность. Для Гамма распределения исследовать динамику асимметрии и эксцесса с ростом k и попытаться объяснить ее, исходя из свойств Гамма-распределения и центральной предельной теоремы. *Да, обратите внимание, наименьший возможный экцесс* (-2) *имеет распределение симметричного Бернулли (честная монетка), а наибольший* (∞) *— вырожденное распределение, сосредоточенное в одной точке. Ответ может показаться на первый взгляд странным, но все дело в масштабе, т.е. в знаменателе*

Задание 2.4. Показать, что выборочная функция распределения в каждой точке является несмещенной и состоятельной оценкой реальной функции распределения в соответствующей точке.

Задание 2.5. Дана выборка из равномерного распределения на отрезке $[0,\,\theta]$. Найти предел (в смысле сходимости п.н.) при $k \to \infty$ последовательности оценок параметра θ, полученных методом моментов по k-му нецентральному моменту.

Задание 2.6. Пусть $X_1, \ldots, X_N$ — выборка из известного непрерывного распределения (с функцией распределения F и плотностью f). В терминах функции распределения и плотности выразить функцию распределения и плотность порядковых статистик $X_{[i]}$). *Порядковая статистика — это статистика, занимающая строго определенное место в ранжированной совокупности, т.е.* $X_{[i]}$ *это* i*-й элемент в упорядоченном массиве* X*-ов. Примеры порядковых статистик — минимум, максимум, медиана, выборочные квантили*

3 Задачи, разобранные 28 марта

Задание 3.1. Для стандартного нормального $\mathcal{N}(0,1)$ распределения найти квантили уровня $0.5, 0.95, 0.975, 0.99$ и 0.995.

Решение.

```
qnorm(c(0.5, 0.95, 0.975, 0.99, 0.995))
[1] 0.000000 1.644854 1.959964 2.326348 2.575829
```

Задание 3.2. Построить оценки для параметра экспоненциального распределения по методу моментов и по методу максимума правдоподобия. Сравнить. Будут ли эти оценки несмещенными?

Задание 3.3. Построить оценки для параметров Гамма-распределения по методу моментов и по ОМП *ОМП оценки выписаны здесь:* `http://en.wikipedia.org/wiki/Gamma_distribution#Maximum_likelihood_estimation`

Задание 3.4. Построить оценки для параметров нормального распределения (по ОМП и по методу моментов)

Задание 3.5. Рассмотрим распределение $U[0,\theta]$. Построить оценки (по ОМП и методу моментов) для параметра θ.

Задание 3.6. Рассмотрим выборку $X_1, \ldots X_N$ из распределения Бернулли с неизвестным параметром p. Является ли функция $\widehat{p}(X_1, \ldots, X_N) = X_1$ состоятельной оценкой параметра p? А несмещенной оценкой?

Задание 3.7. Найти значения моментов для стандартного нормального распределения

Решение. Центральные моменты совпадают с нецентральными, так как стандартное нормальное имеет матожидание 0. Обозначим за M_k k-й момент и выпишем его по определению:

$$M_k = \mathsf{E}\xi^k = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x^k e^{-x^2/2} dx =$$

Применим формулу интегрирования по частям:

$$= \frac{1}{k+1} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx^{k+1} = \frac{1}{k+1} \cdot \frac{1}{\sqrt{2\pi}} \left(x^{k+1} e^{-x^2/2} \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} x^{k+1} de^{-x^2/2} \right) =$$

Двойная подстановка обнуляется:

$$= \frac{1}{k+1} \cdot \frac{1}{\sqrt{2\pi}} \left(- \int_{-\infty}^{\infty} x^{k+1} de^{-x^2/2} \right) = \frac{1}{k+1} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} -x^{k+1} e^{-x^2/2} (-x) dx =$$
$$= \frac{1}{k+1} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x^{k+2} e^{-x^2/2} dx = \frac{1}{k+1} M_{k+2}$$

Таким образом $M_{k+2} = (k+1)M_k$, следовательно, для всех четных k: $M_k = (k-1)!!$, а для всех нечетных $M_k = 0$.