Problem 1. Пусть $x_1, x_2, \ldots, x_N$ – независимая выборка из непрерывного равномерного распределения $\mathrm{U}[0, \theta]$. Требуется найти оценку максимального правдоподобия θ_{ML}, подобрать сопряжённое распределение $p(\theta)$, найти апостериорное распределение $p(\theta|x_1, \ldots, x_N)$ и вычислить его статистики: мат.ожидание, медиану и моду. Формулы для статистик нужно вывести, а не взять готовые. *Подсказка: задействовать распределение Парето.*

Запишем правдоподобие:

$$p(x_1, \ldots, x_N|\theta) = \frac{1}{\theta^n}[0 \leqslant x_{(1)}, x_{(N)} \leqslant \theta],$$

где $x_{(1)}$ и $x_{(n)}$ минимальное и максимальное значения выборки соответственно.

Способ взять производную и приравнять к нулю тут не работает. Но можно заметить, что функция убывающая, а значит максимум по θ достигается в самой левой точке. Но раз

$$x_{(n)} \leqslant \theta \implies \frac{1}{x_{(n)}^n} \geqslant \frac{1}{\theta^n},$$

то $x_{(N)}$ максимайзер и

$$\theta_{ML} = x_{(N)}.$$

Теперь проверим, является ли распределение Парето (нотация такая же, как в 3 задаче) сопряжённым к равномерному:

$$p(x|\theta)p(\theta|a, b) = \frac{1}{\theta}[x \in [0, \theta]]\frac{C}{\theta^{b+1}}[\theta \geqslant a] = \frac{C}{\theta^{b+2}}[\theta \geqslant \max(x, a)].$$

Видно, что функциональный вид распределения не поменялся, а значит это действительно сопряжённое распределение.

Теперь найдём апостериорное распределение:

$$p(\theta|x_1, \ldots, x_N) = \frac{1}{Z}\frac{1}{\theta^n}[0 \leqslant x_{(1)}, x_{(N)} \leqslant \theta]\frac{C}{\theta^{b+1}}[\theta \geqslant a] = \frac{1}{Z'}\frac{1}{\theta^{b+n+1}}[\theta \geqslant \max(x_{(N)}, a)]$$

$$p(\theta|x_1, \ldots, x_N) \sim \mathrm{Pareto}(\theta|\max(x_{(N)}, a), b + n)$$

В дальнейшем вместо $\max(x_{(N)}, a)$ будем писать max.

Теперь считаем статистики: начнём с мат.ожидания. Распишем по определению:

$$\mathbb{E}[\theta|x_1, \ldots, x_N] = \int_{-\infty}^{+\infty} \theta \cdot \frac{(b+n)\max^{b+n}}{\theta^{b+n+1}}[\theta \geqslant \max]\, d\theta = (b+n)\max{}^{b+n}\int_{\max}^{+\infty}\frac{1}{\theta^{b+n}}d\theta = (b+n)\max{}^{b+n}\frac{\theta^{1-b-n}}{1-b-n}\Bigg|_{\max}^{+\infty} =$$

$$= (b+n)\max{}^{b+n}\frac{\max^{1-b-n}}{b+n-1} = \frac{\max\cdot(b+n)}{b+n-1}$$

Для медианы нам нужна CDF для распределения Парето. Посчитаем её:

$$F(x) = \int_{-\infty}^{x} p(y|a, b)dy = ba^b\int_a^x \frac{dy}{y^{b+1}} = ba^b\left(\frac{x^{-b}}{-b} - \frac{a^{-b}}{-b}\right) = 1 - \left(\frac{a}{x}\right)^b$$

Зная, что $F(x_{median}) = \frac{1}{2}$, получаем:

$$1 - \left(\frac{\max}{x_{median}}\right)^{b+n} = \frac{1}{2} \iff \left(\frac{\max}{x_{median}}\right)^{b+n} = \frac{1}{2} \iff x_{median} = \max\cdot\sqrt[b+n]{2}$$

Моду ищем также, как и θ_{ML}: плотность распределения Парета это убывающая функция, а значит максимальное значение она принимает в самой левой точке, то есть max.

Problem 2. Предположим, что вы приезжаете в новый город и видите автобус с номером 100. Требуется с помощью байесовского подхода оценить общее количество автобусных маршрутов в городе. Каким априорным распределением стоит воспользоваться (обоснуйте выбор его параметров)? Какая из статистик апостериорного распределения будет наиболее адекватной (обоснуйте свой выбор)? Как изменятся оценки на количество автобусных маршрутов при последующем наблюдении автобусов с номерами 50 и 150? *Подсказка: воспользоваться результатами предыдущей задачи. При этом обдумать как применить непрерывное распределение к дискретным автобусам.*

Пусть номера автобусов распределены $\mathrm{U}[0, \theta]$. Тогда в качестве prior удобно взять распределение Парето: во-первых, если верить этому, то параметр a отвечает за максимальное значение наблюдения (в нашем случае это максимальный номер автобуса), а параметр b за количество наблюдений (встреч автобусов); во-вторых это conjugate prior к равномерному распределению. Значит $p(\theta) = \mathrm{Pareto}(\theta|100, 1)$. Чтобы перейти от непрерывных распределений, будем брать целую часть.

Что касается статистик: в прошлой задаче мы нашли их вид. Мода всегда будет выдавать максимум из увиденных номеров. Мат.ожидание при одном наблюдении не определено, так что проверку на адекватность не прошло, но при большом количестве будет стремиться к моде. Медиана при первом наблюдении удвоит максимальный увиденный номер, но при большом количестве наблюдейний также будет стремиться к максимому (только опять же надо брать целую часть). Если честно, я не могу сказать, что более адекватно: мода или медиана, поэтому предлагаю использовать обе статистики.

Опять же, если следовать логике, описанной в первом абзаце, то пронаблюдав автобусы с номерами 50 и 150, распределение изменится до $p(\theta|x_1, x_2) = \mathrm{Pareto}(\theta|150, 3)$. В таком случае мода станет равной 150, медиана $150 \cdot \sqrt[3]{2} = 188$, а мат.ожидание $150 \cdot \frac{3}{2} = 225$.

Problem 3. Записать распределение Парето с плотностью $\mathrm{Pareto}(x|a, b) = \frac{ba^b}{x^{b+1}}[x \geq a]$ при фиксированном a в форме экспоненциального класса распределений. Найти $\mathbb{E} \log x$ путём дифференцирования нормировочной константы.

Напомним форму экспоненциального класса распределений:

$$p(x|\theta) = \frac{f(x)}{g(\theta)} \exp\left[\theta^\top u(x)\right].$$

Тогда понятно, что $f(x) = [x \geqslant a]$, $g(\theta) = \frac{1}{ba^b}$, дальее перепишем в терминах θ.

Теперь разберёмся с exp:

$$\frac{1}{x^{b+1}} = \exp\left[\log \frac{1}{x^{b+1}}\right] = \exp[-(b+1)\log(x)] = \exp[\theta^\top u(x)],$$

где

$$u(x) = \begin{bmatrix} \log x \\ 0 \end{bmatrix}, \theta = \begin{bmatrix} -1-b \\ a \end{bmatrix},$$

тогда $g(\theta) = \frac{1}{-(\theta_1+1)\theta_2^{-(\theta_1+1)}}$

Теперь найдём $\mathbb{E} \log x$:

$$\mathbb{E} \log x = \mathbb{E} u_1(x) = \frac{\partial}{\partial \theta_1} \log g(\theta) = \frac{\partial}{\partial \theta_1} \log \frac{1}{-(\theta_1+1)\theta_2^{-(\theta_1+1)}} = \frac{\partial}{\partial \theta_1} \left[-\log(-(\theta_1+1)) + (\theta_1+1)\log\theta_2\right] =$$

$$= -\frac{-1}{-(\theta_2+1)} + \log\theta_2 = \frac{1}{b} + \log a$$