Санкт - Петербургский государственный унивеситет

Кафедра моделирования электромеханических и компьютерных систем

Гриднев Валерий Анатольевич

Выпускная квалификационная работа бакалавра

Оценивание параметров модифицированного бета-распределения

Направление 010900

Прикладная математика и физика

научный руководитель,

доктор физ.-мат. наук,

профессор

Шмыров А.С

Санкт-Петербург

2017 г.

Содержание

1 Введение

Все мы знаем, как быстро в 21 веке развивается вычислительная техника, в связи с этим повышаются возможности применения методов таких наук, как теория вероятностей и математическая статистика. И если во время зарождения теории вероятностей ее применение виделось только в страховании и азартных играх, то сейчас её области использования вместе с мат. статистикой практически безграничны. Прикладные возможности этих наук могут поразить не интересующегося человека своими масштабами. В данной работе решается задача моделирования и оценивания параметров для модифицированного бета-распределения случайной величины.

Важность и актуальность этого исследования вытекает из разнообразия случаев использования бета–распределения для решения множества прикладных задач. Применение начинается от рейтинговой оценки фильмов и тестирования средств биометрической аутентификации, до моделирования экстремальных событий, например таких, как уровни наводнений, скорости вихрей, максимума индексов ценных бумаг. Суть моей задачи состоит в том, чтобы имея неполную информацию, получить максимально эффективную оценку ситуации.

Для решения данной задачи будет использован принцип максимума энтропии. В современной формулировке «принцип максимальной энтропии» был выдвинут Э.Т. Джейнсом, начиная с середины

пятидесятых. Полное описание принципа приводится в его книге [3], конечная формулировка будет приведена в третьей главе(раздел 3.1). До этого понятие энтропия использовалось в физике для описания термодинамических систем. После формулировки Э.Т. Джейнса принцип максимальной энтропии позволил решать сложные задачи статистики, что и используется в нашем исследовании.

При реализации принципа максимума энтропии используется метод множителей Лагранжа, который позволяет перейти от условной оптимизации к безусловной. Данный переход позволяет написать решение задачи оптимизации в параметрическом виде. Однако стоит заметить, что после применения метода Лагранжа остается проблема определения параметров.

2 Постановка задачи и методика решения

Предположим, что решается задача исследования распределения вероятности случайной величины ξ с помощью выборочного метода, то есть на основе выборки $\xi_1, \ldots, \xi_n$. Случайные величины $\xi_1, \ldots, \xi_n$ независимы, каждая из них имеет одинаковое распределение. Мы можем понимать выборку как результат эксперимента по наблюдениям за случайной величиной. При обычных условиях элементы выборки имеют бета–распределение. Однако может оказаться так, что значения выборки из некоторого промежутка недоступны для наблюдения. Для определения неизвестного распределения, мы применим принцип максимума энтропии, обоснованием этому служит то, что само бета–распределение является решением задачи о максимуме энтропии при связях, выражающихся через математические ожидания функции случайных величин. Принцип распространяется на наш случай, параметры связи рассчитываются и в случае недоступности наблюдений, тоже касается и энтропии. В результате полученное распределение назовем модифицированным бета–распределением и поставим задачу оценивания его параметров. Проведя анализ этого распределения мы придем к выводу, что оно относится к известному экспоненциальному семейству, следовательно имеется возможность получения эффективных оценок для его

параметров. Параметры связи мы оценим с помощью статистик типа выборочного среднего. Затем выпишем уравнения, по которым из параметров связи можно получить параметры плотности искомого модифицированного бета–распределения.

3 Бета-распределение

3.1 Необходимые понятия теории вероятностей и математической стаистики

Для дальнейшей работы введем основные понятия теории вероятности и математической статистики, используемые в данной работе. Прежде всего, вспомним такой термин как распределение случайной величины.

Пусть задано вероятностное пространство $(\Omega, F, \mathbb{P})$,
$\xi : \Omega \to \mathbb{R}$—случайная величина, а P_ξ—распределение вероятностей случайной величины ξ. Вероятностную меру P_ξ можно задать с помощью функции F_ξ одного переменного x, положив

$$F_\xi = P(\xi(\omega) < x) \tag{1}$$

тогда

$$P_\xi([a, b]) = P(a \leq \xi(\omega) < b) = F_\xi(b) - F_\xi(a) \tag{2}$$

и поскольку борелевская прямая есть минимальная сигма-алгебра системы полуинтервалов, то по теореме Каратеодори формула Ляпунова (2) определяет распределение P_ξ на $(\mathbb{R}, B(\mathbb{R}))$

Определение 1 Функция F_ξ называется *функцией распределения случайной величины* ξ

Получив закон, описывающий область значений случайной величины и вероятность их появления, обратим внимание на то,

что есть несколько способов задания распределения. Так как мы рассматриваем бета–распределение, то нам важно, что у бета–распределения есть плотность.

Оперделение 2 Случайная величина ξ называется *абсолютно непрерывной* (относительно меры Лебега), если существует функция $f_\xi : \mathbb{R} \to \mathbb{R}_+$,
что $\forall B \in B(\mathbb{R})$

$$P_\xi(B) = \int\limits_B f_\xi(x)dx \tag{3}$$

Функция f_X называется *плотностью распределения.*

Примечание Распределение называется непрерывным, если функция распределения непрерывна.

Отметим, что только абсолютно непрерывные распределения имеют плотность. Именно с плотностью распределения мы будем работать после модификации стандартного бета–распределения.

Также нужно обратить внимание на такие, понятия как параметры распределения и оценка.

Параметры являются теоретическими величинами, недоступными для измерения, но возможными для оценки. В генеральной

совокупности параметры представляют ее количественную характеристику и определяются только в результате теоретического моделирования. Для определения их на практике нужно осуществлять выборочную оценку, которая в свою очередь предполагает статистический подсчет. Статистика нужна либо для оценки параметров распределения случайной величины, либо для описания самой выборки. Статистика основывается на исследовании выборочных значений, и является количественной характеристикой исследуемых параметров.

Для избегания ошибок в неверном толковании данных в ходе эксперимента, очень важно разделять понятия «статистика» и «параметр». Между параметрами и статистикой всегда существуют различия, оценить мы их не можем. Теоретически, чем большего объема наша выборка, тем ближе оцениваемые параметры к выборочных характеристикам. Это не значит, что при увеличении выборки мы уменьшим разницу между ними.

Определение 3 *Несмещенная оценка* – точечная оценка, математическое ожидание которой равно оцениваемому параметру.

Если же значение не совпадает, то такую оценку называют *смещенной*. Забегая вперед, примем во внимание, что оценка будет не только несмещенной , но и эффективной.

Определение 4 *Эффективная оценка* – статистическая оценка параметра, которая при заданной выборке имеет наименьшую дисперсию.

Нахождение функции плотности модифицированного бета–распределения произведем с помощью метода множителей Лагранжа. Аналогичным образом в [1] были найдены функции плотностей для нормального и гамма распределения. Напомним определения этих распределений.

Определение 5 Распределение вероятностей с плотностью вида

$$\gamma(x,\alpha,\lambda) \;=\; \begin{cases} 0 & \text{если } x \le 0 \\ \frac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x} & \text{если } x > 0 \end{cases} \tag{4}$$

называется *Г-распределением.* При этом параметры λ, α - положительные.

Определение 6 Абсолютно непрерывная случайная величина ξ нарывается *нормально распределенной*, если ее плотность f_ξ имеет вид

$$f_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}, \tag{5}$$

где m, σ - действительные параметры и $\sigma > 0$

Для ясности метода множителей Лагранжа, рассмотрим его на примере нормального и гамма-распределений.

Определение 7 Назовем *энтропией случайной величины* ξ (обозначение H_ξ) интеграл Лебега

$$H_\xi = \int f_\xi(x) ln f_\xi(x) dx, \tag{6}$$

если этот интеграл существует.

Принцип диктует, что нужно искать распределение, соответствующее доступной информации, которое максимизирует энтропию. Использовать этот принцип можно в силу вариационного свойства, которое формулируется в [1] следующим образом.

Вариационное свойство: экспоненциальное распределение доставляет максимум энтропии на классе абсолютно непрерывных распределений случайных величин, принимающих неотрицательные значения и имеющие фиксированное математическое ожидание.

Теперь рассмотрим метод множителей Лагранжа для нормального и гамма распределений.
Запишем энтропию

$$H_\xi = \int f_\xi(x) ln f_\xi(x) dx \tag{7}$$

Плотность случайной величины f_ξ должна удовлетворять условиям (1 - условие нормировки, 2,3 - уравнения связи):

$$\int f(x)dx = 1 \tag{8}$$

$$\int xf(x)dx = m \tag{9}$$

$$\int x^2 f(x)dx = m^2 + \sigma^2 \tag{10}$$

Теперь введем множители $\lambda_0, \lambda_1, \lambda_2$ и образуем функцию Лагранжа

$$L = \int (-f(x)lnf(x)) + \lambda_0 f(x) + \lambda_1 x f(x) + \lambda_2 x^2 f(x))dx \tag{11}$$

Продифференцировав подынтегральное выражение по f, получаем для плотности $f^*(x)$ реализующий максимум

$$f^*(x) = -lnf^*(x) - 1 + \lambda_0 + x\lambda_1 + x^2\lambda_2 = 0, \tag{12}$$

откуда

$$f^*(x) = exp(-1 + \lambda_0 + x\lambda_1 + x^2\lambda_2) \tag{13}$$

И теперь, для того чтобы эта формула задавала плотность нормального распределения, нужно, чтобы $\lambda_2 < 0$, тогда (формула) задает нормальную плотность, которая определяется, если заданы дисперсия и математическое ожидание. В итоге получаем

$$f^*(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} \tag{14}$$

Теперь рассмотрим гамма-распределение.

Условие нормировки и уравнения связи:

$$\int_0^\infty f(x)dx = 1 \tag{15}$$

$$\int_0^\infty xf(x)dx = m \tag{16}$$

$$\int_0^\infty lnxf(x)dx = l \tag{17}$$

Как и для нормального распределения, зададимся множителями Лагранжа $\lambda_0, \lambda_1, \lambda_2$ и образуем функцию

$$L = \int_0^\infty (-f(x)lnf(x) + \lambda_0 f(x) + \lambda_1 xf(x) + \lambda_2 lnxf(x))dx \tag{18}$$

Подынтегральное выражение является выпуклой вверх функцией от f и имеет единственный максимум при $f = f^*$, который определяем приравнивая к нулю производную по f от подынтегрального выражения

$$-lnf^*(x) - 1 + \lambda_0 + \lambda_1 x + \lambda_2 lnx = 0 \tag{19}$$

откуда

$$f^*(x) = exp(-1 + \lambda_0 + \lambda_1 x + \lambda_2 lnx) \tag{20}$$

3.2 Бета-распределение

Определение 8 Бета-распределение – двухпараметрическое семейство абсолютно непрерывных распределений.

Бета-распределение задается двумя параметрами, но в отличие от нормального распределения (которое всегда имеет одинаковую форму) оно более гибкое. При $\alpha = \beta = 1$ бета–распределение является равномерным, при $\alpha > 1, \beta > 1$ похоже на нормальное, а при $\alpha < 1 \beta < 1$ имеет форму колодца.

Название бета–распределение связно с бета–функцией Эйлера $B(\alpha, \beta)$, которая определяется по формуле

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1 - x^{\beta-1})dx \tag{21}$$

для $\alpha > 0, \beta > 0$

Плотность бета-распределения в свою очередь имеет вид

$$f(x) \;=\; \begin{cases} 0 & \text{если } \; x \notin (0,1) \\ \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{если } \; x \in [0,1], \alpha, \beta > 0 \end{cases} \tag{22}$$

Условие нормировки

$$\int_0^1 f(x)dx = 1 \tag{23}$$

Для дополнительного анализа рассмотрим связь параметров распределения с моментами. Обратим внимание, что условие нормировки

приводит к соотношению

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \tag{24}$$

Из этого соотношения получим два момента

$$M\xi = \int_0^1 x f(x) dx = \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + 1)}{\Gamma(\alpha)\Gamma(\alpha + \beta + 1)} = \frac{\alpha}{\alpha + \beta} \tag{25}$$

$$M\xi^2 = \int_0^1 x^2 f(x) dx = \frac{\alpha(\alpha + 1)}{(\alpha + \beta + 1)(\alpha + \beta)} \tag{26}$$

4 Модифицированное бета-распределение

4.1 Введение понятия модифицированного бета–распределения. Применение принципа максимума энтропии. Вывод системы интегральных уравнений

Так как бета–распределение является решением задачи о максимуме энтропии (разобрано выше),то и на случай с неполной информацией (бета–распределение на отрезке $[0,1]$ без интервала $< a, b >$) распространяется принцип максимума энтропии. Сгенерировав выборку и отбросив из нее значения из интервала $< a, b >$, мы задаем параметры. В качестве параметров мы возьмем математические ожидания в таком виде

$$Mln(\xi) = l_1, \quad Mln(1-\xi) = l_2 \tag{27}$$

запишем энтропию

$$H_\beta = -(\int_0^a f_\beta(x) ln f_\beta(x) dx + \int_b^1 f_\beta(x) ln f_\beta(x) dx) \tag{28}$$

Сформулируем вариационную задачу для нашего случая: найти неотрицательную функцию f, определенную на множестве

$M = [0, a] \cup [b, 1]$, удовлетворяющую условиям (1 - условие нормировки, 2,3 - уравнения связи)

$$\int_0^a f_\beta(x)dx + \int_b^1 f_\beta(x)dx = 1 \tag{29}$$

$$\int_0^a ln(x) f_\beta(x)dx + \int_b^1 ln(x) f_\beta(x)dx = l_1 \tag{30}$$

$$\int_0^a ln(1-x) f_\beta(x)dx + \int_b^1 ln(1-x) f_\beta(x)dx = l_2 \tag{31}$$

и доставляющую максимум энтропии.

Воспользовавшись методом множителей Лагранжа, получаем функцию плотности, распределение с таким видом плотности назовем модифицированным бета-распределением.

$$f_\beta = Cx^{\lambda_1}(1-x)^{\lambda_2} \tag{32}$$

В данном случае константа и коэффициенты λ_1, λ_2 неизвестны. Подставим эту функцию в уравнение связи

$$\int_0^a Cx^{\lambda_1}(1-x)^{\lambda_2}dx + \int_b^1 Cx^{\lambda_1}(1-x)^{\lambda_2}dx = 1 \tag{33}$$

$$\int_0^a ln(x)Cx^{\lambda_1}(1-x)^{\lambda_2}dx + \int_b^1 ln(x)Cx^{\lambda_1}(1-x)^{\lambda_2}dx = l_1 \tag{34}$$

$$\int_0^a ln(1-x)Cx^{\lambda_1}(1-x)^{\lambda_2}dx + \int_b^1 ln(1-x)Cx^{\lambda_1}(1-x)^{\lambda_2}dx = l_2 \tag{35}$$

Получена система из трех уравнений с тремя неизвестными, так как не в общем случае интервал $< a, b >$ нам известен, а параметры l_1, l_2 считаются из выборки.

Стоит также заметить что стандартное бета–распределение, и полученное модифицированное бета–распределение входят в семейство экспоненциальных распределений. Обратимся к [2] и вспомним определение экспоненциального семейства.

Пусть $\theta = \theta_1, \ldots, \theta_k$ $k-$мерный параметр и плотность $f_\theta(x)$ представляется в виде

$$f_\theta(x) = h(x)exp(\overset{k}{_{j=1}}\alpha_j(\theta)U_j(x) + V(\theta)) \tag{36}$$

где все функции, входящие в правую часть, конечны и измеримы.

Так как наше распределение входит в семейство экспоненциальных распределений, то следовательно мы можем получить эффективные оценки наших параметров. Оценивать мы будем с помощью статистик типа выборочного среднего.

$$l_1^* = \frac{1}{n}\overset{n}{_{k=1}} ln(x_k) \tag{37}$$

$$l_2^* = \frac{1}{n}\overset{n}{_{k=1}} ln(1 - x_k) \tag{38}$$

Оценка математических ожиданий выборочным средним является эффективной и несмещенной оценкой. Это следует из неравенства Рао – Крамера.

4.2 Результаты и выводы

В результате работы было показано, что метод основанный на принципе максимума энтропии допускает довольно простую алгоритмизацию. Эта алгоритмизация была реализована на примере выборки специального вида, компоненты которого имеют модифицированное бета–распределение. Для оценивания моментов определенных уравнениями связи были предложены эффективные оценки. Оценки получены в явном виде. Для определения параметров описывающих плотность модифицированного бета–распределения, были выведены интегральные уравнения вида.

5 Заключение

В данной работе проведено исследование о возможности оценивания параметров распределения с неполной информации. В ходе работы введено и обосновано понятие модифицированного бета-распределения. В результате работы предложен алгоритм получения такого распределения на основе принципа максимума энтропии с использованием метода множителей Лагранжа. Получены эффективные оценки параметров модифицированного бета-распределения. Составлена система интегральных уравнений связывающих параметры плотности модифицированного бета–распределения и эмпирические моменты.

6 Список литературы

1.Шмыров А.С., Шмыров В.А. Теория вероятностей: учебное пособие.2012. стр. 162-180.

2. Боровков А.А. Математическая статистика: Учебник.4-е изд. 2010. Стр.178-200.

3. E.T. Jaynes. Probably theory: The logic of since. 1995. Chapter 11.

4. Боровков А.А. Теория вероятностей. 3-е издание. 1999.