Б.М. Пранов
(Академия Государственной противопожарной службы МЧС России;
e-mail: boris.pranov@gmail.com)

О ПРИМЕНЕНИИ ЭКСТРЕМАЛЬНЫХ СТАТИСТИК В ПРОБЛЕМАХ ТЕХНОСФЕРНОЙ БЕЗОПАСНОСТИ

Аннотация. Показано, что для решения ряда важных проблем техносферной безопасности можно использовать порядковые статистики и статистики экстремальных значений.

Ключевые слова: порядковые статистики, экстремальные статистики, проблемы техносферной безопасности.

B.M. Pranov

ABOUT USE EXTREME STATISTICS IN PROBLEMS OF TECHNOSPHERIC SAFETY

Abstract. It is shown that for the solution of important problems of technospheric safety can use the order statistics and the statistics of extreme values.

Key words: order statistics, statistics of extreme values, problems of technospheric safety.

Статья поступила в редакцию Интернет-журнала 12 октября 2010 г.

Хорошо известно, что изучение экстремальных значений является очень важной стороной практических исследований – как часто максимальное значение в последовательности наблюдений встречается на практике? Задача статистической теории экстремальных значений состоит в анализе наблюденных экстремумов и предсказании тех экстремумов, которые могут иметь место при последующих наблюдениях.

Экстремумы не являются фиксированными величинами – это новые случайные величины, зависящие от исходного распределения и от объема выборки. С подобными вопросами встречаются проектировщики гидротехнических сооружений при прогнозе катастрофических паводков (на реках) и возможной мощности лавин и селей (в ущельях), геофизики и строители – при оценке максимальной силы землетрясений, проектировщики линий электропередач – при расчёте максимальной нагрузки электросетей промышленных предприятий [1, 2].

Экстремальные величины, как и можно предполагать заранее, связаны с малыми вероятностями. Статистический анализ экстремумов предназначен для ответа на вопросы двух типов:

- не выпадает ли за разумно ожидаемые границы какое-либо отдельное наблюдение в выборке из совокупности, распределение которой предполагается известным?

- проявляет ли последовательность экстремальных значений регулярное поведение? В обоих случаях слова "разумно ожидаемые" и "регулярное поведение" должны быть определены строго как математические понятия.

Поскольку предсказать максимальный размер переменной случайной величины мы не можем, для оценки квантиля 99,9 % следует обратиться к теории. На языке математики это означает: необходимо найти аналитическое продолжение эмпирической функции распределения потерь. Область математики, изучающая "выбросы" наблюдений, получила название "теория экстремальных значений" (Extreme Value Theory, EVT). Наиболее важным выводом EVT, с точки зрения нахождения высоких квантилей распределения потерь, является теорема Балкемы-де Хаана и Пикэндса о сходимости функции распределения превышений порогового значения к обобщенному распределению Парето. Согласно данной теореме, условное распределение превышения потерями некоторого достаточно высокого порога асимптотически описывается функцией, носящей название обобщенного распределения Парето (Generalized Pareto Distribution, GPD).

Рассмотрим постановку аналогичных задач, касающихся вопросов пожарной безопасности и других направлений.

Задача 1. Проанализируем последовательность числа пожаров, произошедших в некотором городе за один день: $x_1, x_2, \ldots, x_n$, где n – количество дней наблюдения. Пусть s_k – некоторая случайная величина, связанная с этой последовательностью, например, наибольшая площадь пожара в день с номером k, где $k \in [1 \div n]$. В таком случае исследование закона распределения величины s_k может позволить с заданной вероятностью вычислить минимально необходимое количество сил и средств для ликвидации пожара максимальной площади.

Ясно, что неограниченное увеличение количества средств защиты нерентабельно. В связи с этим возникает ряд формулировок – минимизация ошибки второго рода при определении минимально необходимого количества средств защиты, определение минимума необходимого числа средств при заданном риске не справиться с пожаром максимальной площади за нормативное время и другие разновидности формулировок.

Отметим ряд особенностей постановки этой задачи. Можно в качестве случайной величины s_k рассматривать максимальный ущерб, количество пострадавших человек, в том числе погибших. Тогда постановка задачи приобретает совсем другой смысл, хотя математическая формулировка остается прежней.

Задачу 1 можно решать при условии, что последовательные значения случайной величины s_k образуют:

а) независимую последовательность;

б) марковскую цепь (или имеют более глубокий период временной зависимости).

Распределение величины s_k можно определять при условии, что последовательность случайных величин $x_1, x_2, \ldots, x_n$ с высокой точностью подчиняется

а) некоторому простому теоретическому распределению (например, распределению Пуассона);

б) некоторому эмпирическому распределению.

Интересно сравнить отклонения расчётов в случаях а) и б).

Итак, пусть s_1, s_2, … – последовательность независимых и одинаково распределенных случайных величин, а M_k – максимум первых k из этих величин:

$$M_k = \max(s_1, ..., s_k).$$

Большая часть классической теории экстремальных значений имеет дело с распределением M_k, особенно с его свойствами при $n \to \infty$. Очевидно, что термин "максимум" можно заменить на термин "минимум" в связи с возможностью замены знака величин на обратный. Функция распределения величины M_k по очевидным соображениям равна

$$P\{M_k \leq x\} = P\{s_1 \leq x, ..., s_k \leq x) = F^k(x),$$

где $F(x)$ обозначает общую функцию распределения для s_i.

Ясно, что для практического рода задач эта функция распределения $F(x)$ должна находиться путем оценивания статистических наблюдений (поскольку, например, нет никаких теоретических соображений для регистрируемой максимальной величины площади пожара).

Большая часть книги Гумбеля [1] имеет дело с распределением M_k в целом ряде типовых случаев и с множеством родственных вопросов (например, относительно других порядковых статистик, размаха и т.д.).

В связи с этим может возникнуть вопрос о желательности получения асимптотических результатов. Один из доводов в пользу такого рода исследования представляется особенно убедительным. В обычной теории центральной предельной теоремы асимптотически нормальное распределение суммы многих независимых нормально распределенных случайных величин получают независимо от того, какова их исходная функция распределения.

Фактически, чтобы применять асимптотическую теорию, вовсе не обязательно знать эту функцию распределения очень точно. Подобная ситуация имеет место и в теории экстремальных значений. Невырожденное асимптотическое распределение (соответствующим образом нормализованное) M_k обязательно должно принадлежать одному из трех единственно возможных общих семейств независимо от исходной функции распределения $F(x)$.

Кроме того, нет никакой необходимости знать функцию распределения $F(x)$ полностью, чтобы определить, к какой предельной форме (если таковая существует) она приводит, то есть к какой "области притяжения" она принадлежит. В действительности это определяется только поведением функции $F(x)$ для больших x, так что об асимптотических свойствах максимума можно сказать многое, основываясь лишь на довольно ограниченной информации о свойствах функции распределения $F(x)$.

Центральный результат впервые был получен Фишером и Типпетом в 1928 г. и позднее был доказан в полной общности Гнеденко в 1943 г. Оказывается, каждое так называемое максимум-устойчивое предельное распределение с точностью до линейной замены аргумента имеет одну из трех параметрических форм, которые называются распределениями экстремальных значений:

тип I. $G(x) = \exp(-e^{-x}), -\infty < x < \infty;$

тип II. $G(x) = 0$ при $x \le 0;$

$G(x) = \exp(-x^{-\alpha}),$ для некоторого $\alpha > 0$, $x > 0$;

тип III. $G(x) = \exp(-(-x)^{-\alpha}),$ для некоторого $\alpha > 0$, $x \le 0$;

$G(x) = 1$ при $x > 0$.

Задача 2. Условия остаются такими же, как и в задаче 1, только величина s_k представляет собой какую-либо характеристику α-процентной случайной величины x_k. Исследуя оценку распределения величины s_k при различных значениях α, можно изучать вероятностные характеристики пожаров различных номеров и, возможно, дать им теоретические обоснования с помощью научных методов.

Задача 3. В качестве случайной величины s_k можно выбирать площадь (максимальную или α-процентиль максимальной), уровень загрязненности водоема или другую величину. Так что приведенная постановка задачи имеет прямое отношение к анализу как пожарной безопасности, так и техносферной.

Порядковые и экстремальные статистики достаточно хорошо изучены, формулы для их вычисления приведены во многих учебниках по теории вероятностей и математической статистике [4]. К сожалению, формулы для них настолько сложны, что почти во всех интересных случаях все распределения не вычисляются точно, а задаются в виде интегралов, так что для практических применений следует использовать приближенные вычисления этих интегралов или пользоваться упрощенными оценками из книги [1].

Любопытно отметить, что первые задачи, связанные с экстремальными величинами, возникли в связи с анализом паводков. Их важность с точки зрения экономики была осознана давно, так как сельское хозяйство прошлых времен опиралось исключительно на изобилие воды, а водные пути являлись основным видом коммуникаций. Жизнь и имущество людей должны быть защищены от вреда, причиняемого наводнениями. Невозможно переоценить значение теории контроля паводков, позволяющей заблаговременно принимать меры предосторожности, что является одной из важнейших задач МЧС.

Литература

1. **Гумбель Э.** Статистика экстремальных значений. М.: МИР, 1965. 452 с.

2. **Ван Данциг Д.** Математические проблемы, возникшие в связи с разрушительным наводнением 1953 г. // В кн. "Международный математический конгресс в Амстердаме 1954 г." (Обзорные доклады). М.: Физматгиз, 1965. С.106-133.

3. **Todhunter J.** History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace (Reprint of 1869 ed.), New York, Chelsea, 1949.

4. **Кендалл М., Стьюарт А.** Теория распределений, т.1. М., Наука, 1966. 566 с.