УДК 004.622

А.Н. Мущук, ассист. (БГТУ, г. Минск)

МЕТОДЫ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ

Для чего нам нужна статистика? Если говорить об ИТ сфере, то статистика используется на многих этапах разработки ПО, от планирования до реализации ключевого функционала. Активно используется в тестировании, например, для предсказывания модулей с потенциально наибольшим количеством ошибок, А/Б тестирования и т. д. В основе машинного обучения, которое является одним из наиболее перспективных направлений ИТ сферы, также лежит статистика.

Статистика – это набор математических методов и инструментов, позволяющих анализировать и структурировать данные, для их наглядного представления, в частности, и для получения возможности их рационального использования в целом. Она делится на две категории:

1. Описательная статистика. Предлагает методы резюмирования данных путем преобразования необработанных наблюдений в значимую информацию, которую легко интерпретировать и распространять.

2. Логическая статистика. Предлагает методы изучения экспериментов, выполненных на маленьких образцах данных, и умозаключения для всего набора информации.

Другими словами, чтобы преобразовать полученную информацию в имеющие смысл идеи, применяется описательная статистика. Затем применяется логическая статистика, чтобы, изучив полученные выборки данных, дать вывод с проведением параллели на всю совокупность данных.

Перейдем к описательной статистике. Как уже было сказано, хоть и немного другими словами, это методы описания выборок, исследуемых по количественному признаку x, с помощью их различных числовых характеристик.

Преимущество данных методов заключается в следующем. Несколько простых и достаточно информативных статистических показателей, если они известны, во-первых, избавляют нас от просмотра сотен, а порой и тысяч значений данных, а во-вторых, позволяют получить более или менее точную оценку характеристик распределения признака в генеральной совокупности.

Описывающие выборку показатели разбиваются на несколько групп; в своем большинстве они имеют аналоги в виде числовых характеристик случайных величин в теории вероятностей.

Показатели положения описывают положение вариантов выборки на числовой оси. Сюда относят:

а) минимальную и максимальную варианту (значение);

б) выборочное среднее арифметическое значение (выборочное среднее), выборочные моду и медиану. Данные значения указывают на «центральную» точку распределения выборки – наиболее значимую для нас варианту в поставленной задаче.

Выборочное среднее $x_в$ является той точкой, сумма отклонений значений x от которой равна нулю. По сути, это просто среднее арифметическое. Это единственная точка, которая обладает данным свойством, оно выделяет ее среди всех других.

Выборочная мода $Mo_в$ – варианта, которая чаще всего встречается в исследуемой выборке, т. е. имеет наибольшую частоту. Если выборочное распределение имеет несколько мод, то говорят, что оно мультимодально. Следует помнить, что при мультимодальном распределении моды не должны иметь строго одинаковые значения. Если выборка имеет несколько показателей, которые явно выбиваются из общей статистики, но при этом имеют разные значения, они все еще являются модами.

Выборочная медиана $Me_в$ – варианта, которая делит упорядоченный статистический ряд на две равные части по числу попадающих в них вариант.

Рассмотрим пример, допустим у нас есть упорядоченная выборка значений 1,3,4,4,4,6,6. Выборочное среднее будет иметь значение (1+3+4+4+4+6+6)/7= 4. Выборочная мода также будет равна четырем, поскольку это наиболее часто встречающееся значение. Медианным значением, в нашем случае, будет являться четвертый элемент выборки, поскольку он является разделителем упорядоченного ряда. В случае, если ряд состоит из четного числа значений, медианное считается как среднее арифметическое между значениями, стоящими посередине, пример: в ряде 12,15,17,17, медианным значением будет (15 + 17) / 2 = 16.

Показатели разброса описывают степень разброса данных относительно своего центра. Здесь обычно используются:

а) стандартное отклонение S и выборочная дисперсия $Dв = S^2$, характеризующие рассеяние вариантов вокруг их среднего выборочного значения $x_в$ и являются самыми распространенными механизмами для измерения и описания разброса величин. Именно эти параметры часто используются как меры изменчивости некоторого исследуемого показателя (случайной величины X). Чем больше Dв и S, тем сильнее разбросаны значения x относительно среднего. Дисперсия

вычисляется путем определения, насколько далеко от среднего значения расположены наблюдения в рамках одного и того же набора данных. Однако вычисление дисперсии идет следующим образом: разница между значением каждого варианта и средним возводится в квадрат, после чего сумма всех таких значений делится на количество этих вариантов Среднеквадратичное отклонение- корень из значения дисперсии. Суть его в том, что если мы отклонимся на это значение влево или вправо на графике значений, мы получим интервал, на котором будут сконцентрированы наиболее вероятные значения выборки; Также стоит отметить, что дисперсия служит исключительно для нахождения стандартного отклонения в описательной статистике.

Рассмотрим на той же упорядоченной выборке. Дисперсия будет считаться следующим образом: $Dв=((1-4)^2+(3-4)^2+(4-4)^2+(4-4)^2+(4-4)^2+(6-4)^2+(6-4)^2)/7=(9+1+0+0+0+4+4)/7=18/7=2,57$.

Стандартное отклонение соответственно считается как корень из 2,57 и равен 1,6.

б) размах выборки – разность между максимальной и минимальной вариантами: $x_{макс} – x_{мин}$;

в) коэффициент вариации $v=S/x_{в}*100\%$. Это отношение стандартного отклонения к средней арифметической для выборки, которое выражено в процентах. Даёт понять, насколько на самом деле велик разброс в данных, независимо от масштаба измерений. Однако, данный показатель не годится для данных, измеренных по интервальной шкале, вроде температуры, времени и т.д.

Так же можно отметить выборочный эксцесс ($Ex_{в}$), который показывает, насколько большая разница между пиковым значением и минимальным, другими словами, он показывает, насколько полученный график будет сглаженным, и выборочную асимметрию $As_{в}$, которая характеризует меру скошенности упорядоченного графика выборки влево или вправо относительно наивысшего значения. Соответственно выделяют правостороннюю, где график справа будет больше вариант чем слева, и левостороннюю, где слева вариант больше, чем справа.

Для определения распределения исследуемой величины, которое нужно знать для проведения последующего анализа, можно использовать гистограммы, графики, полигоны частот и другие способы визуализации информации.

О законе распределения также можно судить по выборочным числовым характеристикам случайной величины.

Большинство методов статистического анализа данных разработано для случайных величин, распределенных по нормальному зако-

ну. Распределение исследуемой величины в генеральной совокупности можно рассматривать как близкое к нормальному, если:

1. Выборочные $x_в$, $Me_в$, $Mo_в$ равны или незначительно отличаются друг от друга.

2. Минимальное и максимальное значения x ($x_{макс}$ и $x_{мин}$) примерно равноудалены от $x_в$.

3. Выборочные $Ex_в$ и $As_в$ близки к нулю.

Подводя итоги, описательная статистика, да и статистика в целом, это, местами, несложная, но весьма хорошая отрасль для работы с данными, способная в умелых руках и при правильном применении облегчить процесс разработки путем упрощения понимания и работы с информацией за счет ее структурирования и облегченного для понимания представления. Ни один крупный проект, на данный момент, не обходится без участия специалиста по данным. И, хотя они не участвуют напрямую в процессе кодирования программного продукта, тем не менее, навыки обработки данных, в том числе статистическими методами, никак нельзя недооценивать.

ЛИТЕРАТУРА

1 Charles Wheelan. Naked Statistics / W. W. Norton & Company; First Edition 2013. C. 35-59.

2 Питер Брюс, Эндрю Брюс. Практическая статистика для специалистов Data Science: Пер. с англ. / П. Брюс, Э. Брюс. – СПб.: БХВ-Петербург, 2018. С. 26-45.

УДК 004:371. 301.5:378.663

Н.И. Потапенко, ст. преп. (БГТУ, г. Минск)

МОРФИЗМ В ВЕБ-ДИЗАЙНЕ

Относительно короткий срок становления и развития веб-дизайна отмечается постоянными изменениями в проектировании и дизайне. Постоянное развитие цифровых технологий затрагивает как техническую составляющую, так и оформительскую в сфере производства веб-изданий. Изменяются способы и формы подачи контента, инструменты взаимодействия с пользователем. Мобильные технологии вносят свои поправки в представления о том, каким должен быть сайт или мобильное приложение.

Веб-дизайн как направление в искусстве не живет своей отдельной жизнью. Веб-сайт – отражение нашей реальности через призму восприятия веб-дизайнера, заказчиков, потребителей, общих тенденций во всех сферах нашей жизни.