Маркетинговая аналитика

Раздел 4. Анализ данных?

4.7.2. Оценка средних значений. Оценка долей. Доверительные интервалы

Когда истинное значение параметров неизвестно, принято использовать оценки этих параметров. Оценка могут быть точечными или интервальными. Рассмотрим сначала точечные. Вообще говоря, существует бесконечно много способов оценить тот или иной параметр, поэтому важно, чтобы наша оценка обладала «хорошими» свойствами, такими как: состоятельность, несмещенность и эффективность. Поставим задачу следующим образом. Пусть неизвестна некоторая генеральная совокупность, например, зарплата каждого человека в Москве. Но известна некоторая выборка, взятая из генеральной совокупностей. Для примера, предположим, что выборка представляет собой 5 человек и их зарплаты: Петя: 50к, Вася: 60к, Сеня: 70к, Маша: 80к и Саша: 100к. Требуется оценить среднее и суммарное значение зарплат в Москве, если известно, что в Москве проживает ~12кк. Принято считать, что генеральная совокупность, как и выборка, порождена некоторым распределением с неизвестными параметрами, эти самые параметры нам и требуется оценить. Для оценки среднего выборки используется выборочное среднее (сумма всех элементов, поделенная на их количество). В нашем случае среднее выборки оценивается как: 50к+60к+70к+80к+100к/5=72к. Сумма же всех равна: 360к. Таким образом, мы получили оценку среднего и суммы для нашей выборки. Оценка среднего генеральной совокупности совпадает со средним выборочным, а вот для оценки суммы генеральной совокупности, необходимо выборочное среднее умножить на объем генеральной совокупности, в нашем случае 72к\*12кк = 864ккк. Таким образом, в месяц в Москве уходит 864 миллиарда на зарплаты людям :)

Часто у объектов бывают важные свойства и возникает задача оценить долю объектов с заданным свойством. Предположим, для примера, что в качестве генеральной совокупности рассматриваются дома в Москве: дома бывают многоэтажные и малоэтажные. Пусть нам неизвестно в точности, сколько домов в Москве, но дана выборка (например, определенный район), объемом 1000 и количеством многоэтажных домов равным 700. Для построения оценки доли многоэтажных домов в Москве, достаточно просто поделить количество многоэтажных домов в нашей выборке на объем выборки: 700/1000 = 0.7. Всё просто!

Однако, являются ли полученные оценки верными на 100%? Конечно, нет. Полученная доля многоэтажных домов во втором примере может сильно отличаться, если подсунуть в выборку другой район. А средняя зарплата Москвичи могла бы быть повыше, если бы в выборке присутствовали одни генеральные директоры…Для того, чтобы понять, насколько полученное значение близко к истине, пользуются стандартными ошибками полученных оценок, которые, по сути, показывают, насколько далеко наше значение от истины. Не будем останавливаться на том, как именно это используется, но интуиция подсказывает, что чем больше объем выборке, тем точнее оценка!

Главное полезное применение стандартной ошибки - это построение так называемых доверительных интервалов. Доверительный интервал – это метод интервальной оценки интересующего нас параметра. Зачем он вообще нужен? Поскольку точечные оценки не так точны, то знать в каком интервале на самом деле лежит истинное значение оцениваемого нами параметра крайне полезно. Для построения доверительного интервала необходимо зафиксировать уровень уверенности alpha. Эта альфа показывает, с какой вероятностью истинное значение параметра лежит в интервале. Чаще всего принято принимать альфу = 95%. Таким образом, для построения доверительного интервала необходимо зафиксировать/посчитать три величины: точечную оценку параметра, стандартную ошибку параметра и, с заданным уровнем альфа, посчитать квантиль (уровня 1-alpha) распределения (нормального/Стьюдента или любого другого). Значения квантилей всегда можно найти в таблицах или с помощью калькуляторов на просторе интернета. Когда все значения вычислены, можно записать интервал:

[точечная оценка – квантиль \* стандартная ошибка; точечная оценка + квантиль \* стандартная ошибка]

Доверительные интервалы применимы как для среднего значения с суммой, так и для оценки долей, разве что стандартная ошибка считается по-разному.

Несколько фактов о доверительных интервалах (интерпретация):

* Широкий доверительный интервал указывает на то, что оценка неточна, а узкий наоборот – на точность оценки.
* Ширина доверительного интервала зависит от размера стандартной ошибки, которая в свою очередь зависит от объема выборки
* Можно проверить, ложится ли вероятное значение для параметра популяции в пределы доверительного интервала. Если да, то результаты согласуются с этим вероятным значением. Если нет, тогда маловероятно (для 95% доверительного интервала шанс почти 5%), что параметр имеет это значение.

## ***Список источников***

1. Горяинов, Е. Р., Панков, А. Р., Платонов, Е.Н., Прикладные методы анализа статистических данных, 2012 – 310с
2. http://statistica.ru/theory/doveritelnye-intervaly/