Маркетинговая аналитика

Раздел 4. Анализ данных?

4.7.3. Сравнение средних значений в 2 и более группах. Дисперсионный анализ.

На практике часто возникает необходимость сравнить между собой несколько выборок. Что значит сравнить? Имеется в виду, понять, насколько они однородны, то есть похожи. Нас будет интересовать, действительно ли среднее значение выборок одинаковы или нет. Разделим случаи на два принципиально разных, когда имеет две выборки и когда выборок больше, чем две. Это важно, т.к. подходов для случая двух выборок больше и они проще, чем для случая множества выборок.

Итак, для пусть даны две выборке. Первое, что необходимо сделать, поставить задачу: сформулировать гипотезы. Основная гипотеза выглядит так: средние двух выборок (пусть выборки будут X и Y) равны. Альтернативная же гипотеза выглядит так: средние двух выборок не равны. Для проверки подобных гипотез существует множество различных критериев. Глобально они все делятся на параметрические и непараметрические. Параметрические критерии стоит использовать, когда известно распределения, из которых получены выборки (чаще всего необходимо нормальность выборок). Непараметрические же подходят для любых выборок (конечно же, если выборки независимы, иначе смысла в решении поставленной задачи нет). Иногда накладываются дополнительные ограничения на применимость, например, равенство дисперсий нулю, но мы не будем останавливать на них подробно, а посмотрим, как пользоваться критериями. У каждого критерия есть так называемая статистика – специальное число, используя которое можно принять решение о принятии или отвержении нулевой (основной) гипотезы. Статистика считается по определенной формуле. Например, в случае t-критерия Стьюдента, который является основным параметрическим критерием для проверки равенства средних двух выборок, используется такая формула:

t =

Где, X и Y с чертой – выборочные средние первой и второй выборки соответственно, s\_1 и s\_2 – стандартные отклонения, n\_1 и n\_2 – объемы выборок, t – сама статистика.

Все статистики имеют определенное распределение (чаще всего нормальное или Стьюдента, либо же собственное). В данном случает статистика имеет распределение Стьюдента с n\_1+n\_2-2 степенями свободы (нужно для того, чтобы знать, что смотреть в таблице). Необходимо также зафиксировать уровень значимости альфа – вероятность совершения ошибки первого рода (вероятность того, что основная гипотеза верна, но была отвергнута в пользу альтернативной). Обычно это довольно маленькое число, часто равное 0.05. По заданному уровню альфа, можно построить множество значение (критические области) для статистики, с помощью таблицы или калькулятора. После чего необходимо посмотреть, принадлежит ли посчитанная статистика критической области: если да, то нулевая гипотеза отвергается, если нет, то нулевая гипотеза принимается.

Альтернативный способ проверки заключается в подсчете p-value с помощью статистики и сравнением его с уровнем значимости. P-value – это вероятность получить такие или еще более экстремальные значения статистики. Если p-value близко к нулю (меньше уровня значимости альфа), то нулевая гипотеза отвергается, в ином случае – принимается.

Непараметрические критерии игнорируют распределения выборок и оперируют рангами элементов (по сути, им важны не значения элементов, а то на каком месте они стояли бы в отсортированном порядке). Такой подход является более универсальным, однако не всегда таким же точным. Как пример такого критерия – критерий Уилкоксона-Манна-Уитни.

Рассмотрим случай больше двух выборок. Может возникнуть вопрос, почему бы не использовать критерии для двух выборок и не сравнивать выборки попарно? Такой подход попарного сравнения в корне неверен и вызывает эффект множественных сравнений (подробно об этом можно почитать здесь *multiple comparisons;*Гланц 1999, с. 101-104).

Для множественного сравнения используют дисперсионный анализ. Дисперсионный анализ – статистический метод, предназначенный для оценки влияния различных факторов на результат эксперимента, а также для последующего планирования аналогичных экспериментов. Бывает однофакторным и многофакторным. Мы рассмотрим случай однофакторного дисперсионного анализа и поймем, как сравнить средние значения нескольких выборок.

Рассмотрим подробно, как всё посчитать. Предположим, дано три выборки. В каждой по десять наблюдений. Обозначим за наблюдение j в группе i, за – среднее значение в группе, а за – среднее значение всех элементов. Тогда каждое наблюдение можно разложить на следующее составляющие:

Второе слагаемое – отклонения групповых средних от общего среднего значения, а третье – отклонения отдельных наблюдений от среднего значения группы, к которой они принадлежат.

Тогда разброс внутри групп можно рассчитать, как:

А разброс между группами:

В приведенных обозначениях буква W означает “within” (внутри), а B – “between” (между). Теперь можно нормализовать величины и получить дисперсии внутри и между групп.

N – число наблюдений (в нашем случае 3 \* 10 = 30), k – число групп (в нашем случае 3). В знаменателях стоят количество степеней свободы, и такая нормализация позволяет сравнивать эти дисперсии с помощью F-критерия (Фишера).

F – статистика, имеющее распределение Фишера. Зная статистику, количество степеней свободы (соответственно и распределение), задав уровень значимость альфа, мы можем провести тест и определить, принимать нулевую гипотезу (о равенстве всех средних всех групп) или нет.

## ***Литература***

1. Горяинов, Е. Р., Панков, А. Р., Платонов, Е.Н., Прикладные методы анализа статистических данных, 2012 – 310с
2. <http://www.machinelearning.ru/wiki/index.php?title=%D0%9F%D1%80%D0%BE%D0%B2%D0%B5%D1%80%D0%BA%D0%B0_%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D1%85_%D0%B3%D0%B8%D0%BF%D0%BE%D1%82%D0%B5%D0%B7>
3. <https://r-analytics.blogspot.com/2013/01/blog-post.html>