Словарь начинающего аналитика

Полезен тем, что:

- Содержит расшифровку основных терминов, которые будут встречаться в курсе
- Всегда будет под рукой, и к нему можно обратиться в любой момент при обучении/работе

Для удобства термины расположены по порядку, исходя из последовательности лекций

Лекция 4

t-тест (критерий Стьюдента) — это статистический метод, который позволяет сравнивать средние значения двух выборок и на основе результатов теста делать заключение о том, отличаются они друг от друга статистически или нет.

Альтернативная гипотеза ($н_1$) — это единственное утверждение, являющееся логическим отрицанием нулевой гипотезы. Часто альтернативная гипотеза означает наличие связи между изучаемыми переменными.

Вероятность — это числовая характеристика возможности наступления какого-либо события. Например, вероятность того, что на кубике выпадет число 5, равна 1/6. Так же, как и для любого другого числа на кубике.

Выборка (выборочная совокупность) — часть генеральной совокупности элементов, которая охватывается экспериментом (наблюдением, опросом).

Генеральная совокупность — это собрание всех единиц или объектов, по отношению к которым планируется изучение определённой проблемы и принятие выводов. В состав генеральной совокупности входят все объекты, которые планируется изучать.

Дизайн эксперимента — это подробное техническое задание по тестированию выдвигаемой гипотезы. Необходим для того, чтобы понимать, что тестируем, на какой выборке и какие метрики считаем.

Дисперсия — это один из основных показателей в статистике. Он отражает меру разброса данных вокруг средней арифметической. Характеризует, насколько частные значения отклоняются от средней величины в данной выборке.
Чем больше дисперсия, тем больше отклонения или разброс данных.

Доверительный интервал — это определённый диапазон, который служит для оценки неизвестного параметра с высокой степенью надёжности. Например, нужно узнать, какому количеству людей известно о торговой марке фирмы. Когда будет подсчитан доверительный интервал, можно будет, например, сказать, что с 95% долей вероятности доля потребителей, знающих о данной торговой марке, находится в диапазоне от 27% до 34%.

Зависимость данных — это свойство данных, при котором изменение одних данных влечёт за собой изменение других данных. Итак, если взять очень много значений, мало зависящих друг от друга, и расположить на линии, получится характерный «колокол» Гауссового распределения. Опустим с самой верхней точки перпендикуляр — и вот она, нормаль. Слева и справа будет примерно одинаковое количество данных. В центре то, что встречается чаще всего.

Категориальная предикторная переменная — это переменная, принимающая одно из заданных значений (категорий), которые используются при предсказании откликов одной или более зависимых переменных. Например, «Пол. Категории: мужчины, женщины»; «Группа крови. Категории: I, II, III, IV».

Количественная переменная — это переменная, которая может принимать любые числовые значения в некотором диапазоне. Например, температура воздуха или систолическое давление.

Корреляционный анализ данных — это метод обработки статистических данных, заключающийся в изучении коэффициентов (корреляции) между переменными. При этом сравниваются коэффициенты корреляции между одной парой
или множеством пар признаков для установления между ними статистических взаимосвязей.

Корреляция — это взаимосвязь двух или нескольких случайных параметров. Когда одна величина растёт или уменьшается, другая тоже изменяется. Например, существует корреляция между температурой воздуха и потреблением мороженого. Чем жарче погода, тем больше мороженого покупают люди.
И наоборот.

Линейная регрессия — это модель зависимости переменной *x* от одной или нескольких других переменных (факторов, регрессоров, независимых переменных) с линейной функцией зависимости. Относится к задаче определения «линии наилучшего соответствия» через набор точек данных и применяется при анализе данных и в машинном обучении.

Медиана — это показатель (значение в совокупности), делящий ранжированные данные (отсортированные по возрастанию или убыванию) на две равные части. Значения в одной половине меньше, а в другой больше медианы. Например, в выборке {3, 5, 5, 9, 11} медианой является число 5.

Мода — это значение в анализируемой совокупности данных, которое встречается чаще других. Например, в выборке {6, 2, 6, 6, 8, 0} модой является число 6.

Научный метод — это комплекс способов получения новых знаний и методов решения задач в границах любой науки.

Непрерывные случайные величины — это случайная величина, которая может принимать все значения из некоторого конечного или бесконечного промежутка. Число возможных значений непрерывной случайной величины бесконечно. Например, расстояние, которое пролетит снаряд при выстреле — это непрерывная случайная величина, значения которой принадлежат некоторому промежутку [а; в].

Номинальная переменная — это переменная, значения которой представляют категории без естественного упорядочения: например, подразделение компании, где работает наёмный сотрудник. Примеры номинальных переменных включают регион, почтовый индекс или религию.

Нормальное распределение (распределение Гаусса, распределение Гаусса — Лапласа) — это набор значений, который создаст одинаковый рисунок (с допущениями, конечно) слева и справа от перпендикуляра, находящегося в месте наиболее часто встречающегося значения. Например, если взять обыкновенный песок, поместить в любую ёмкость и высыпать на землю, получится горка, которая и будет представлять нормальное распределение песчинок от центра к краям. Это происходит потому, что все песчинки имеют почти одинаковую (хотя и не совсем, если присмотреться) форму и массу, силы трения между ними практически одинаковы, и рассыпаются они под действием одной и той же силы.

Нулевая гипотеза ($н_0$) — это предположение о том, что никакой связи между изучаемыми событиями нет, и по умолчанию она считается верной, пока не будет доказано обратное.

Ошибка второго рода — это ситуация, когда при проверке статистических гипотез принята неправильная нулевая гипотеза.

Ошибка первого рода — это ситуация, когда при проверке статистических гипотез отвергнута правильная нулевая гипотеза.

Плотность вероятности — это один из способов задать распределение случайной величины. Отношение числа реализаций ожидаемого события, которые могут произойти на единице площади или объёма, к общему числу равновозможных реализаций.

Порядковая (ординальная, полуколичественная) переменная — это переменная, измеренная в шкале порядка. Позволяет ранжировать (упорядочить) объекты, если указано, какие из них в большей или меньшей степени обладают качеством, выраженным данной переменной. Однако они не позволяют определить,
«на сколько больше» или «на сколько меньше» данного качества содержится в переменной. Например, степень тяжести заболеваний и повреждений, шкала Рихтера для определения силы землетрясения, балльная система оценивания в школе.

Случайная величина — это математическое понятие, служащее для представления случайных явлений, когда для них может быть определена вероятность, то есть мера возможности наступления.

Стандартное отклонение — это мера разброса данных вокруг средней. Показывает, как распределены значения относительно среднего в выборке. В жизни функции стандартных отклонений используются для определения стабильности продаваемой продукции, создания цены, корректировки или формирования ассортимента и так далее.

Статистика — это наука, которая занимается вопросами сбора, измерения, мониторинга, анализа массовых статистических (количественных или качественных) данных и их сравнения.

Статистический анализ — это ряд математических приёмов обработки количественной информации, с помощью которых выявляются основные тенденции распределения показателей и степень корреляции между отдельными показателями.

Статистическая гипотеза о данных — это предположение о виде распределения
и свойствах случайной величины, которое можно подтвердить или опровергнуть применением статистических методов к данным выборки.

Статистическая значимость — это оценённая мера уверенности в «истинности» результата, репрезентативности выборки (выборка обладает всеми свойствами исходной популяции, значимыми с точки зрения задач исследования).

Статистический анализ — это заключительный этап исследования, представляющий собой процесс изучения, сопоставления, сравнения полученных данных (в т. ч. и с другими данными), их обобщения, истолкования
и формулирования научных и практических выводов.