Введение в статистику

Гришин Михаил

22 апреля 2016 г.

Содержание

Полезные сведения из предыдущих лекций

Математическое ожидание

Среднее ожидаемое значение случайной величины:

$$E[X] = \int_{-\infty}^{\infty} x \, dF(x).$$

q-квантиль

Такое значение случайной величины x_q, что вероятность события «X меньше или равен x» в точности равно q:

$$F(x_q) = q,$$

Для непрерывного распределения будет представлять обратную функцию распределения.

Полезные сведения из предыдущих лекций

Медиана

Медиана - 0.5 квантиль, то есть такое значение, которое делит область значений СВ на 2 части, попадание в которые является равновероятным.

Статистическая гипотеза

Статистическая гипотеза - некоторе предположение о виде и параметрах закона распределения, породившем наблюдаемую выборку. Проверка статистической гипотезы заключается в расчете некоторой выборочной статистики и сравнении ее с ожидаемыми результатами при условии верности проверяемой гипотезы.

Полезные сведения из предыдущих лекций

Параметрический метод

Параметрический метод в статистике - метод проверки гипотезы, который использует априорные предположения о законе распределения тестовой статистике. Эквивалентно: параметрические методы - методы, в основе которых лежит некоторая конечномерная модель.

Робастность

Под робастностью в статистике подразумевается устойчивость метода или оценки к наличию в данных выбросов - нетипичных значений, которые либо порождены иным процессом, либо являются следствием ошибки наблюдения.

Основные понятия

Непараметрическая статистика - совокупность методов проверки статистических гипотез и получения оценок, которые не используют предположения о том, что наблюдаемая выборка получена из какого-то заранее известного распределения.

Схожим образом, метод является непараметрическим, если он является бесконечномерным. Сложность непараметрической модели растет вместе с размером выборки.

Основной вопрос: когда следует применять параметрические методы, а когда лучше использовать непараметрические?

Выбор между параметрическим и непараметрическим методом

Аргументы «за» для параметрических методов:

- в силу **ЦПТ** статистики, построенные на выборках достаточно большого размера из генеральной совокупности, даже если она не является нормальной, будут асимптотически иметь нормальное распределение;
- параметрические тесты умеют «управляться» с выборками, имеющими различную дисперсию;
- мощность критерия - параметрические тесты, как правило, имеют более высокую мощность, таким образом парметрические тесты, как правило, более успешны в обнаружении значимых эффектов.

Выбор между параметрическим и непараметрическим методом

Аргументы «за» для непараметрических методов:

- исследуемую задачу лучше характеризует не среднее, а медиана. Как мы отмечали ранее - медиана является робастной статистикой, в отличие от среднего;
- маленький объем выборки, для которого ЦПТ может не выполняться;
- наличие выбросов в исходных данных - нетипичных значений, которые либо порождены иным процессом, либо являются следствием ошибки наблюдения.
- выборка представлена в виде ранговых наблюдений или наблюдения преставлены на ординальной шкале.

Выбор между параметрическим и непараметрическим методом

Аргументы «за» для непараметрических методов:

- исследуемую задачу лучше характеризует не среднее, а медиана. Как мы отмечали ранее - медиана является робастной статистикой, в отличие от среднего;
- маленький объем выборки, для которого ЦПТ может не выполняться;
- наличие выбросов в исходных данных - нетипичных значений, которые либо порождены иным процессом, либо являются следствием ошибки наблюдения.
- выборка представлена в виде ранговых наблюдений или преставлены на ординальной шкале.

Выбор между параметрическим и непараметрическим методом

Неформально, процедуру определения «пригодности» выборки для параметрического теста (с определенными оговорками) можно описать так:

- в качестве быстрого теста можно посмотреть на график распределения - визуально оценить скошенность распределения и наличие выбросов;
- рассчитать базовые описательные статистики и сравнить их значения между собой;
- рассчитать значения критерия согласия с нулевой гипотезой о нормальности распределения.

Выбор между параметрическим и непараметрическим методом

Для примера быстрого анализа исходной выборки на принадлежность к нормальному распределению рассмотрим ноутбук median.ipynb.

Примеры критериев согласия, которые могут быть использованы, если исходное предположение о нормальности не подтвердилось, приведены в ноутбуке gof.ipynb.

Непараметрические тесты

В основном, непараметрические тесты используются для проверки гипотезы о равенстве распределения между генеральными совокупностями.

Примеры приведены в ноутбуке nonparametrictests.ipynb.

Также интерес представляют непараметрические меры связи между переменными. Меры связи и способы их оценки приведены в ноутбуке nonparametricmeasures.ipynb.

Бутстреп

Бутстреп - класс статистических методов для получения выборочного распределения оценок в случае, если точную оценку получить проблематично без введения дополнительных ограничений.

Метод основан на приближении теоретического распределения выборками с повторением. Описание метода и примеры использования приведены в ноутбуке bootstrap.ipynb.

Непараметрические оценки плотности

Для непараметрической оценки плотности распределения случайной величины существует класс т.н. «ядерных» методов, которые используют ядерную функцию для того, чтобы получить гладкую оценку функции распределения без параметризации распределения в генеральной совокупности.

Непараметрические оценки плотности рассмотрены в ноутбуке KDE.ipynb.

Спасибо за внимание!