Статистика

Основные понятия

- Выборка
  - Случайный элемент X — это функция, которая случайным исходам сопоставляет не обязательно числа (множества, векторы, функции). Таким образом, это обобщение понятия случайной величины
  - Набор x = [x_l, ..., x_n] — выборка, если x содержит n независимых реализаций случайного элемента X
  - Случайная выборка — это такой же вектор, только составленный из случайных величин
- Модель
  - Опыт, знания -> структура модели -> модель -> оценка -> искомая величина
    - Пример: хотим измерить время падения камня с пизанской башни. Известно, что камень падает вниз под действием силы тяжести.
      Знаем, что можем ввести систему координат и найти уравнение движения камня. Найти ускорение, скорость.
      Можем теперь построить модель (какую-то). Например, метод М-К, и найти оценку на распределение искомой величины.
  - Как-то описывает механизм генерации данных. Нет жёсткого определения " генерации".
  - "Правильных" моделей нет, все модели " плохие", но могут хорошо соответствовать нашему условию/ эксперименту.
- Эмпирическое распределение
  - Это дискретное распределение, построенное по частотам элементов в выборке.
  - $x_n = [x_1, \dots, x_n],\ \mathcal{P}_n^* : \mathbb{P}(x_i) = \frac{1}{l_i}; l_i = \#\{x_i\}$
- Теорема Гливенко-Кантелли
  - Супремум модуля разности эмпирической функции распределения и обычной сходится к нулю почти наверное на бесконечности.
- Теорема Колмогорова (про непрерывную функцию распределения)
  - Если функция распределения непрерывна, то супремум модуля разности обычной и эмпирических ф.р. по распределению сходится к распределению Колмогорова.
- Plug-in оценка
  - Нужно оценить какую-то характеристику phi(P(x)_n). Точного распределения мы [обычно] не знаем, но можем оценить его как-то (выборкой, например).
    Теперь меняем P(x)_n на P(x)_n^* и получаем оценку по выборке.
- Смещение оценки
  - $b(\varphi^*) = \mathbb{E}\varphi^*(X_{[n]}) - \varphi(\mathcal{P})$
- Разброс
  - $\mathbb{D}\varphi^*(X_{[n]})$

Интервальные оценки

- Доверительные интервалы [Confidence intervals]
  - Это пара статистик (phi*_l, phi*_r) с уровнем доверия gamma таких, что
    $\mathbb{P}(\phi_L^*(X_{[n]}) \le \phi(\mathcal{P}_X) \le \phi_R^*(X_{[n]})) = \gamma$
    - Иногда не ас. называют точным
    - Если сходимость только на бесконечности, то интервал асимптотический
    - Левый
      $\phi_L^*(X_{[n]}) = -\infty$
    - Правый
      $\phi_R^*(X_{[n]}) = \infty$
  - Вероятность тут стоит понимать в частотном смысле -- никаких мер на множествах тут нет, но, если мы сгенерируем N выборок, то примерно gamma*N будут накрывать истинное значение параметра
  - Эфронов (или просто "доверительный интервал")
    - Построение:
      считаем бутстраповскую выборку статистик и берём квантили
    - **Используют только это на практике**
  - BCa-интервал, навороченный. (От "Bias Correction & acceleration")
    - Построение: используя метод складного ножа для phi.
- Достоверные интервалы (байесовский случай). [Credible intervals]

Свойства оценок ('хорошие')

- Несмещённость.
  Матожидание оценки равно параметру
  - Прямая
  - Асимптотическая
  - Слабое свойство.
    Можно аггрегировать оценки от разных экспериментов, если они не смещены.
- Состоятельность.
  Оценка сходится к параметру по вероятности
  - Прямая
    - Проверка состоятельности:
      пусть мы хотим проверить не оригинальную оценку phi, а функцию от неё g(phi). Тогда, если g непрерывна, то по теореме Манна-Уайльда(continuous mapping theorem) из какой-либо сходимости phi*(x_n) -> phi(\mathcal{P}) будет следовать аналогичная сходимость g(phi*(x_n)) -> g(phi(\mathcal{P}))
    - Асимптотическая. На бесконечности только сойдётся к параметру. Довольно слабое и не говорит о скорости сходимости оценки параметра к истинному значению
  - Сильное, показывает скорость сходимости
    - Асимптотическая нормальность.
      Разность оценки и параметра сходится к нормальному распределению *как-то* при стремлении размера выборки к бесконечности
      $\sqrt{n}(\varphi^*(X_{[n]}) - \varphi) \to \xi \sim \mathcal{N}(0, \sigma_{\varphi^*}^2)$
      - Проверка ас. нормальности через дельта-метод. Верно ли, что из n**0.5(phi* - phi) -> eta \sim D (распределение) следует n**0.5(g(phi*) - g(phi)) -> <smth>?
        Условие: дифференцируемость g(phi).
        Раскладываем функцию в Тейлора с первой производной, берём небольшое приращение и затем приравниваем приращение (phi*-phi). Получаем, что по распределению n**0.5(g(phi*) - g(phi)) -> g'(phi) eta.
- Эффективность.
  Оценка имеет наименьшую дисперсию в каком-то классе.
  - Прямая
  - Асимптотическая
- Робастность

Метод МК. Выборочные характеристики

- plug-in оценки
  - Выборочное среднее
  - Выборочная (исправленная) дисперсия/ стандартное отклонение
  - k-выборочный момент/k-центральный выборочный момент
- Метод М-К
  - Есть выборка из какого-то неизвестного распределения. Мы хотим для неё найти какую-то характеристику. Поскольку истинного распределения мы не знаем, то можем всегда заменить его эмпирическим.
    Идея оценок методом подстановки в этом и заключается: мы берём характеристику и считаём её для эмпирического распределения и говорим, что это оценка истинной хар-ки
  - Метод МК сам по себе является обобщением плагин-оценок
- Вариационный ряд
  - Построение
    - Отсортированная выборка
  - Статистический ряд
    - Это вариационный, если есть повторяющиеся значения. Для каждого уникального считаем кратность
- Положения
  - Квартиль — то же самое, только по четвертям (от лат. quartus — четверть). Обозначаются Q_i
  - a-квантиль — это значение
    $x_a : P(x \le x_a) = a$
  - Медиана — 0.5-квантиль
- Разброса
  - Интерквартильный размах
    $IQR = Q_3 - Q1$
  - Размах (амплитуда)
    - Разница между мин. и макс. значениями
  - Выборочное и исправленное выборочное стандартное отклонение
    - Стандартное отклонение, посчитанное за незнанием истинного распределения
    - Если дисперсия не определена (например, Коши), то нет смысла использовать
  - Медианное стандартное отклонение
    - Считается так же, как и обычное, только отклонение не от выборочного среднего, а выборочной медианы
    - Существует всегда

Параметрические модели. Оценка параметров

М-оценки. Робастность

- Обобщение метода максимального правдоподобия

Бутстрап

- plug-in-оценка разброса характеристики.
  Из-за неспособности постоянно семплировать из истинного распределения, мы семплируем из эмпирического и находим оценки искомой характеристики. Далее можем посчитать смещение/разброс
- Не может определить смещение, обусловленное плохой моделью
- Ошибка
  - Устранимая
    - N выборок, а не беск.
  - Неустранимая
    - Из-за эмпирического распр. вместо истинного

Проверка гипотез

- Простая и сложная гипотезы
- Гипотезы о характеристиках
- Гипотезы согласия
- Гипотезы однородности
  - A/B тест
  - A/A тест

Дисперсионный анализ. ANOVA (ANalysis Of VAriance)

- Множественная проверка гипотез
  - Проблема в том, что, если проверять не одновременное множественное равенство характеристик с.в., а равенство по отдельности, то начинает накапливаться ошибка первого рода.

Линейная регрессия.

$x \in \mathbb{R}^{n \times k}, y \in \mathbb{R}^n$

- Гипотезы
  - Оценка у среднего
  - Оценка y_n+1
  - Оценка значимости всех коэффициентов кроме константного
- Теорема Гаусса-Маркова
  - Четыре условия:
    - forall i E(eps_i) = 0
    - Гомоскедастичность (константность дисперсии для всех ошибок: forall i var(eps_i) = sigma^2
    - cov(eps_i, cov_j) = 0 для i != j
    - rank(x) = k
      Тогда оценка beta^* является наилучшей в классе линейных несмещённых оценок
- Нормальные ошибки

Корреляционный анализ

- Таблицы сопряженности.
  - Разбиваем всевозможные значения на ячейки и количество попадания значений в каждую ячейку.
  - Пример: даны два вектора оценок судей (светские улыбки), состоящий из 1 и 0. Хотим узнать, скоррелированы (независимы ли) оценки судей между собой. Строим таблицу из количества всевозможных пар.
- Подтема 2

Байесовские методы