

Семинары по статистике 2021

1 Введение: принципы работы с данными в статистике

Пусть имеется *выборка из генеральной совокупности размера n* — такая формулировка уже означает **приятие специальных условий статистической модели**, а именно:

Наличие некоторой случайной величины (вариант – случайного вектора) $\xi : \Omega \rightarrow \mathbb{R}$, функция распределения F_ξ которой нам не полностью известна, а характеризуется лишь принадлежностью к некоторому классу \mathfrak{F} распределений,

Наличие модели повторений (то есть последовательности ξ_1, ξ_2, \dots независимых и распределенных так же как ξ), задающих меру на пространстве \mathbb{R}^∞ .

Наличие конечной последовательности чисел $\xi_1(\omega_1), \xi_2(\omega_2), \dots, \xi_n(\omega_n)$, которая, собственно, и называется *выборкой*.

1.1 Типичные задачи математической статистики

Типичные задачи математической статистики таковы: по заданной выборке

Определить функцию распределения F_ξ

Определить плотность распределения f_ξ (в предположении, что таковая существует).

Определить числовой функционал $T(\xi)$, например, определить математическое ожидание $E(\xi)$ (или разобраться, существует ли таковой)

Непараметрическая регрессия По выборке пар $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$ значений случайного вектора (ξ, η) определить функцию регрессии $r(x) = E(\eta | \xi = x)$.

Линейная регрессия в корреляционной теории. Пусть $\xi = (\alpha, \beta)$ двумерный гауссовский вектор, причем $\beta = c\alpha + d + \epsilon$, где c, d вещественные числа, а ϵ – величина с нулевым математическим ожиданием. Оценить по выборке из вектора $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$ значения c, d .

Основная задача корреляционной теории. По выборке, отвечающей многомерному гауссовскому распределению, определить набор его параметров.

Напомним, что за словом «определить» подразумевается дать статистический способ описания с указанием вероятностей попадания искомого объекта в заданный класс значений.

1.2 Неформально о практических навыках

Методы математической статистики требуют достаточно длинных, порою громоздких вычислений. Разумеется, что человечество использует для этих целей компьютеры и возникает естественный вопрос о наилучших программных пакетах для таких вычислений (их немало, но они быстро устаревают в основном из-за развития операционных систем и графики). Достаточно долгое время ведущие позиции для работы с данными занимал MATLAB, однако коммерциализация разработчиков в настоящее время не позволяет его использовать широко. Для учебных целей вполне годится свободно распространяемый пакет OCTAVE, который с точки зрения пользователя есть тот же MATLAB (различия видны только при работе с очень большими массивами данных). В настоящий момент исследователи BigData в основном используют PYTHON и R, причем эти два языка программирования для задач статистики успешно конкурируют (о чем имеется специальное рассуждение <https://opensource.com/article/16/11/python-vs-r-machine-learning-data-analysis>). С точки зрения обучения методам совершенно неважно, какой именно язык выбрать, главное чтобы была возможность оперировать с данными не отвлекаясь на технические детали. Основные приемы вычислений с данными давно алгоритмизированы и на практике требуется лишь понимание того, какую функцию из библиотеки процедур надо применить. Однако серьезная трудность сосредоточена именно в понимании: всегда есть несколько возможностей действия и неточное знание границ применимости метода легко приведет к бессмысленному ответу. Например, формула для оценки математического ожидания $E(\xi)$ чрезвычайно проста (взять среднее значение по выборке), но мы-то знаем, что математическое ожидание в каком-либо классе \mathfrak{F} распределений вполне может и не существовать — а как тогда быть с полученным ответом? Основной принцип заключается в том, что вдобавок к библиотечной функции (реализующей статистическую формулу из справочника) нужно еще знать несколько приемов для контроля своих действий, обсуждение смысла применения формул статистики и контрольных действий составляет значительную часть нашего курса. Вот два важнейших навыка при практической работе методами математической статистики.

Симуляция данных

Подразумевается, что данные для задачи возникают извне: это заказ общества специалисту по обработке. Предположения о том как устроены данные не всегда оговаривают специфику, например в измерениях может сохраняться значительная ошибка, природа которой заказчику неизвестна и он о ней может ничего и не знать. Исследование того, как повлияет конкретная ошибка на ответ остается эксперту, причем часто ответ сложно дать в аналитических терминах. Здесь часто применяют метод, известный под названием boot-strap, заключающийся в генерации синтетических данных (например, включающих ошибку конкретного вида) и повторений статистических вычислений применительно к синтетическим данным. Необходимый навык здесь — умение построить при помощи компьютера выборку значений случайной величины, иными словами, умение создать выборку по известному распределению $F_\xi(x)$. К этому умению, в свою очередь, прилагается умение проконтролировать свои действия, убедиться в том, что синтетические данные действительно распределены как надо и действительно независимы. Для некоторых типов распределений генераторы соответствующих выборок уже реализованы в библиотеке процедур, но безусловно надо понимать, как действовать и в нестандартном случае.

Формулы преобразования законов распределения

Итак, для выборки $x_{i_1}, x_{i_2}, \dots, x_{i_n}$ статистическая формула предложит вычисление значения $Z(x_{i_1}, x_{i_2}, \dots, x_{i_n})$, ясно, что для другой выборки из той же генеральной совокупности вычисленное значение может отличаться. Таким образом, $Z(x_{i_1}, x_{i_2}, \dots, x_{i_n})$ задает случайную величину, потому что аргументы берутся из пространства \mathbb{R}^n с вероятностной мерой: $(x_{i_1}, x_{i_2}, \dots, x_{i_n}) \in \mathbb{R}^n$.

Таким образом мы сталкиваемся с исследованием случайных величин, построенных по случайному вектору (который, в свою очередь, построен согласно схеме повторных независимых повторений. Чтобы пройти по этому пути необходимо уметь каждый раз получать формулу для соответствующих распределений.

1.3 Гипотезы о распределениях. Предварительный графический анализ

Речь идет об умении с помощью наглядно демонстрируемых особенностей интерпретировать формальные свойства данных. Этот этап при работе с данными формально говоря, не входит в математическую статистику, однако на практике он очень важен, поскольку проверяет (на эвристическом уровне) предположение о том, что изучаемые данные возникли в процессе независимых повторений эксперимента со случайным исходом.

Традиционный, всем знакомый подход, состоит в использовании гистограмм. Здесь есть над чем подумать: гистограммы строятся, исходя из некоторого разумного числа бинов на оси, причем разумным числом бинов традиционно называют то, которое подходит для одномодальных распределений, типа гауссова распределения. В частности, широко распространены следующие рекомендации для выбора числа бинов k у выборки размера n

Правило Большого Пальца $k = \left\lceil \frac{\sqrt{n}}{3} \right\rceil$ (фольклор)

Правило Стерджеса $k = 1 + \lceil \log_2 n \rceil$ (H.Sturges, 1926)

Другие рекомендации построения гистограмм состоят в выборе оптимального размера h интервала бинирования, тогда число бинов возникнет как результат деления разброса всей выборки на величину бина.

Правило Скотта $h = \left\lceil s^* \cdot \frac{3.49}{\sqrt[3]{n}} \right\rceil$ (D.Scott, 1979)

Правило Фридмана $h = \left\lceil \left(q_{0.75}^* - q_{0.25}^* \right) \frac{2}{\sqrt[3]{n}} \right\rceil$ (Freedman and Diaconis, 1981)

здесь квадратные скобки обозначают целую часть, а s^* и q_t^* — вычисленные по выборке соответственно эмпирическую дисперсию¹ и t -квантили. В любом случае видно, что количество бинов в типовой гистограмме совсем немного. Вдобавок, формула Скотта предлагает ориентироваться на формулу величины эмпирической дисперсии, применять которую к неизвестным данным опасно.

¹О типичных статистических характеристиках речь пойдет далее. Пока же примем, что эмпирическую дисперсию можно определить через среднее \bar{x} выборки формулой

$$S^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 = \frac{1}{n} \sum_{k=1}^n x_k^2 - (\bar{x})^2$$

Пример: иногда статистические формулы выдают странные ответы

Выберем достаточно реалистичный объем данных, например, $N = 1000$. Сгенерируем N случайных (равномерных) чисел α_i на интервале $[0, 1]$ и рассмотрим данные $\omega_i = 1/\alpha_i$. Нет никаких проблем с тем, чтобы вычислить арифметическое среднее всех α_i и всех ω_i . Однако, если вы проделаете этот эксперимент несколько раз, то получающиеся ответы вас удивят. Объясните, в чем здесь дело.

1.3.1 Задача: указать какие формулы могут соответствовать данным

Все данные в задаче получены компьютерной симуляцией, записаны в текстовом формате по колонкам, файл `Zoopark.txt` выложен на Яндекс-диск <https://disk.yandex.ru/d/snty0o9oGSMP7A>. Список возможных плотностей к задаче следующий: $\gamma_1 + \gamma_2 + \gamma_3$, $|\eta|$, $\sqrt{\gamma}$, e^η , $|\frac{\gamma_1}{\gamma_2}|$. Здесь обозначения γ указывает на показательное распределение, γ_i независимы, и η указывает на гауссово распределение с математическим ожиданием 0.

1. Для выборок из разных генеральных совокупностей и набора соответствующих им плотностей указать взаимные соответствия.
2. Для каждой выборки указать (эвристические) аргументы в пользу существования или несуществования первых двух моментов у исследуемой случайной величины.

Разумеется, такая задача — учебная: для симуляции данных использовались точные формулы, задающие закон распределения. В реальной жизни в данных всегда присутствуют сторонние эффекты, насколько они существенны покажет дальнейший количественный статистический анализ. На деле искусство обработчика заключается в умении предвидеть модель (то есть схему повторных независимых повторений) на основе понимания происхождения данных, графический анализ важен, но им не ограничиваются. На предварительном для применения статистической теории этапе обычно есть несколько возможностей и с помощью количественных методов статистики надо еще их сравнивать как альтернативные гипотезы.

Некоторые указания к задаче

Одним из различающих свойств распределений является асимптотика убывания их плотности распределения на бесконечности. Оценить асимптотику, используя только гистограммы, совсем даже непросто: действительно, гистограммы строятся, исходя из малого числа бинов на оси, и для оценки асимптотики убывания на бесконечности этих бинов в типовой гистограмме будет совсем немного.

Лучше обстоит дело, если обратиться к эмпирической функции распределения F_n^* , построенный по выборке график которой содержит n характерных точек-ступенек. Связь с асимптотикой убывания плотности функции распределения $f_\xi(x)$ ясна из решения следующей вспомогательной задачи:

- Как убывает на бесконечности функция $1 - F_\xi(x)$ ($F_\xi(x)$ — функция распределения), если соответствующая плотность распределения $f_\xi(x)$ убывает как $y = \frac{1}{x^m}$?

Для визуального анализа асимптотик убывания ключевое соображение связано с отрисовкой графиков в логарифмической шкале по одной или обоим осям. Вот серия вспомогательных упражнений для понимания этого практического приема.

- Как выглядит график зависимости $y = \frac{1}{x^m}$ в билигарифмической шкале?. Указать коэффициент наклона.
- На Рис(1) три кривые характеризующие разные типы убывания — полиномиального (синий), экспоненциального (красный) и сверхэкспоненциального (черный) изображены в линейной, полулогарифмической и билигарифмической шкалах. Объяснить какая картинка соответствует какому типу изображению.

1.4 Как самому создать статистический тест

Вот как можно создать тест для проверки гипотезы о том, что *выборка* y_1, y_2, \dots, y_n *размера* n *взята из генеральной совокупности распределений с плотностью*² $f_\xi(x)$.

Схема применения теста:

²увы, но действия, реализующие этот самодельный тест, зависят от распределения. То есть это неудобный тест, по сравнению, скажем, с тестом Колмогорова-Смирнова.

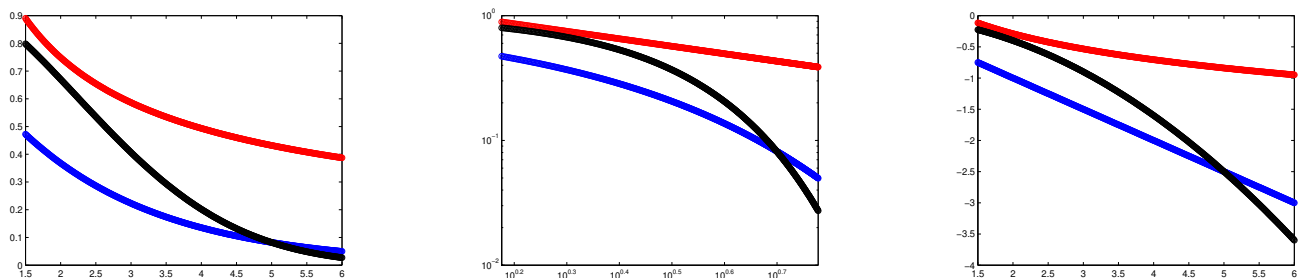


Рис. 1: Вид графиков полиномиального (синий), экспоненциального (красный) и сверхэкспоненциального (черный) убывания в разных шкалах: линейной, полулогарифмической и билוגарифмической шкалах.

1. На отрезке $[\min(y_k), \max(y_k)]$ строим нормализованную (то есть из частот, а не из числа попаданий в сегмент) гистограмму из m столбиков.
2. По некоторому (см. ниже) правилу подсчитывается расстояние $\widehat{D}_{[n,m]}$ между частотами из нормализованной гистограммы и теоретическими вероятностями.
3. Для малого ϵ критерий на соответствующем уровне значимости $1 - \epsilon$ применяется так: если расстояние $\widehat{D}_{[n,m]} > q_{1-\epsilon}$, то гипотезу отвергают (измеренное расстояние слишком велико, а это маловероятно, если выборка и вправду взята из распределения с.в. $f_\xi(x)$), здесь конечно $q_{1-\epsilon}$ — это квантиль распределения статистики $D_{[n,m]}$

Таким образом, для заданных $f_\xi(x)$, n , m осталось только составить таблицу (приближенных) значений квантилей $q_{1-\epsilon}$ распределения статистики $D_{[n,m]}$. Вы должны рассчитать значения квантилей для $\epsilon = 0.1, 0.05$ и указать (приблизительно) форму функции плотности распределения статистики $D_{[n,m]}$.

Но откуда же взять значения, предположительно зависящих от n , m и $f_\xi(x)$ квантилей? Этот этап был бы самым трудоемким при аналитическом выводе, но в наше компьютерную эпоху можно поступить иначе.

1.4.1 Упражнение

Догадайтесь, как в указанной ситуации получить (приблизительно) значения квантиля, например, для $t = 0.95$?

1.4.2 Задание с вариантами теста

Варианты индексируются значениями опций А (тип распределения), n (размер выборки), С (способ построения гистограммы), D (способ вычисления $\widehat{D}_{[n,m]}$)

Варианты плотностей $f_\xi(x)$ гипотетического распределения

1. А=0. Равномерное распределение на $[0, 1]$
2. А=1. Квадратный корень из равномерного распределения на $[0, 1]$

Варианты значений n

1. $n = 1000$
2. $n = 100$

Варианты значений С

1. С=0. Число столбиков m выбирается по правилу Большого Пальца
2. С=1. Число столбиков m выбирается по правилу Стерджеса

Варианты правил D для вычисления $\widehat{D}_{[n,m]}$

В следующих формулах w_i – частота попадания в i -й сегмент бинирования, то есть в $(x_{i-1}, x_i]$, $i = 1, \dots, m$, а соответственно $p_i = P(\xi \in (x_{i-1}, x_i])$.

1. D=1. $D_{[n,m]} = \max_i |w_i - p_i|$

2. D=2. $D_{[n,m]} = \sum_i |w_i - p_i|$

3. D=3. $D_{[n,m]} = \sqrt{\sum_i (w_i - p_i)^2}$

4. D=4. $D_{[n,m]} = \sum_i \frac{|w_i - p_i|}{p_i}$

5. D=5. $D_{[n,m]} = \sum_i \frac{(w_i - p_i)^2}{p_i(1-p_i)}$

1.4.3 Задача

Как (и надо ли) изменить критерий, если проверяется гипотеза о том, что заданная выборка взята из (соответственно см. вариант значения A)

1. A=0. Равномерного распределения на $[0, B]$ $B \neq 1$?

2. A=1. Квадратного корня из равномерного распределения на $[0, B]$ $B \neq 1$?

2 Практические методы непараметрической статистики

2.1 Описание теста

Тест Колмогорова-Смирнова (KS) применим к распределениям непрерывной случайной величины ξ (но не вектора!), то есть можно считать, что выборка состоит из не повторяющихся числовых величин. По n -выборке строится эмпирическая функция распределения $F_n^*(x)$: для действительного $y \in \mathbb{R}$ положим $F_n^*(y)$ равным $1/n$ умноженному на число тех x_i в наборе x_1, x_2, \dots, x_n , которые меньше либо равны y . Для сравнения полученной ступенчатой функции с функцией распределения $F_\xi(x)$ используется мера D_n :

$$D_n = \sup_x |F_n^*(x) - F_\xi(x)|$$

(ее версии D_n^+ и D_n^- использовались Смирновым для мер в областях $\{x | F_n^*(x) > F_\xi(x)\}$ и $\{x | F_n^*(x) < F_\xi(x)\}$)

$$\lim_{n \rightarrow \infty} P(\sqrt{n}D_n > z) = K(z)$$
$$K(z) = \begin{cases} 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 z^2}, & z > 0 \\ 0 & z \leq 0, \end{cases}$$

Построив (приблизительно) график $K(z)$ легко убедиться, что это монотонная функция с $K(0) = 1$ и $K(+\infty) = 0$. Сходимость по n к пределу очень быстрая (следует из доказательства теоремы Колмогорова, которое мы не разбирали), тем самым нам почти известны функции распределения случайных величин D_n . В терминах функции $K(z)$ можно сформулировать правило, когда на данном уровне значимости надо отвергнуть гипотезу о соответствии распределения и выборки, то есть утверждать о маловероятности наблюдаемого в эксперименте значения D_{observed} :

$$P(D_n > D_{\text{observed}}) \approx K(D_{\text{observed}} \cdot (\sqrt{n} + 0.12 + 0.11/\sqrt{n}))$$

Странное выражение $(\sqrt{n} + 0.12 + 0.11/\sqrt{n})$ отражает поправки насколько точно при конкретном n вероятность $P(D_n \sqrt{n} > z)$ описывается предельной формулой $K(z)$.

Чаще всего критерий Колмогорова-Смирнова используется для следующей задачи: *опровергается ли на данном уровне значимости, что две выборки длиной в n_1 и n_2 отсчетов взяты из одной генеральной совокупности (а про гипотетическое распределение не говорится ничего, кроме его непрерывности)?* В ее (почти очевидном) решении также задействовано расстояние D_{n_1, n_2} между двумя ступенчатыми графиками и далее работает практически та же самая формула, но в учитывающем разницу длин выборок виде:

$$P(D_{n_1, n_2} > D_{\text{observed}}) \approx K(D_{\text{observed}} \cdot (\sqrt{M} + 0.12 + 0.11/\sqrt{M})) \quad M = \frac{n_1 n_2}{n_1 + n_2}$$

В конкретных компьютерных реализациях обычно не приходится вычислять расстояние самому — соответствующая библиотечная функция проделает все за вас.

Из-за того, что метод основан на максимальном отклонении двух функций распределения, ясно, что максимальное отклонение наблюдается скорее в центральной части распределения (ближе к медиане неизвестного распределения), а не на концах, где графики выполаживаются. Результатом является то, что тест Колмогорова-Смирнова эффективно разбирается с отличиями типа сдвига распределения, но хуже чувствует разницу на хвостах распределений. Важная модификация теста Колмогорова-Смирнова называется тестом Андерсона-Дарлинга и основана на вычислении по выборке вместо $\sup_x |F_n^*(x) - F_\xi(x)|$ величины

$$\tilde{D}_n = \sup_x \frac{|F_n^*(x) - F_\xi(x)|}{\sqrt{F_\xi(x)(1 - F_\xi(x))}}$$

или (в зависимости от версии)

$$\tilde{\tilde{D}}_n = \int_{-\infty}^{+\infty} \frac{|F_n^*(x) - F_\xi(x)|}{\sqrt{F_\xi(x)(1 - F_\xi(x))}} dx$$

Для этих мер также рассчитаны распределения и тест Андерсона-Дарлинга по факту оказался более чувствительным к отклонениям на хвостах распределения, чем в середине. Бессмысленно ставить вопрос какой тест лучше потому что наиболее правильным будет применить несколько тестов — не отвергнет ли гипотезу хоть один из них. Впрочем, здесь появляется трудный вопрос об уровне значимости применения нескольких тестов сразу и о приближениях в используемых расчетных формулах. В нашем вводном курсе мы это не рассматриваем, но в принципе это вполне математически корректный вопрос.

2.2 Упражнения на критерий Колмогорова-Смирнова

2.2.1 Еще раз о распознавании гипотез

Как уже сказано выше стандартная задача непараметрической статистики возникает, когда даны две выборки (возможно разного объема) из, вообще говоря, разных генеральных совокупностей и надо на данном уровне значимости сделать статистические выводы совпадают ли эти две генеральные совокупности. длины выборки и формы распределения В частности вернемся к задаче 1.3.1³.

1. Применить компьютерные симуляции и на уровне значимости 0.95 методом Колмогорова-Смирнова возникшие у вас из предварительного рассмотрения пять гипотез о распределениях в задаче 1.3.1. Разумеется, здесь встает вопрос о параметрах симуляции, Общий подход будет рассмотрен позже, пока приведем несколько значений, которые надо использовать для подстановки в симуляции (куда какой разберитесь, пожалуйста, самостоятельно): 1, 0.3 1.5,
2. Компьютерные эксперименты, показывающие роль параметров распределения и приближений предельного перехода в критерии KS. Заодно это дает опыт практического применения критерия KS (поскольку теория нам в принципе известна).
 - (a) Подготовьте 1000 выборок длины 100 значений случайной величины, распределенной как $\mathcal{N}(0, 1)$ и сосчитайте 95% границу отклонения эмпирических функций распределения от теоретической функции распределения. Вычислите и нарисуйте график границ, где по критерию К-С должны лежать 95% эмпирических функций распределения выборки длины 100 значений значений случайной величины, распределенной как $\mathcal{N}(0, 1)$. Сравните величину этих границ с тем, что предлагает предельная функция Колмогорова (то есть при $n \rightarrow \infty$).
 - (b) Подготовьте 1000 выборок длины 25 значений случайной величины, распределенной как $\mathcal{N}(0, 1)$ и сосчитайте 95% границу отклонения эмпирических функций распределения от теоретической функции распределения. Вычислите и нарисуйте график границ, где по критерию К-С должны лежать 95% эмпирических функций распределения выборки длины 25 значений значений случайной величины, распределенной как $\mathcal{N}(0, 1)$. Сравните величину этих границ с тем, что предлагает предельная функция Колмогорова (то есть при $n \rightarrow \infty$).
 - (c) Подготовьте 1000 выборок длины 100 значений случайной величины, распределенной по закону Коши с параметром 1 и сосчитайте 95% границу отклонения эмпирических функций распределения от теоретической функции распределения. Вычислите и нарисуйте график границ, где по критерию К-С должны лежать 95% эмпирических функций распределения выборки длины 100 значений значений случайной величины, распределенной по закону Коши с параметром 1. Сравните величину этих границ с тем, что предлагает предельная функция Колмогорова (то есть при $n \rightarrow \infty$).
 - (d) Подготовьте 1000 выборок длины 50 значений случайной величины, распределенной по закону Коши с параметром 1 и сосчитайте 95% границу отклонения эмпирических функций а что обратить внимание распределения от теоретической функции распределения. Вычислите и нарисуйте график границ, где по критерию К-С должны лежать 95% эмпирических функций распределения выборки длины 50 значений значений случайной величины, распределенной по закону Коши с параметром 1. Сравните величину этих границ с тем, что предлагает предельная функция Колмогорова (то есть при $n \rightarrow \infty$).

2.2.2 Замечание

В компьютерной реализации решения имеется не вполне очевидный шаг: вычисление точной верхней границы расстояний между построенной по выборке ступенчатой функцией и построенной по явному уравнению непрерывной функцией $F_{\xi}(x)$ требует, вообще говоря, сравнения значений в бесконечном наборе аргументов. Но непрерывная функция $F_{\xi}(x)$ нигде не убывает, поэтому на каждой ступеньке максимальное расстояние между графиками заведомо достигается либо на правом конце ступеньки, либо равно пределу расстояний при аргументах стремящихся к левому концу ступеньки. Но необходимо ли нам вычислять пределы при учете расстояний? Обратите внимание, что ступеньки каждый раз поднимаются на высоту $1/n$ поскольку мы в этой задаче занимаемся выборками в классе непрерывных распределений и получить два совершенно одинаковых значения в такой выборке можно лишь с вероятностью ноль. Отсюда следует, что сверку расстояний надо вести все-таки по всем точкам разрыва ступенчатой функции, но делать это грамотно с учетом также и величины $1/n$.

³Речь о задаче «для выборок из *разных* генеральных совокупностей и набора соответствующих им плотностей указать взаимные соответствия»

3 Практические методы непараметрической статистики.

Критерий знаков

Достаточно частое использование непараметрической статистики связано со следующей задачей: *опровергается ли на данном уровне значимости, что две выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_n равной длины взяты из одной генеральной совокупности (а про гипотетическое распределение не говорится ничего, кроме непрерывности)?* Мы уже видели, как к этому вопросу можно подступиться критерием Колмогорова-Смирнова, однако есть и более грубые, но простые методы, такие как критерий знаков.

3.1 Описание теста

Если сравниваемые выборки не пересекаются и взяты из одной генеральной совокупности, то вероятности появления положительных и отрицательных разностей $x_i - y_i$ должны быть равны. Нулевые разности *не должны наблюдаться* в силу непрерывности изучаемой случайной величины, таким образом

$$P(x_i - y_i > 0) = P(x_i - y_i < 0) = \frac{1}{2}$$

следовательно, надо проверять, что число положительных разностей имеет биномиальное распределение с параметром $p_+ = \frac{1}{2}$, альтернативой чему выступают условие: $p_+ > \frac{1}{2}$ или $p_+ < \frac{1}{2}$. При заданном уровне значимости α гипотезу о $p_+ = \frac{1}{2}$ надо отклонить, если наблюдается аномально большое или аномально малое число r положительных разностей. В терминах распределения Бернулли для отклонения гипотезы на уровне значимости α должно быть выполнено одно из неравенств:

$$2^{-n} \sum_{k=r}^n \binom{n}{k} \leq \frac{\alpha}{2} \quad 2^{-n} \sum_{k=0}^r \binom{n}{k} \leq \frac{\alpha}{2}$$

Более внимательный взгляд на тест показывает, что тут проверяют лишь **гипотезу о равенстве медиан у двух независимых компонент непрерывного случайного вектора, проецируя эту задачу на проверку гипотезы о значении параметра биномиальной случайной величины, определяемой числом появления положительных разностей между компонентами в выборке случайного вектора**. В случае дискретных случайных величин нужна дополнительная аккуратность в формулировке гипотез и альтернатив. Рассмотренный выше тест имел в виду пару альтернатив $H_0 : p_+ = 1/2$ и $H_1 : p_+ \neq 1/2$.

Поскольку реально тестируется лишь дискретное биномиальное распределение, можно рассматривать и иные пары альтернатив для значений его параметра p_+ :

$H_0 : p_+ \leq 1/2$ и $H_1 : p_+ > 1/2$. Для отклонения гипотезы H_0 нужно неравенство $2^{-n} \sum_{k=r}^n \binom{n}{k} \leq \alpha$

$H_0 : p_+ \geq 1/2$ и $H_1 : p_+ < 1/2$. Для отклонения гипотезы H_0 нужно неравенство $2^{-n} \sum_{k=0}^r \binom{n}{k} \leq \alpha$

При больших значениях n прямой подсчет биномиальных коэффициентов неудобен и потому используют аппроксимации биномиального закона гауссовым и подсчитывают в терминах отклонения функций распределения в метрике L_2 , это в свою очередь приводит к сравнению наблюдаемых величин с квантилями $F_q[k, l]$ распределения Фишера-Снедекора:

$H_0 : p_+ \leq 1/2$ и $H_1 : p_+ > 1/2$. Для отклонения гипотезы H_0 нужно неравенство $\frac{r}{n-r+1} \geq F_{1-\alpha}[2(n-r+1), 2r]$

$H_0 : p_+ \geq 1/2$ и $H_1 : p_+ < 1/2$. Для отклонения гипотезы H_0 нужно неравенство $\frac{n-r}{r+1} \geq F_{1-\alpha}[2(r+1), 2(n-r)]$

$H_0 : p_+ = 1/2$ и $H_1 : p_+ \neq 1/2$. Для отклонения гипотезы H_0 должно быть выполнено одно из двух предыдущих неравенств, но с заменой показателя квантиля $1 - \alpha$ на $1 - \alpha/2$

3.2 Примеры и задачи

Как и всегда, вычисления за нас делает компьютер, вот ссылка на соответствующую функцию языка python `statsmodels.stats.descriptivestats.sign_test`

- Компьютерный эксперимент: заметит ли на уровне $\alpha = 0.05$ критерий знаков разницу двух равномерных распределений соответственно на отрезках $[0.0, 0.95]$ и $[0.0, 1.0]$ при выборках длиной 100? При выборках длиной 1000?

По факту этот довольно грубый непараметрический критерий используют в медицинской статистике, где эксперименты на людях и вдобавок дорогие. При этом вместо независимых повторений одного эксперимента рассматривают показатели у разных людей — весьма спорно, в каких случаях можно ли это отождествлять. В данном случае грубый тест заметит только сильные отличия.

- Рассматривают влияние курса БАД на снижение веса у десяти испытуемых

вес до	68	80	92	81	70	79	78	66	57	76
вес после	63	85	87	79	74	75	76	65	56	71

Проверить на уровне $\alpha = 0.1$, влияет ли курс и как?

4 Практические методы непараметрической статистики.

Ранговые методы

В предыдущем разделе рассматривалась задача о двух выборках : опровергается ли на данном уровне значимости, что две выборки x_1, x_2, \dots, x_m и y_1, y_2, \dots, y_n взяты из одной генеральной совокупности. В принципе этот вопрос весьма важен для клинической медицинской диагностики и неудивительно, что возникло несколько методик исследования. Варианты решения при помощи KS и критерия знаков уже обсуждались, рассмотрим еще достаточно эффективный подход Уилкоксона-Манна-Уитни (Wilcoxon, Mann, Whitney), он также входит, например, в библиотеку языка Python (см. `scipy.stats.mannwhitneyu` `scipy.stats.wilcoxon`). Применение в медицинской практике: для каждого больного вычисляется величина изменения признака. Все изменения упорядочивают по абсолютной величине (без учета знака). Затем рангам приписывают знак изменения и суммируют эти «знаковые ранги» - в результате получается значение статистики Уилкоксона.

Полное объяснение деталей соответствующего критерия достаточно громоздко, но стоит указать некоторые идеи, на которых этот критерий основан.

4.1 Идея теста, результаты и алгоритм

Каждая из двух выборок возникла в модели независимых повторений, следовательно в соответствующих генеральных совокупностях с точки зрения мер нет никакой разницы, если внутри выборки менять значения перестановками. Таких перестановок в первой выборке $m!$, во второй — $n!$ (подход Уилкоксона, как и критерий знаков, в своей классической форме предполагает непрерывные распределения и потому выборки с повторениями ненаблюдаемы). Если выполняется гипотеза H_0 о том, что эти генеральные совокупности на самом деле одна и та же, то выборки можно слить в одну⁴ и рассматривать $(m+n)!$ не меняющих меру перестановок в объединении. Что меняется при перестановках понять легко: если в одном варианте пара значений шла «правильно» т.е. по возрастанию, то после перестановки она может стать уже «неправильной», т.е. не возрастающей. Идея критерия состоит в том, чтобы сравнивать показатели относящиеся к «правильным/неправильным» парам, возникающим в исходных выборках x_1, x_2, \dots, x_m и y_1, y_2, \dots, y_n . Оказывается, что сравнивать эти показатели удобно рассмотрев упорядочение значений по возрастанию в объединенной выборке. Показатели R_1, R_2 , которые будут основой критерия, они вводятся так: для каждого значения x_i из первой выборки возникает порядковый номер r_i этого значения в объединенной упорядоченной выборке. Это дает $R_1 = \sum_{k=1}^m r_k$, аналогичным образом определяют $R_2 = \sum_{k=1}^n r_k$. В зависимости от размера выборки имеются два подхода:

1. Для маленьких выборок $n \leq 25$ критерий Уилкоксона, в котором критическая область явно сосчитана по комбинаторным формулам. Тест Уилкоксона также иногда называется критерием Манна-Уитни для малых выборок.
2. Для больших выборок U-критерий Манна-Уитни, использующий аппроксимацию распределения соответствующей статистики

Именно, Манн, Уитни в середине XX века показали, что при больших m, n распределение статистики U

$$U = \min \left(R_1 - \frac{1}{2}m(m+1), R_2 - \frac{1}{2}n(n+1) \right)$$

быстро сходится к гауссову закону с матожиданием $mn/2$ и дисперсией $(m+n+1)mn/12$ и, стало быть, если значения U оказались маргинальными (= лежащими на хвостах функции гауссовой плотности), то нужно отбрасывать нулевую гипотезу H_0 . Более точно, при заданном уровне значимости α гипотеза H_0 отклоняется, если вычисленное по выборками значение

$$Z^* = \frac{U - \frac{mn}{2}}{\sqrt{\frac{mn}{12}(m+n+1)}}$$

по абсолютной величине больше квантиля $q_{1-\alpha/2}$ для нормального закона $\mathcal{N}(0, 1)$ (это т.н. двусторонний критерий, предполагающий альтернативу «выборки исходно взяты из разных, отличающихся распределениями, генеральных совокупностей» для H_0).

⁴Понятно, что вне этой гипотезы объединение выборок может вообще говоря привести к самым парадоксальным последствиям!

Замечание В изложенном тесте речь шла не о выборке пар у случайного вектора, а о сравнении свойств двух выборок — то есть постановка вопроса отличается от аналогичной в ранее изложенном тесте знаков.

Часто эти же тесты можно встретить и проблемы сравнения медиан m_x, m_y у двух выборок. Формально, разумеется, объединять выборки в таком случае нельзя, но поскольку в построении статистик реально используются не сами числа, а их ранги, то трюк заключен в формулировке альтернатив: H_0 утверждает, что медианы $m_x \leq m_y$, альтернатива H_1 утверждает, что $m_x > m_y$. В этом случае нужен *односторонний* критерий.

4.2 Упражнения и задачи

1. Для выборок S_04_001.dat и S_04_003.dat критерием Манна и Уитни на уровне значимости 0.05 проверить гипотезу, что они взяты из одного распределения.
2. Самостоятельно разобраться как устроена статистика (формула) для одностороннего критерия Манна-Уитни. Соответствующее правило тоже опирается при больших m, n на квантили гауссовского закона: но в чем именно отличие правила, для пары альтернатив $m_x \leq m_y$ и $m_x > m_y$ от правила для пары альтернатив $m_x < m_y$ и $m_x \geq m_y$?
3. Для выборок S_04_001.dat и S_04_003.dat подходящим критерием на уровне значимости 0.01 проверить гипотезу, что медиана первого распределения не превосходит медианы второго.

4.3 О преобразованиях и рангах

В изложенном тесте использовали порядковые номера в выборке вместо рассмотрения исходных значений. Этот прием полезен и в других ситуациях, например при оценивании по выборкам меры зависимости двух разных случайных величин.

Как известно, в теории вероятностей есть понятие независимости, а зависимость означает нарушение свойства независимости. Если ограничиваться случайными величинами с конечными первым и вторым моментом, то для зависимых случайных величин возникает достаточный признак зависимости — коэффициент корреляции, который допускает геометрическую трактовку в терминах косинуса угла между векторами в линейном пространстве таких случайных величин. Если, например, два вектора параллельны и, стало быть, пропорциональны (зависимы), то коэффициент корреляции оказывается по модулю равным 1, а по построению этого линейного пространства со скалярным произведением, независимые с.в. окажутся ортогональными и тем самым, коэффициент корреляции будет нулем.

Таким образом, оценка коэффициента корреляции по выборке — практически важная и часто встречающаяся задача. Интуитивно кажется довольно очевидной оценка Пирсона (Pearson) для ρ , сконструированная по plug-in оценкам математического ожидания и дисперсии компонент случайного вектора (ξ, η) :

$$\hat{\rho} = \frac{\sum_k [(x_k - \bar{x})(y_k - \bar{y})]}{\sqrt{\sum_k (x_k - \bar{x})^2 \sum_k (y_k - \bar{y})^2}}$$

(подразумевается, что выборка значений вектора состоит из пар (x_i, y_i)).

Однако эта интуитивно понятная формула на практике обладает статистически неприятными свойствами, что и показывают первые два упражнения из раздела 4.4 ниже. Из них понятно, что проблемное поведение связано с эффектом масштаба выброса. Идея, похожая на ту, что использовалась в критериях Уилкоксона, Манна и Уитни подсказывает эффект уничтожения масштаба по схеме: выборка пар значений (x_i, y_i) вектора (α, β) трансформируются в выборку пар $F_n^*(x_i), G_n^*(y_i)$, где F_n^*, G_n^* соответственно эмпирические функции распределения координат вектора. Иными словами, вместо коэффициента корреляции $\rho_{\alpha, \beta}$ вычисляют по формуле Пирсона оценку для $\rho_{F_n^*(\alpha), G_n^*(\beta)}$. Ответ называется оценкой коэффициента корреляции по Спирмену (Spearman) $\hat{\rho}_s$. Результат преобразования случайного вектора (α, β) к вектору $(F_n^*(\alpha), G_n^*(\beta))$ называется (двумерной) копулой и достаточно часто встречается в статистическом анализе многомерных данных.

Часто встречается также еще одна характеристика зависимости, построенная в духе критерия знаков — она называется корреляционным коэффициентом τ Кендалла (Kendall Correlation Coefficient) и вычисляется по формуле

$$\hat{\tau} = \frac{2}{n(n-1)} \sum_{i < j} \text{sign}(x_i - x_j) \text{sign}(y_i - y_j)$$

Так же как и для привычного коэффициента корреляции в случае линейной связи двух выборок случайных величин $\hat{\tau}$ принимает значения ± 1 , напротив для выборок из независимых с.в. значение $\hat{\tau}$ мало (в смысле концентрации значений вблизи нуля).

В языке `python` все необходимые вычисления, разумеется, реализованы и включены в `scipy.stats`. **Неприятность: коэффициент корреляции Кендалла определен в разных местах по-разному. В частности в `python` он определяется по умолчанию НЕ ТАК. А как? Разберитесь по `scipy.stats` и будьте внимательны, если вдруг придется реально вычислять его.**

При вычислениях мы получаем случайную величину, которая помогает решать вопрос о гипотезе независимости. Следовательно, на практике необходимо знать, попало ли наше вычисление соответствующей статистики в критическую область, установленную для данной значимости критерия. Поэтому библиотечная функция также вычисляет соответствующий квантиль и показывает следует ли отвергнуть гипотезу о независимости. Разберитесь самостоятельно, как использовать эти правила для всех трех оценок зависимости, указанных выше.

В случае непрерывного случайного вектора (α, β) имеющего плотность можно еще указать аналитические формулы для коэффициентов корреляции Спирмена ρ_s и Кендалла ρ_τ , а именно, рассмотрим двумерную копулу, построенную по случайному вектору — это случайный вектор с носителем $[0, 1] \times [0, 1]$ и с двумерной дифференцируемой функцией распределения $C(x, y)$. Тогда

$$\rho_s = 12 \iint_{[0,1]^2} [C(x, y)] dx dy - 3$$

$$\rho_\tau = 1 - 4 \iint_{[0,1]^2} \left[\frac{\partial C}{\partial x}(x, y) + \frac{\partial C}{\partial y}(x, y) \right] dx dy$$

а выборочные формулы выше будут соответственно оценками этих величин. Здесь мы это оставляем без доказательства, ссылка на статью doi: 10.1016/j.jspi.2006.06.045

4.4 Упражнения

Приготовьте выборку X из 999 значений случайного вектора (α, β) с независимыми и равномерно распределенными на $[0, 1]$ компонентами.

1. Добавьте в исходную выборку X пару значений (выброс) $(100, 100)$ — это будет множество данных Y и вычислите для множества данных Y значения коэффициента корреляции по формуле Пирсона.
2. Добавьте в выборку X пару значений (выброс) $(5, 5)$ — это будет множество данных Y и вычислите для множества данных Y значения по формуле Пирсона.
3. Добавьте в исходную выборку X пару значений (выброс) $(100, 100)$ — это будет множество данных Y и вычислите для множества данных Y значения коэффициента корреляции по формуле Спирмена.
4. Добавьте в выборку X пару значений (выброс) $(5, 5)$ — это будет множество данных Y и вычислите для множества данных Y значения по формуле Спирмена.
5. Сравните с ответами по формуле Пирсона.