Семинары по статистике 2021

1 Введение: принципы работы с данными в статистике

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

Наличие некоторой случайной величины (вариант – случайного вектора) $\xi : \Omega \to \mathbb{R}$, 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 $\xi_1, \xi_2, \ldots$ независимых и распределенных так же как ξ), задающих меру на пространстве $\mathbb{R}^\infty$.

Наличие конечной последовательности чисел $\xi_1(\omega_1), \xi_2(\omega_2), \ldots \xi_n(\omega_n)$, 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 $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$ 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 $\beta = c\dot{\alpha} + d + \epsilon$, где c, d вещественные числа, а ϵ – величина с нулевым математическим ожиданием. Оценить по выборке из вектора $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$ 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 $E(\xi)$ черезвычайно 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 содержаться 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 строится, 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 $k = \left[\frac{\sqrt{n}}{3}\right]$ (фольклор)

Правило Стерджеса $k = 1 + [\log_2 n]$ (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 $h = \left[s^* \cdot \frac{3.49}{\sqrt[3]{n}}\right]$ (D.Scott, 1979)

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

здесь квадратные скобки обозначают целую часть, а s^* и q^*_t — вычисленные по выборке соответственно эмпирическую дисперсию[1] 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.

[1] О 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 строится, 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 билогарифмической 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 билогарифмической шкалах. Объяснить какая картинка соответствует какому типу изображению.

1.4 Как самому создать статистический тест

Вот как можно создать 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:

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

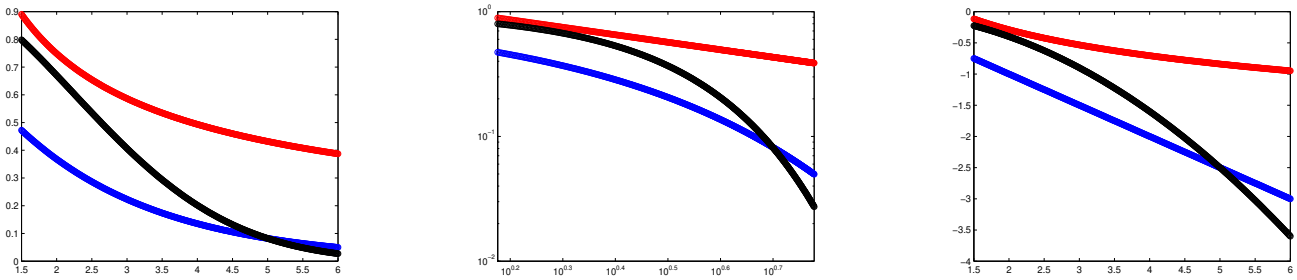

Рис. 1: Вид графиков полиномиального (синий), экспоненциального (красный) и сверхэкспоненциального (черный) убывания в разных шкалах: линейной, полулогарифмической и билогарифмической шкалах.

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 (приближенно) 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 (приближенно) значения квантиля, например, для $t = 0.95$?

1.4.2 Задание с вариантами теста

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

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

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

Варианты значений n

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

Варианты значений C

1. C=0. Число столбиков m выбирается по правилу Большого Пальца
2. C=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. $D_{[n,m]} = \max\limits_i |w_i - p_i|$
2. D=2. $D_{[n,m]} = \sum\limits_i |w_i - p_i|$
3. D=3. $D_{[n,m]} = \sqrt{\sum\limits_i (w_i - p_i)^2}$
4. D=4. $D_{[n,m]} = \sum\limits_i \frac{|w_i - p_i|}{p_i}$
5. D=5. $D_{[n,m]} = \sum\limits_i \frac{(w_i - p_i)^2}{p_i(1-p_i)}$

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 числу 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 $\{x|F_n^*(x) < F_\xi(x)\}$)

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

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}:

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

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 длиною 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:

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

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. Впрочем, здесь появляется трудный вопрос об уровне значимости применения нескольких тестов сразу и о приближениях в используемых расчетных формулах.

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 аи формы распределения В частности вернемся к задаче 1.3.1[3].

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): 0, 0.3 1.5,

2. Компьютерные эксперименты, показывающие роль параметров и приближений

   (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 $\mathcal{N}(0,1)$.

   (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 $\mathcal{N}(0,1)$.

   (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.

   (d) Подготовьте 1000 выборок длины 50 значений случайной величины, распределенной по закону Коши с параметром 1 и сосчитайте 95% границу отклонения эмпирических функцийа 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.

2.2.2 Замечание

Вычисление точной верхней грани расстояний между построенной по выборке ступенчатой функцией и построенной по известному 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 $1/n$.

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