Сборник долгосрочных заданий по курсу математической статистики

(проект v. 0.9.1)

Д.Б. Фомин, А.Б. Чухно

7 ноября 2019 г.

Оглавление

Введение

При подготовке материалов для долгосрочных домашних заданий по курсу математической статистики авторы подбирали задания так, чтобы в ходе выполнения заданий студенты могли

- ▷ закрепить знание основных понятий, теорем и утверждений курса;
- ▷ проводить собственные исследования на интересных и необычных примерах;
- ▷ получить практический опыт применения статистических методов;
- ▷ освоить современные программные средства статистической обработки данных и работы с графикой.

Домашнее задание 1

Вероятностные распределения

Данное домашнее задание посвящено повторению основ теории вероятностей.

Задание состоит из следующих пунктов:

1. Выбор одного дискретного распределения и одного непрерывного распределения.
2. Описание основных характеристик распределения.
3. Поиск примеров событий, которые могут быть описаны выбранными случайными величинами.
4. Описание способа моделирования выбранных случайных величин.

1.1 Выбор одного дискретного распределения и одного непрерывного распределения

Для выполнения домашнего задания необходимо выбрать одно из предложенных дискретных распределений (см. табл. 1.1) и одно из предложенных непрерывных распределений (см. табл. 1.2).

Выбор распределений необходимо согласовать с преподавателем, ведущим практические занятия.

В таблицах 1.1, 1.2 звездочкой обозначены распределения не входящие в обязательный минимум. В тоже время, при желании студента можно дополнительно рассмотреть распределения помеченные знаком «*» и/или иные не рассмотренные в таблицах 1.1, 1.2, что может лишь положительно повлиять на итоговую оценку студента как по домашним работам, так и за весь курс.

Таблица 1.1: Дискретные распределения

№	Распределение	Закон распределения
1	Бернулли	$p^x \cdot (1-p)^{1-x}$, $x \in \{0,1\}$, $0 < p < 1$
2	Биномиальное	$p(x) = \binom{n}{x} p^x q^{n-x}$, $x \in \{0,1,\ldots,n\}$, $n \geq 1$, $0 < p < 1$, $q = 1-p$
3	Геометрическое	$p(x) = pq^x$, $x \in \mathbb{N} \cup \{0\}$, $0 < p < 1$, $q = 1-p$
4	Отрицательное биномиальное	$p(x) = \binom{z+m-1}{z} p^m q^z$, $z \in \mathbb{N} \cup \{0\}$, $m \in \mathbb{N}$, $0 < p < 1$, $q = 1-p$
5	Дискретное равномерное	$p(x) = n^{-1}$, $x \in \{a,\ldots,a+n-1\}$
6	Пуассона	$p(x) = \frac{\mu^x}{x!} e^{-\mu}$, $\mu > 0$
7	Гипергеометрическое распределение	$p(x) = \frac{\binom{M}{x}\binom{N-x}{N_M}}{\binom{N}{n}}$, $N, M, n \in \mathbb{N} \cup \{0\}$, $M \leq N, n \leq N$, $x \in \{\max(0, M+n-N), \min(M,n)\}$
8	Логарифмическое распределение	$p(x) = -\ln(q)^{-1} \cdot p^x \cdot x^{-1}$, $x \in \mathbb{N}$, $0 < p < 1$, $q = 1-p$
9	Ципфа*	
10	Бореля-Таннера*	
11	Дзета*	
12	Пойа*	

Таблица 1.2: Непрерывные распределения

№	Распределение	Закон распределения
1	Нормальное	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$, $x, \mu, \sigma \in \mathbb{R}$, $\sigma > 0$
2	Лапласа	$f(x) = \frac{1}{2} \cdot \lambda \cdot \exp\{-\lambda \lvert x - \mu \rvert\}$, $x, \mu, \lambda \in \mathbb{R}$, $\lambda > 0$
3	Коши	$f(x) = \frac{\lambda}{\pi\left(\lambda^2 + (x-\mu)^2\right)}$, $x, \mu, \lambda \in \mathbb{R}$, $\lambda > 0$
4	Экспоненциальное	$f(x) = \lambda e^{-\lambda x}$, $x \in \mathbb{R}$, $x > 0$
5	Гамма	$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$, $x, \lambda, \alpha \in \mathbb{R}$, $x, \lambda, \alpha > 0$
6	Эрланга*	$f(x) = \frac{\lambda^m}{(m-1)!} x^{m-1} e^{-\lambda x}$, $x, \lambda \in \mathbb{R}$, $x, \lambda > 0$, $m \in \mathbb{N}$
7	Рэлея*	
8	Максвелла*	
9	Логнормальное*	
10	Парето*	
11	Равномерное	
12	Треугольное	
13	χ^2	

1.2 Описание основных характеристик распределения

Для каждого из выбранного распределения необходимо вывести (любым возможным способом) его основные характеристики:

- ▷ математическое ожидание,
- ▷ дисперсию.

Привести (если есть) производящую и характеристическую функцию, функцию распределения. В случае наличия трудностей с выводом тех или иных функций по согласованию с преподавателем, ведущим практические занятия, их можно не приводить.

Для каждой из выбранных распределений графически изобразить

- ▷ плотность (для непрерывных распределений),
- ▷ гистограмму вероятностей (для дискретных распределений),
- ▷ функцию распределения.

1.3 Поиск примеров событий, которые могут быть описаны выбранными случайными величинами

Для каждого из выбранных распределений необходимо

- ▷ привести типичные интерпретации распределения (как распределение определяется);
- ▷ известные соотношения между распределениями;
- ▷ нетипичные интерпретации распределения (примеры событий, описываемых рассматриваемым распределением).

К типичным интерпретациям будем относить математические модели, описываемые данным распределением. К нетипичным интерпретациям отнесем способы применения распределений при решении практических задач. Заметим, что **необходимо формально обосновать хотя бы одну типичную и одну нетипичную интерпретацию**.

Пример 1.1 *Рассмотрим пример для распределения Пуассона.* ***Типичной интерпретацией*** *для него является следующая ситуация. Каждый раз,*

подходя к кассе и попадая в очередь, вы , наверняка, задавались вопросом: "Как долго мне стоять в этой очереди?"Или же, излагая данный вопрос на языке теории вероятностей "С какой вероятностью я пройду к кассе за t минут, если передо мной n человек?". Пусть также выполнены очевидные, но необходимые с точки зрения теории постулаты:

1. *за малый промежуток времени кассир не сможет обслужить больше одного покупателя;*

2. *количества обслуженных клиентов за непересекающиеся промежутки времени не зависят друг от друга,*

3. *Среднее количество $E\xi$ покупателей, которых обслужил кассир, за временной промежуток длины l, пропорционально с параметром λ длине этого промежутка. $E\xi \approx \lambda \cdot l$*

Тогда, для вычисления вероятности быть обслуженным кассиром за время t, воспользуемся следующим рассуждением: временной промежуток длины t, в течение которого хочется отстоять очередь, разделим на m одинаковых отрезочков Δt_i, $i = 1, \ldots .m$ при достаточно большом m, чтобы выполнялся постулат 1.

Коль скоро в каждый малый промежуток времени может обслуживаться не более чем один покупатель, то среднее число покупателей в этом промежутке равно вероятности события, что покупатель будет обслужен. Это следует из того, что мат. ожидание бернуллиевской случайной величины равно вероятности её успеха. То есть вероятность p, что в одном из наших маленьких отрезочков Δt_i произошло обслуживание покупателя, примерно равна $\frac{\lambda}{m}$.

Тогда вероятность p_n, что было обслужено n покупателей, примерно будет равна $p_{n,m} \approx C_m^n \left(\frac{\lambda}{m}\right)^n \left(1 - \frac{\lambda}{m}\right)^{m-n}$, тоесть имеет биномиальное распределение с параметрами $Bi\left(m, \frac{\lambda}{m}\right)$. Ясно, что при увеличении числа m примерная вероятность будет приближаться к искомой. Осталось заметить, что для биномиального распределения с такими параметрами будет выполняться теорема Пуассона, следовательно $p_{n,m} \to p_n = \frac{\lambda^n \cdot e^{-\lambda}}{n!}$. Это ситуация является одной из типичных, где возникает распределение Пауссона.

Нетипичную интерпретацию *рассмотрим для экспоненциального распределения. Пусть имеется некоторое видео в интернете. Рассмотрим процесс появления комментариев под ним. Для начала стоит заметить, что чем больше существует видео, тем меньше комментариев под ним пишут в единицу времени, при этом также будем предполагать, что каждый оставляющий новый комментарий делает это независимо от остальных комментариев, и под конец, предполагая, что в достаточно малый промежуток времени может быть написано не более одного комментария.*

Обозначим через X_s – число комментариев написанных под видео за время s. В описанных выше условиях распределение числа комментариев будет иметь следующее свойство: для $t > s$ $X_t - X_s$ $\Pi(\lambda(t-s))$. Параметр λ – интенсивность появления комментариев.

В данной модели ставится вопрос, а как распределено время между появлением соседних комментариев. Попытаемся ответить на данный вопрос.

Обозначим через t_n – момент появления n-го комментария, тогда $X_{t_n} = n$ и $X_{t_n-0} = n-1$. Момент времени t_n – непрерывная случайная величина. Событие $(t_n < x)$ заключается в том, что к моменту времени x будет написано не менее n комментариев $(t_n < x) = \bigcup\limits_{k=n}^{\infty}(X_x = k) = \overline{\bigcup\limits_{k=0}^{n-1}(X_x = k)}$, тогда

$$P(t_n < x) = 1 - P\Big(\bigcup_{k=0}^{n-1}(X_x = k)\Big) =$$

т.к. при каждом k события $(X_x = k)$ несовместны то

$$P(t_n < x) = 1 - \sum_{k=0}^{n-1} P(X_x = k)\Big) =$$

$$= 1 - \sum_{k=0}^{n-1} \frac{(\lambda x)^k e^{-\lambda x}}{k!} =$$

$$= \int_0^x \frac{t^{n-1} e^{-\lambda t} \lambda^n}{(n-1)!}\, dt$$

Последнее равенство проверяется взятием по частям интеграла.

В итоге получили, что момент появления n-го комментария будет иметь распределение Эрланга, и в частности при $n = 1$ получим экспоненциальное распределение с параметром λ. Объединяя всю эту информацию, можно сказать, что время между появлениями комментариев распределено экспоненциально. таким образом в модели, касающейся казалось бы дискретных объектов (числа комментариев), проявит себя экспоненциальное распределение. Это пример нетипичной интерпретации.

Замечание 1.2 *Примеры нетипичной интерпретации могут быть вашего авторства (что необходимо будет отметить при подготовке отчета о контрольной работе) либо найдены в каком-либо источнике. В случае, если это не ваше предположение, необходимо дать правильную ссылку на материал.*

Замечание 1.3 *Необходимо внимательно отнестись к выбору нетипичной интерпретации так как в последующих домашних контрольных работах необходимо будет провести верификацию хотя бы для одной нетипичной интерпретации.*

1.4 Описание способа моделирования выбранных случайных величин

Основной задачей данной контрольной работы является практическое моделирование выбранных случайных величин (построении случайной выборки из заданных распределений).

В нашем курсе вопрос построения не рассматривается и остается на самостоятельное изучение. Данное направление является хорошо освященным в литературе (см. например [1, 2]).

При построении случайной выборки можно пользоваться существующими библиотеками выработки случайных чисел из заданных распределений.

Замечание 1.4 *При использовании библиотек для построения случайной выборки из заданных распределений необходимо обратить внимание, что распределение может быть заданно отличным от способов, представленных в таблицах 1.1, 1.2.*

Замечание 1.5 *Важно! Можно предложить свой способ построения случайной выборки из заданных распределений. Обязательно необходимо доказать, что предложенный способ корректен.*

В случае использования стандартных или сторонних библиотек необходимо привести обоснование того, что используемая функция действительно позволяет построить выборку из заданного распределения.

Пусть стоит задача построения случайной выборки случайной величины ξ, имеющей равномерное распределение на множестве $\{0, 1, 2\}$. Необходимо понимать, что использование стандартных функций языка C «`rand() % 3`» может быть некорректно по многим причинам.

Замечание 1.6 *В отчете должен быть представлен код, с помощью которого производилось моделирование случайной величины.*

Домашнее задание 2

Основные понятия математической статистики

Данное домашнее задание является продолжением предыдущего домашнего задания посвящено закреплению пройденного материала по основам математической статистики.

1. Моделирование выбранных случайных величин.
2. Построение эмпирической функции распределения.
3. Построение вариационного ряда выборки.
4. Построение гистограммы и полигон частот.

2.1 Моделирование выбранных случайных величин

Для каждой из выбранных случайных величин необходимо построить по 5 реализаций выборок следующий объемов n: $n \in \{5, 10, 100, 1000, 10^5\}$. То есть всего 50 различных выборок. В бумажной версии отчета должны быть приведены выборки для $n = 5$ и $n = 10$.

2.2 Построение эмпирической функции распределения

Для каждой выборки необходимо построить эмпирическую функцию распределения. Графики необходимо привести в отчете. На одном графике необходимо отобразить эмпирическую функцию распределения и график функции распределения случайной величины.

Для каждого $n \in \{5, 10, 100, 1000, 10^5\}$ необходимо найти точную верхнюю границу разности каждой пары эмпирических функций распределения (то есть для каждого n необходимо найти $\binom{5}{2}$ чисел).

2.3 Построение вариационного ряда выборки

Для каждого n необходимо построить вариационный ряд выборки. В отчете привести пример для $n = 5$ и $n = 10$.

Найти выборочную квантиль уровня 0.1, 0.5, 0.7. Сравнить их с квантилями рассматриваемых распределений.

2.4 Построение гистограммы и полигон частот

Для каждого распределения и для каждого n необходимо построить и привести в отчете:

- ▷ гистограмму частот
- ▷ полигон частот
- ▷ сравнение с плотностью распределения для непрерывных распределений и функцией вероятности для дискретных распределений.

Домашнее задание 3

Оценки

Третье долгосрочное домашнее задание посвящено разделу математической статистики, связанным с нахождением неизвестных параметров распределений.

Задание состоит из следующих пунктов:

1. Нахождение *выборочного среднего* и *выборочной дисперсии*
2. Нахождение параметров распределений событий
3. Работа с реальными данными.

3.1 Нахождение выборочного среднего и выборочной дисперсии

Для каждого из распределений для каждой выработанной во втором домашнем задании выборки найти выборочное среднее и выборочную дисперсию.

Ответить на вопросы

▷ какими свойствами данные оценки обладают?

Также необходимо сравнить истинные значения параметров и значения их оценок. Объясните полученные результаты с точки зрения теории.

3.2 Нахождение параметров распределений событий

Для каждого из выбранных распределений необходимо предложить оценку X для оцениваемого параметра θ и выполнить следующее:

▷ Проверить, является преложенная оценка несмещённой.

- ▷ Проверить, является ли предложенная оценка X состоятельной для оцениваемого параметра θ.

- ▷ Проверить, является ли преложенная оценка оптимальной, эффективной. Если нет, то (по возможности) построить оптимальную оценку для параметра θ.

В данном разделе можно рассмотреть разные способы построения известных статистических оценок а также провести их сравнение. Какими известными свойствами обладают полученные оценки.

Также необходимо сравнить истинные значения оцениваемых параметров и значения их оценок. Объясните полученные результаты с точки зрения теории.

3.3 Работа с данными

Для выбранного нетипичной интерпретации, обоснованной в первой домашней работе найти данные, соответствующие данной интерпретации. При этом необходимо привести источники данных, а также сами данные (или постоянную ссылку на данные, если они взяты из открытых источников.)

В случае, если рассматриваемые данные не соответствуют интерпретации из первой домашней работы, необходимо привести обоснование выбора данных.

Для полученных данных необходимо проделать такую же работу как и с построенными выборками, а именно:

1. привести значение выборочного среднего и выборочной дисперсии.

2. привести значение преложенной оценки X и (в случае их несовпадения) значение оптимальной оценки.

Домашнее задание 4

Проверка статистических гипотез

4.1 Проверка гипотез о виде распределения

Задание состоит из следующих пунктов:

1. Для каждой из рассматриваемых распределений необходимо проверить гипотезу о виде распределения.

Проверка гипотезы о виде распределения должна состоять из следующих этапов:

- ▷ Описание применяемого критерия.
- ▷ Описание свойств рассматриваемого критеря; описание его положительных сторон и недостатков.
- ▷ Выбор уровня значимости (рассмотреть следующие: 0.1, 0.05).
- ▷ Ответ на вопрос о соответствии выборки рассматриваемому распределению.

Замечание 4.1 *Применение статистических критериев необходимо проводить для выборок объема не меньше 1 000.*

Замечание 4.2 *Желательно проиллюстрировать применение критерия с помощью графиков, диаграмм или любых иных средств.*

По результатам проведенных исследований необходимо сопоставить получившиеся результаты с теорией.

4.1.1 Задание для рассматриваемых распределений

Для каждого из рассматриваемых распределений необходимо рассмотреть следующие статистики:

- ▷ Критерий согласия Колмогорова-Смирнова (для абсолютно непрерывных распределений)
- ▷ Критерий согласия хи-квадрат
- ▷ Критерий согласия Колмогорова-Смирнова для сложной гипотезы (для абсолютно непрерывных распределений)
- ▷ Критерий согласия хи-квадрат для сложной гипотезы

Так как известны параметры распределений, для которых были построены реализации выборки, то проверка гипотез о виде распределения с использованием критерия согласия Колмогорова-Смирнова и критерия согласия хи-квадрат происходит и использованием соответствующих статистик и их предельных распределений.

Описание статистик критерия и их распределений можно найти, например, в [1, 3, 4].

Замечание 4.3 *В [5–7] предлагается вместо статистики D_n использовать следующий вид статистики с поправкой Большева*

$$S = \frac{6nD_n + 1}{6\sqrt{n}},$$

которая также имеет распределение Колмогорова, но сходится к нему быстрее, что, согласно [5–7], позволяет использовать ее при меньших объемах данных.

Также при малых объемах выборки можно пользоваться рекуррентными соотношениями для Критерия Колмогорова-Смирнова (см. раздел дополнительные задания на стр. 23).

При применении критерия согласия хи-квадрат для случая непрерывных распределений как и бесконечных дискретных (как и некоторых конечных) необходимо применять предварительную группировку наблюдений. В литературе часто встречается эвристическое правило Старджесса для определения «оптимального» числа интервалов. Вопросы выбора числа интервалов со списком литературы можно найти в [3].

В виду того, что при малых значениях вероятностей и недостаточном объеме данных статистика хи-квадрат может давать погрешности, рекомендуется пользоваться методом равных вероятностей для построения классов [8],

который заключается в том, что ожидаемые вероятности попадания наблюдения в определенный класс оказались равными.

Следующей задачей является рассмотрение сложной гипотезы. Будем считать, что известен вид распределения, но не известны его параметры.

Случай сложных гипотез для критериев согласия Колмогорова-Смирнова и хи-квадрат состоит из следующих этапов:

- ▷ построение оценки неизвестного параметра методом максимального правдоподобия;
- ▷ вычисление значения статистики, соответствующей рассматриваемому критерию;
- ▷ вычисление критической границы критерия в зависимости от выбранного уровня значимости.

Будем считать, что построение оценки методом максимального правдоподобия и вычисление значений статистики при фиксации параметра является простой и уже решенной задачей.

Рассмотрим вопрос вычисления критической границы. Сначала рассмотрим более простой случай — статистика хи-квадрат. В случае, когда определяется m параметров распределений число степеней свободы статистики хи-квадрат, к которой стремится статистика Пирсона, снижается на m.

При проверке сложных гипотез в случае критерия Колмогорова-Смирнова, когда по выборке сначала оцениваются параметры закона, с которым проверяется согласие, непараметрические критерии согласия теряют свойство свободы от распределения [4–7]. При проверке сложных гипотез условные распределения статистик непараметрических критериев согласия (и критерия Колмогорова) зависят как от вида наблюдаемого закона, соответствующего справедливой проверяемой гипотезе, так и от типа оцениваемого параметра и числа оцениваемых параметров.

При этом, различия в предельных распределениях той же самой статистики при проверке простых и сложных гипотез существенны. Только лишь для небольшого количества распределений получены численные значения предельных значений статистик, которые можно найти в [4–7].

В случае, если для рассматриваемого распределения не известно предельных значений, можно воспользоваться следующим подходом: по одной выборке достаточного объема необходимо оценить неизвестный параметр, а по другой проверить гипотезу о виде распределения, как предложено в [5].

4.2 Проверка параметрических гипотез

Данное задание посвящено вопросу различения двух простых параметрических гипотез, а также закреплению основных понятий, и практическому при-

менению методов математической статистики для решения поставленной задачи.

Задание состоит из следующих пунктов:

1. Выбор данных.
2. Постановка задачи.
3. Вычисление функции отношения правдоподобия.
4. Вычисление критической области.
5. Вычисление минимального необходимого количества материала при фиксации минимального возможного значения ошибок первого и второго рода.

4.2.1 Выбор данных

Необходимо найти два набора данных, соответствующих распределению с различными параметрами. Это могут быть оценки к фильмам, статистики игр разных команд и т.п.

В случае наличия сложностей можно выбрать одно из выбранных в предыдущих заданиях распределений и рассмотреть две выборки с разными (но известными) параметрами.

4.2.2 Постановка задачи

Необходимо описать постановку задачи. **Что является гипотезой H_0, что H_1? Что такое ошибка первого и второго рода, функция мощности?**

4.2.3 Вычисление функции отношения правдоподобия

Необходимо описать вид функции $l\left(\overline{X}\right)$ отношения правдоподобия, описать способ ее вычисления.

4.2.4 Вычисление критической области/количества материала

Рассмотрим один из самых сложных вопросов данного задания — вычисление критической области.

Для оценки ошибок первого и второго рода по материалу или вычислении необходимого материала при фиксированных ошибках необходимо знать распределение статистики в случае верности гипотезы H_0 — $l\left(\overline{X}\middle| H_0\right)$ и в

случае верности гипотезы H_1 — $l\left(\overline{X}\middle| H_1\right)$. Для большинства распределений это сделать достаточно сложно.

В случае, если не удается вычислить распределение статистики $l\left(\overline{X}\right)$ в случае верности разных гипотез, предлагается рассмотреть асимптотический подход к различению гипотез.

Прологарифмировав функцию отношения правдоподобия получим сумму одинаково распределенных независимых случайных величин вида

$$z_i = \ln \frac{f_1\left(X_i\right)}{f_2\left(X_i\right)}.$$

Используя Ц.П.Т. можно легко получить распределение статистики $\ln l\left(\overline{X}\right)$ в случае верности каждой из гипотез.

Замечание 4.4 *Необходимо внимательно подходить к выбору данных так как Ц.П.Т. выполняется не всегда. Дополнительно желательно с использованием критерия согласия проверить гипотезу о нормальности рассматриваемой статистики в случае каждой из гипотез.*

Замечание 4.5 *Для применения Ц.П.Т. хоть и не требуется вычисление вычисление явного вида распределения статистики отношения правдоподобия, тем не менее необходимо вычислить значения первых двух моментов логарифма статистики отношения правдоподобия в случае верности каждой из гипотез.*

Имея две нормально распределенные случайные величины с разными параметрами, задача вычисления ошибок первого/второго рода как и вычисление минимально необходимого количества материала для достижения нужных ошибок первого и второго рода решается достаточно легко.

Замечание 4.6 *Применение метода необходимо проиллюстрировать с использованием ЭВМ.*

Домашнее задание 5

Линейная регрессия и метод наименьших квадратов

Будем считать, что понятие линейной регрессии и метод наименьших квадратов знаком студентам. Теоретические вопросы связанные с применением метода наименьших квадратов и его оптимальности можно найти, например в [1].

5.1 Теоретическое введение

При обработке экспериментальных (или статистических) данных часто требуется проводить кривые заданного вида, проходящие поблизости от заданных точек.

Имеются результаты серии экспериментальных измерений

$$(x_1, y_1)\,,\ \ (x_2, y_2)\,,\ \ \ldots,\ \ (x_n, y_n)\,,$$

причем обе величины x и y измерены в одних и тех же экспериментах; известно, что ошибки измерений величины y – независимые нормально распределенные случайные величины с одинаковыми дисперсиями и нулевыми математическими ожиданиями, величина x измерена с пренебрежимо малой ошибкой (т.е. x – неслучайная величина).

Необходимо восстановить линейную регрессионную зависимость по результатам измерений.

Задача линейного регрессионного анализа состоит в восстановлении функциональной зависимости $y\,(x) = a_0 + a_1 x$ по результатам измерений

$$(x_1, y_1)\,,\ \ (x_2, y_2)\,,\ \ \ldots,\ \ (x_n, y_n)\,.$$

Уравнение (эмпирическая регрессия) $\widehat{y} = \widehat{a}_0 + \widehat{a}_1 x$, определяет прямую линию, которая является оценкой истинной линии регрессии. Необходимо вычислить точечные и интервальные оценки $\widehat{a}_0, \widehat{a}_1$ для параметров a_0, a_1 по

результатам эксперимента и проверить значимость полученного уравнения регрессии.

Коэффициенты регрессии можно вычислить различными способами:

▷ минимизируя сумму квадратов отклонений:

$$E\left(\widehat{a}_0, \widehat{a}_1\right) = \sum_{i=1}^{n} \left(\widehat{a}_0 + \widehat{a}_1 x_i - y_i\right)^2;$$

▷ численно решая систему уравнений:

$$\frac{\partial E\left(\widehat{a}_0, \widehat{a}_1\right)}{\partial \widehat{a}_i} = 0, \ i \in \overline{0,1};$$

▷ некоторыми другими способами.

5.2 Постановка задачи

Необходимо для выбранных данных (выбираются самостоятельно или предлагаются преподавателем) найти оценки a_0, a_1 минимизируя сумму квадратов отклонений.

При самостоятельном поиске данных рекомендуется вычислять коэффициент корреляции Пирсона, который характеризует существование линейной зависимости между двумя величинами.

Пусть даны две выборки $x^m = (x_1, \cdots, x_m)$, $y^m = (y_1, \cdots, y_m)$; коэффициент корреляции Пирсона рассчитывается по формуле:

$$r_{xy} = \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m} (x_i - \bar{x})^2 \sum_{i=1}^{m} (y_i - \bar{y})^2}} = \frac{cov(x,y)}{\sqrt{s_x^2 s_y^2}},$$

где $\bar{x}, \bar{y}$ – выборочные средние x^m и y^m, s_x^2, s_y^2 – выборочные дисперсии, $r_{xy} \in [-1, 1]$. При этом, $|r_{xy}| = 1 \ \Rightarrow \ x, y$ — линейно зависимы. Таким образом, предлагается выбирать данные, имеющие $r_{xy} \geq 0.7$.

Дополнительные необязательные задания

Точное распределение статистики D_n при $n \leq 20$

Критерий Колмогорова является асимптотическим и использование статистики Колмогорова возможна при объема данных $n \geq 20$. Помимо предельного результата Колмогоров в работе 1993 года предложены рекуррентные соотношения для конечных n. Стоит задача найти значение распределения статистики D_n при $n \leq 20$.

Об уточнении критерия Колмогорова-Смирнова

В работе [9] предложен способ уточнения критерия согласия Колмогорова-Смирнова. Стоит задача в оценке погрешности рассмотренных в работе статистик по сравнению с их неуточненными версиями в зависимости от уровня значимости.

Вычисление трудоемкости статистического метода анализа криптографического алгоритма и вероятности нахождения ключа

Вместо домашнего задания 5 возможно рассмотреть один из статистически методов анализа криптографических алгоритмов (линейный, разностный, корреляционный) и найти основные параметры метода криптографического анализа алгоритма.

Оценивание стоимости недвижимости

Вместо домашнего задания 5 возможно рассмотреть задачу оценки стоимости недвижимости с использованием леммы Неймана-Пирсона по материалам [10] (может найдете еще что).

Проверка гипотезы о равномерности распределения генератора случайных чисел игры DOOM

На сайте `https://github.com/id-Software/DOOM/blob/master/linuxdoom-1.10/m_random.c` опубликован исходный код генератора случайных чисел, используемого в играх DOOM и DOOM II. На какой длине выборки можно отличить случайную величину, выработанную этим генератором от истинно-случайной последовательности.

Правила оформления материалов

Итоговые материалы предоставляются в формате PDF. Так как большинство домашних заданий связано желательно (но не обязательно) все домашние задания представлять в одном файле.

Сам текст необходимо готовить одним из предложенных ниже способов:

- ▷ система компьютерной верстки TeX (LaTeX, XeLaTex и т.п.);
- ▷ Jupyter Notebook (можно сразу и код писать и текст с использованием команд LaTeX);
- ▷ текстовый процессор MS Word, Page, Libreoffice Writer **с использованием редакторов формул** (наименее предпочтительный вариант).

Замечание 5.1 *Любой написанный код необходимо включать в текст итогового отчета для того, чтобы можно было проверить правильность работы программы.*

Замечание 5.2 *Все написанные программы, используемые данные, исходники отчетов необходимо заархивировать и прислать архив при сдаче домашней контрольной работы. Можно также залить на GitHub, Bitbucket. Тогда достаточно прислать ссылку.*

Литература

[1] Ивченко Г.И. Медведев Ю.И. *Введение в математическую статистику.* УРСС, Москва, 2010.

[2] В.В. Некруткин. *Моделирование распределений.* СПбГУ, 2014. `http://statmod.ru/wiki/_media/books:vv:simulation_v4.pdf`.

[3] *Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Методические рекомендации. Часть I. Критерии типа* χ^2. ГОССТАНДАРТ РОССИИ, 2001.

[4] *Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Методические рекомендации. Часть II. Непараметрические критерии.* ГОССТАНДАРТ РОССИИ, 2001.

[5] С.Н. Постовалов и др. Б.Ю. Лемешко, С.Б. Лемешко. *Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход.* НИЦ ИНФРА-М, 2015. `https://ami.nstu.ru/~headrd/seminar/publik_html/Statistical_Data_Analysis.pdf`.

[6] Модели распределений статистик непараметрических критериев согласия при проверке сложных гипотез с использованием оценок максимального правдоподобия. Ч.i. 2009. `http://ami.nstu.ru/~headrd/seminar/publik_html/Models_Part_I.pdf`.

[7] Модели распределений статистик непараметрических критериев согласия при проверке сложных гипотез с использованием оценок максимального правдоподобия. Ч.ii. 2009. `http://ami.nstu.ru/~headrd/seminar/publik_html/Models_Part_II.pdf`.

[8] Стьюарт А. Кендалл М. Дж. *Статистические выводы и связи.* Наука, 1973.

[9] Л. Н. Большев. Асимптотически пирсоновские преобразования. *Теория вероятн. и ее примен.*, 8(2), 1963. `http://mi.mathnet.ru/tvp4657`.

[10] Marcus Berliant. A characterization of the demand for land. *Journal of Economic Theory*, 33(2):289 – 300, 1984.