Домашнее задание 1.

Характеристики вероятностных распределений

1. Описание основных характеристик распределения

Для каждого из выбранного распределения необходимо выписать его основные характеристики:

- ▷ функция распределения,
- ▷ математическое ожидание,
- ▷ дисперсия,
- ▷ квантиль уровня γ,

Все выписываемые характеристики должны сопровождаться теоретическими выкладками.

2. Поиск примеров событий, которые могут быть описаны выбранными случайными величинами

Для каждого из выбранных распределений необходимо

- ▷ привести пример интерпретации распределения – описания события, исходы в котором подчиняются выбранному распределению;
- ▷ известные соотношения между распределениями;

К интерпретациям будем относить математические модели, описываемые данным распределением. .

Пример 1.1 *Рассмотрим пример для распределения Пуассона.* ***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, $i = 1, \dots m$ 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.

Еще один пример ***интерпретации*** 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 $t > s$ $X_t - X_s$ $\Pi\left(\lambda(t-s)\right)$. 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

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

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

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

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

$$= \int\limits_0^x \frac{t^{n-1} e^{-\lambda t} \lambda^n}{(n-1)!}\, dt$$

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.

3. Описание способа моделирования выбранных случайных величин

Для каждого из двух распределений (дискретное и непрерывное) необходимо описать способ моделирования выборок с заданными распределениями.

Полагая, что у каждого есть источник непрерывных случайных величин, распределённых равномерно на отрезке $[0, 1]$ (random), необходимо описать и **обосновать** процедуру получения нужного распределения на основе равномерной выборки. Данное направление является хорошо освященным в литературе (см. например [1, 2]).

Замечание 1.2 *В отчете должен быть представлен код, с помощью которого производилось моделирование случайной величины или процедура получения выборки, описанная с помощью псевдокода.*

Домашнее задание 2.

Основные понятия математической статистики

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

1. Генерация выборок выбранных случайных величин.
2. Построение эмпирической функции распределения.
3. Построение гистограммы и полигона частот.
4. Вычисление выборочных моментов.

1. Генерация выборок выбранных случайных величин

Для каждой из выбранных случайных величин необходимо построить по 5 выборок следующих объемов $n = \{5, 10, 100, 200, 400, 600, 800, 1000\}$.

2. Построение эмпирической функции распределения

Для каждой сгенерированной выборки необходимо построить график эмпирической функции распределения

$$\mathscr{F}_n(t) = \frac{\sum\limits_{i=1}^{n} I\left(x_i < t\right)}{n}.$$

Графики необходимо привести в отчете. На одном графике необходимо отобразить эмпирические функции распределения для каждого из объемов выборки независимо и график функции распределения случайной величины.

Для каждой пары построенных эмпирических $\mathscr{F}_n(x)$ и $\mathscr{F}_m(x)$, $n, m \in \{5, 10, 100, 200, 400, 600, 800, 1000\}$ необходимо вычислить

$$D_{m,n} = \sqrt{\frac{nm}{m+n}} \sup_{x \in \mathbb{R}} \left|\mathscr{F}_n(x) - \mathscr{F}_m(x)\right|.$$

3. Построение гистограммы и полигона частот

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

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

Необходимо пояснить полученные графики. Какие теоремы из курса математической статистики они иллюстрируют?

4. Вычисление выборочных моментов

Для каждого сгенерированной выборки необходимо выписать значение выборочного среднего $\overline{X}$ и выборочной дисперсии $\overline{S}^2$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\overline{S}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

Какими свойствами данные оценки обладают?

Также необходимо сравнить значения полученных оценок с истинными значениями математического ожидания и дисперсии.

Домашнее задание 3.

Построение точечных оценок параметра распределения

Третье долгосрочное домашнее задание посвящено построению точечных оценок неизвестных параметров и исследованию их свойств.

Задание состоит из следующих пунктов:

1. Получение оценок методом моментов и методом максимального правдоподобия.

2. Поиск оптимальных оценок

3. Работа с реальными данными*.

1. Получение оценок методом моментов и методом максимального правдоподобия

Для каждого из распределений (дискретное и непрерывное) необходимо получить оценки неизветного параметра методом моментов и методом максимального правдоподобия.

Для каждой выборки, сгенерированной в пункте 2.1, необходимо привести значения полученных оценок.

2. Поиск оптимальных оценок

Для каждого из распределений (дискретное и непрерывное) необходимо предложить параметрическую функцию $\tau(\theta)$, для которой существует оптимальная оценка.

Замечание 3.1 *В случае, если параметрическая функция $\tau(\theta)$ не равна оцениваемому параметру θ, необходимо (если это возможно) также построить оптимальную оценку для θ.*

Для каждой выборки, сгенерированной в пункте 2.1, необходимо привести значения полученных оценок.

3. Работа с данными*

Данное задание является дополнительным и не входит в обязательную программу.

Для выбранного интерпретации, обоснованной в первой домашней работе найти данные, соответствующие 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.

В качестве источников данных можно пользоваться следующими сайтами или любыми другими найденными датасетами (в том числе собранными самими):

- ▷ `kaggle.com`
- ▷ `opendata.socrata.com`
- ▷ `https://github.com/awesomedata/awesome-public-datasets`
- ▷ `https://cloud.google.com/bigquery/public-data/`
- ▷ `http://archive.ics.uci.edu/ml/index.php`
- ▷ `https://www.data.gov`
- ▷ `https://academictorrents.com/browse.php`
- ▷ `https://www.quandl.com/search`
- ▷ `https://data.gov.ru`
- ▷ `https://data.mos.ru`
- ▷ `https://data.gov.spb.ru`

Домашнее задание 4.

Проверка статистических гипотез

1. Проверка гипотезы о виде распределения

Для каждой выборки, сгенерированной в пункте 2.1 необходимо рассмотреть следующие статистики:

- ▷ Критерий согласия Колмогорова (Смирнова),
- ▷ Критерий согласия хи-квадрат,
- ▷ Критерий согласия Колмогорова (Смирнова) для сложной гипотезы (в условиях когда неизвестен параметр распределения),
- ▷ Критерий согласия хи-квадрат для сложной гипотезы (в условиях когда неизвестен параметр распределения).

При известных параметрах распределений, проверка гипотез о виде распределения с использованием критерия согласия Колмогорова (или критерия Смирнова) и критерия согласия хи-квадрат происходит с использованием соответствующих статистик и их предельных распределений.

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

Для снижения требований к объему выборки можно вместо статистики D_n, для применения критерия Колмогорова, 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].

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]. Необходимо **применить критерий хи-квадрат** с различными вариантами группировки значений. Для каждой выборки, сгенерированной в пункте 2.1, выписать значение статистики Пирсона при различных вариантах разбиения и соответствующие им значения квантилей распределения хи-квадрат.

Следующей задачей является **проверка 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.

При проверки гипотезы о виде распределения с использованием критерия хи-квадрат – число степеней свободы статистики хи-квадрат, к которой стремится статистика Пирсона, снижается на m, где m – число оценивемых параметров распределения.

При проверке сложных гипотез с использованием 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].

2. Проверка гипотезы об однородности выборок

На основе построенных значений $D_{m,n}$ в пункте 2.2 сделать вывод об однородности сгенерированных выборок.

3. Задание для данных, описываемых распределение*

Данное задание является дополнительным и не входит в обязательную программу.

Как правило, при наличии данных имеется лишь предположение о виде распределения. В этом случае для данных необходимо проверить проверить критерии согласия для сложных гипотез.

Домашнее задание 5.

Различение статистических гипотез

Данное домашнее задание посвящено вопросу различения двух простых гипотез, а также закреплению основных понятий.

Задание состоит из следующих пунктов:

1. Описание критерия отношения правдоподобия
2. Вычисление функции отношения правдоподобия.
3. Вычисление критической области.
4. Вычисление минимального необходимого количества материала при фиксации минимального возможного значения ошибок первого и второго рода.

Необходимо ответить на вопросы:

- ▷ Что является гипотезой H_0, что H_1?
- ▷ Что такое ошибка первого и второго рода, функция мощности?

1. Вычисление функции отношения правдоподобия

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

2. Вычисление критической области

Рассмотрим один из самых сложных вопросов данной контрольной работы — 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

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

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

Замечание 5.1 *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.*

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

Замечание 5.2 *Применение метода необходимо проиллюстрировать с использованием ЭВМ.*

Дополнительные необязательные задания

Точное распределение статистики D_n при $n \le 20$

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

Об уточнении критерия Колмогорова-Смирнова

В работе [8] 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

Рассмотрение одного из статистически методов анализа криптографических алгоритмов (линейный, разностный, корреляционный) и нахождение его основных параметров.

Оценивание стоимости недвижимости

Рассмотреть задачу оценки стоимости недвижимости с использованием леммы Неймана-Пирсона по материалам [9] (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?

Может ли воспользоваться последовательным анализом Вальда?

О других критериях

Математическая статистика хоть и молодая, но уже достаточно развитая наука. В куре математической статистики мы рассмотрели лишь основные понятия. В частности, известно достаточно большое количество критериев не рассмотренных в нашем курсе (см. напр. [**?**]). Интересно воспользоваться иными статистиками, отличными от рассмотренных в нашем курсе. Какие они имеют преимущества и недостатки?

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

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