Министерство образования и науки Российской Федерации

Санкт-Петербургский Политехнический Университет Петра Великого

—

Институт кибербезопасности и защиты информации

КУРСОВАЯ РАБОТА

«Исследование распределения Гомперца»

по дисциплине «Теория вероятностей и математическая статистика»

Выполнил

студент гр. 4851003/10001 Богданова Л. А.

Ассистент преподавателя Штыркина А. А.

Санкт-Петербург

2023

**История**

Бенджамин Гомперц (1779–1865) – лондонский математик и актуарий Эта функция была впервые представлена ​​в его статье от 16 июня 1825 года Функция Гомперца свела значительный набор данных в таблицах дожития в единую функцию. Он основан на предположении, что уровень смертности экспоненциально увеличивается с возрастом человека. Результирующая функция Гомперца предназначена для количества людей, живущих в данном возрасте, в зависимости от возраста. Ранее работы по построению функциональных моделей смертности были выполнены французским математиком Абрахамом де Муавром (1667–1754) в 1750-х годах. Однако де Муавр предположил, что уровень смертности был постоянным. Расширение работы Гомперца было предложено английским актуарием и математиком Уильямом Мэтью Мейкхэмом (1826–1891) в 1860 году, который добавил постоянный фоновый уровень смертности к экспоненциально растущему уровню смертности Гомперца.

**Эмпирическая функция распределения**

Эмпирическая функция распределения для выборки задаётся формулой , где . Таким образом, значение эмпирической функции распределения для числа соответствует количеству элементов выборки, меньших данного числа.

Для того, чтобы построить эмпирическую функцию распределения, следует:

1. Определить количество элементов в выборке
2. Для каждого элемента выборки определить, какое количество элементов меньше данного
3. Осуществить деление полученных значений на размер выборки
4. Построить график функции, отметив на оси абсцисс элементы выборки, а на оси ординат – полученные значения для каждого из них

Используя созданную функцию на языке питон, построим графики эмпирических функций распределения для случайных подвыборок из исходной выборки.

Изображение выглядит как текст, снимок экрана, График, диаграмма

Автоматически созданное описание Изображение выглядит как текст, График, диаграмма, линия

Автоматически созданное описание Изображение выглядит как текст, диаграмма, График, линия

Автоматически созданное описание

Видим, что при увеличении количества элементов подвыборки график её эмпирической функции распределения становится более гладким. Для более точного сравнения формы функций отобразим их на одном графике.

Изображение выглядит как текст, диаграмма, снимок экрана, линия

Автоматически созданное описание

Линии на графике имеют схожую форму, из чего можем сделать вывод, что полученные случайным образом подвыборки являются достаточно репрезентативными.

Построим теоретическую функцию распределения Гомперца со случайными параметрами (b=0.09, loc=10), и сравним полученный график с эмпирической функцией распределения, построенной на основе полной выборки:

Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание

Видим, что форма графиков не совпадает. Попробуем подобрать более удачные параметры, например, b=0.9, loc=6:

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Формы графиков стали более схожими, из чего можно сделать вывод, что параметры распределения близки к выбранным. Но теоретическая функция с заданными аргументами принимает нулевое значение в точке x=0, а все элементы выборки лежат правее 2.4. Следовательно, функция имеет смещение. Попробуем построить график с учётом этого факта, взяв в качестве параметров b=1, loc=3:

Изображение выглядит как текст, линия, снимок экрана, диаграмма

Автоматически созданное описание

Видим довольно хорошо совпадающие графики. Наблюдаемое расхождение можно объяснить недостаточной репрезентативностью выборки и подбором параметров методом проб и ошибок. При вычислении параметров с помощью оценок предполагается получить более точное совпадение.

**Гистограмма**

Гистограмма - кусочно-постоянная функция , принимающая на интервале значение , где  — число элементов выборки, попавших в интервал , , ;

Для построения гистограммы выборки нужно произвести следующие действия:

1. Разбить область значений выборки на произвольные интервалы
2. Для каждого из них подсчитать количество элементов выборки, попавших в него
3. Осуществить деление на количество элементов выборки и длину интервала
4. Отметив на оси абсцисс выбранные границы интервалов, на оси ординат – полученные значения, построить столбчатую диаграмму

Сделаем подвыборки из основной выборки, содержащие различное количество элементов, и построим их гистограммы:

Изображение выглядит как текст, снимок экрана, дисплей, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

Построим график плотности распределения для теоретической функции распределения, заданной в варианте, со случайно выбранными аргументами (loc = 4, b = 3):

Изображение выглядит как текст, снимок экрана, дисплей, линия

Автоматически созданное описание

Видим, что, как и функция распределения, плотность вероятности смещена относительно нашей выборки. Учтя смещение и опытным путём подобрав параметры(loc = 3, b=1, смещение 2.5), построим график плотности распределения Гомперца и сравним его с гистограммой, построенной на основе всей выборки:

–

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание

Видим, что гистограмма примерно повторяет своей формой функцию распределения.

Исходя из примерного совпадения графиков, можем предположить, что истинные значения параметров могут приблизительно совпадать с использованными для построения графика плотности вероятности.

**Описание функции распределения**

Выбранный вариант задания подразумевает исследование функции распределения Гомперца с двумя параметрами – параметром формы (b) и параметром смещения (loc). Функция распределения Гомперца имеет вид: *F(x) = .*

*bce^(c+bx-ce^(bx))*

Параметр loc - коэффициент при возведении числа e в отрицательную степень. Соответственно, логично предположить, что в общем случае, чем меньше параметр loc, тем быстрее изменение функции и теб более “крутым” является её график.

Параметр b в общем случае увеличивает значение производной при возрастании, и, значит, его увеличение приводит к увеличению крутизны графика.

Построим график теоретической функции распределения, используя созданную функцию, аргументами которой являются параметры распределения. Аргументы выберем случайным образом.

Изображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описание

Рассмотрим, как в действительности изменение параметров влияет на вид функции распределения. Зафиксируем параметр loc, и построим графики функции распределения при различных значениях b.

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Видим, что, в соответствии с описанием параметра b, его увеличение приводит к увеличению «крутизны» функции (а, значит, и увеличению её производной), причём при малых значениях параметра (близких к 0) это изменение более значимо.

Теперь зафиксируем параметр b, проводя изменение параметра loc. Построим график функции распределения при различных значениях.

Изображение выглядит как текст, снимок экрана, линия, График

Автоматически созданное описание

Сравнивая полученные графики, можем заметить, что увеличение исследуемого параметра снижает «крутизну» графика и «отодвигает» точку, в которой рост функции практически прекращается.

**Точечные оценки**

Метод моментов

Начальный выборочный момент 1-го порядка: 2.692817603333332

Начальный теоретический момент 1-го порядка:

Указанный интеграл выражается через гамма-функцию, и итоговое значение для пределов интегрирования от 0 до разброса функции:

E =

Однозначно выразить параметр параметры из получившейся формулы невозможно, но можно использовать полученное уравнение для оценки истинных значений параметров методом подбора.

Полученные значения параметров, приводящие к наиболее полному соответствию: b=1.15005, loc=3.1675. При этом значение интеграла: 0,21164, что с высоким уровнем точности соответствует требуемому значению с учётом смещения.

**Метод наибольшего правдоподобия**

Оценка максимального правдоподобия параметра a равна

Где m число наблюдений (случаев).

Для получения оценки максимального правдоподобия параметра b необходимо решить другое уравнение. Запишем

Подставляя найденное значение b в формулу для , получим оценку .

То, что найденные оценки доставляют логарифму правдоподобия максимум, а не минимум доказывается с помощью того факта, что матрица, составленная из вторых частных производных, является отрицательно определённой.

Значение оценки параметра b было посчитано с помощью Excel, = -2.023, тогда =-2.613.

При выборке из 1 элемента: b=1,15, loc=3,02

Изображение выглядит как текст, снимок экрана, линия, диаграмма

Автоматически созданное описание

Несмещенность

Оценка параметра θ называется несмещенной, если ее математическое ожидание совпадает с самим оцениваемым параметром, т. е.

Все оценки параметров, полученные методом максимального правдоподобия, содержат в себе значение другого параметра, причем мы не можем выразит оценку второго параметра из доступных нам уравнений в явном виде. Также оценки, которые мы всё-таки можем выразить, не содержат выборочного среднего и дисперсии, а значения выборки содержатся в степени экспоненты, поэтому мы не можем оценить несмещенность полученных оценок.

Состоятельность

Оценка параметра θ называется состоятельной если она сходится по вероятности к параметру θ, т. е. если ∀ℇ > 0 выполняется:

Оценки, полученные методом моментов являются состоятельными по теореме.

В достаточно широких условиях оценки максимального правдоподобия оказываются состоятельными, причем практически без ограничения общности можно считать, что если для параметра θ в принципе существует состоятельная оценка, то оценка максимального правдоподобия этого параметра также является состоятельной.

Наиболее распространенной причиной несостоятельности оценки максимального правдоподобия является неоднозначность, связанная с тем, что максимум функции правдоподобия достигается для нескольких значений θ одновременно. Если это обстоятельство имеет место при каком угодно увеличении объема данных наблюдения, т. е. при n стремящемся к бесконечности, то очевидно, что оценка максимального правдоподобия не состоятельна.

**Интервальные оценки**

Оценим параметры распределения выборки с помощью интервальной оценки с уровнями доверия 𝛾1 = 0,95, 𝛾2 = 0,99.

Для первого параметра:

n = 300 (объем выборки), – выборочное среднее, - корень из несмещённой дисперсии.

Значение получили из таблицы значений функции Лапласа.

Получаем интервал с уровнем доверия 𝛾 = 0,95.

Значение получили из таблицы значений функции Лапласа.

Получаем интервал с уровнем доверия 𝛾 = 0,99.

**Понятие статистических критериев**

**Гипотезы о параметрах распределения:**

По теореме Неймана-Пирсона критерий отношения правдоподобия является наиболее мощным критерием в классе всех критериев проверки простой гипотезы.

Найдем отношение правдоподобия для параметра b

Теперь вычислим логарифм отношения правдоподобия:

мы должны принять гипотезу Н0, если ⩽ C

или, обозначим

Тогда мы должны принять гипотезу Н0, если , и отвергнуть в противном случае.

Тогда α =

Аналогично, 1-β =

Выдвинем гипотезы:

Подставим численные значения в отношение правдоподобия: Согласно асимптотическому свойству метода отношения правдоподобия при n→∞ .

Имеем распределение со степенью свободы 2 и уровнем значимости 0.05, тогда

=> верна гипотеза .

**Вычисление ошибок 1-го и 2-го рода.**

Уровень значимости критерия проверки H0 – вероятность ошибочно отвергнуть гипотезу H0, когда она верна.

– для непрерывного распределения, – плотность распределения наблюдаемой случайной величины при условии, что верна гипотеза H0.

Мощность критерия - вероятность правильно принять гипотезу H1.

– для непрерывного распределения, – плотность распределения наблюдаемой случайной величины при условии, что верна гипотеза H1.

Гипотезы о параметрах распределения

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Принята по результатам расчета | Вывод |
| Верна на самом деле |  |  | Вероятность правильного решения: 1 – α . |
|  | Ошибка первого рода (ложная тревога, ложное срабатывание). Вероятность: α . |
|  |  | Вероятность правильного решения β’ = 1 – β . |
|  | Ошибка второго рода (пропуск цели, пропуск события, ложноотрицательное срабатывание).  Вероятность ошибки: β . |

Для уровня значимости :

Отклонить : результаты вычислений предполагают, что .

Отказ отклонить : мы принимаем .

Ошибка I рода: с 5% риском ошибки мы примем , хотя действительный результат .

Ошибка II рода: так как у нас отсутствовали необходимые свидетельства, мы принимаем , хотя в действительности .

**Гипотезы о виде распределения.**

Используя критерий Колмогорова и критерий , проверим гипотезу о принадлежности выборки распределению из варианта.

**Критерий Колмогорова**

Критерий Колмогорова применяют, если исследуемая случайная величина принимает бесконечно много значений. По выборке (𝑥1, 𝑥2, …, 𝑥𝑛) с непрерывной теоретической функцией распределения 𝐹(𝑥) строим эмпирическую функцию распределения. Необходимо оценить, насколько далеко построенная эмпирическая функция распределения отклоняется от теоретической функции распределения. Расстояние: ρ = 𝑠𝑢𝑝||.

Критерий Колмогорова:

где 𝐹𝑛 (𝑥𝑖) – эмпирическая функция распределения, 𝐹(𝑥𝑖) - теоретическая функция распределения.

Наибольшее отклонение 𝜆 = 17,32 ∗ 0,21367207481690864 = 3,7

Для уровня значимости 𝛼 = 0,05 по таблице критических значений получим, что 𝜆кр = 1,36.

Так как полученное значение больше критического, мы вынуждены отвергнуть гипотезу H0, то есть критерий согласия Колмогорова опровергает нашу гипотезу о принадлежности рассматриваемой выборки к распределению Гомперца.

Критерий Пирсона (критерий ).

Критерий – статистический критерий для проверки гипотезы H0, что наблюдаемая случайная величина подчиняется некому теоретическому закону распределения. Он является одним из самых распространённых критериев гипотез. Данный критерий можно употреблять, когда все результаты испытаний можно разделить на конечное число категорий. Все возможные результаты 𝑛 независимых испытаний разделены на 𝑘 категорий. Обозначим вероятность того, что результат испытания попадает в 𝑖 − ую категорию, как 𝑝𝑖. А число испытаний, результаты которых попали в 𝑖 − ую категорию, обозначим 𝑛𝑖. Далее в общем случае требуется разбить область значений случайной величины 𝜉 на 𝑘 интервалов (−∞, 𝑑1 ), [𝑑1, 𝑑2), … , [𝑑𝑘+1, +∞), а затем определить теоретические вероятности 𝑝𝑖 = 𝐹(𝑑𝑖) − 𝐹(𝑑𝑖−1) попадания в интервал [𝑑𝑖−1, 𝑑𝑖) и числа 𝑛𝑖 элементов выборки, попавших в эти интервалы.

Разделим область значений случайной величины на k=10 интервалов. ] – количество наблюдений в j-м интервале;

= F() – F() – вероятность попадания наблюдения в j-ый интервал;

= – ожидаемое число попаданий в j-ый интервал (j = 1, 2, …, k);

Для уровня значимости α = 0.05, k = 10 количества рассматриваемых интервалов и m = 2 оцениваемых параметров:

(по таблице распределения Пирсона)

Значит, как и в случае критерия Колмогорова, мы вынуждены отвергнуть гипотезу H0, то есть критерий согласия опровергает нашу гипотезу о принадлежности рассматриваемой выборки к распределению Гомперца.

**Гипотезы об однородности выборок.**

Наиболее близкими по виду к распределению Гомперца является экспоненциальное распределение.

Метод обратного преобразования (Преобразование Н. В. Смирнова) — способ генерации случайных величин с заданной функцией распределения, путём модификации работы генератора равномерно распределённых чисел.

Пусть F(x) является функцией произвольного распределения. Покажем, как, имея генератор выборки из стандартного непрерывного равномерного распределения, получить выборку из распределения, задаваемого функцией распределения F(x)

Для этого требуется знать обратную функцию распределения.

Найдем обратную функцию к функции распределения Гомперца: F(x) = => F-1(x) =

Преобразование, позволяющее из равномерного распределения получить экспоненциальное распределение: , где х – значения выборки с равномерным распределением.

Преобразование, позволяющее из равномерного распределения получить экспоненциальное: Сгенерируем выборку с распределением Гомперца из равномерно распределенной выборки по представленной ранее формуле.

Критерий Смирнова (встречается также название «двухвыборочный критерий Колмогорова-Смирнова») применяется для проверки однородности двух выборок, т.е. гипотезы о том, что выборки взяты из одной и той же генеральной cовокупности. Основан на сопоставлении накопленных частот (эмпирических функций распределения), рассчитанных в двух выборках по всем значениям обоих выборок. Находят максимальную по модулю разность накопленных частот выборок и сравнивают её с критическим значением.

По критерию Смирнова находят D = max|W1(z)-W2(z)|

Здесь W1(z) и W2(z) - накопленные частости для первой и второй выборок по вариационному ряду всех значений обеих выборок. Если D < Dα , то гипотезу о тождественности выборок не отвергают, т.е. принимают на уровне значимости ⍺.

Критическое значение

λ⍺ при n+m > 35 - табличное значение критерия Колмогорова.

Для нахождения частностей двух выборок, посчитаем количество каждого элемента в каждой выборке. В выборке 1 самое часто повторяющееся число встречается 5 раз, а в выборке 2 самое редко попадающееся число встречается 1 раз. Для определения частностей, поделим частоты на объемы выборок, т.е. 1/300 = 0,003333333 и 5/300.

Теперь определим D. D = 7/300 – 1/300 = 0,013.

Подсчитаем Dα. Для уровня значимости α = 0.05, Dα = 1,358\* = 0,110880236.

Полученное значение D оказалось меньше критического значения Dα, значит принимает критерий Н0 об однородности выборок. Значит выборки принадлежат одной генеральной совокупности.