НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

Е.А. Буровский
Ю.Б. Гришунина

ЗАДАЧИ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ И ИХ РЕШЕНИЕ С ИСПОЛЬЗОВАНИЕМ ЯЗЫКА ПРОГРАММИРОВАНИЯ PYTHON

Учебное пособие

ИЗДАТЕЛЬСКИЙ ДОМ
ВЫСШЕЙ ШКОЛЫ ЭКОНОМИКИ

МОСКВА, 2022

УДК 519.2
ББК 22.172
Б91

Рецензенты:
доктор физико-математических наук, профессор кафедры
вычислительных методов факультета ВМК МГУ им. М.В. Ломоносова
Н.В. Соснин
PhD, доцент Департамента прикладной математики МИЭМ НИУ ВШЭ,
заведующий международной Лабораторией статистической
и вычислительной геномики НИУ ВШЭ
В.Л. Щур

Буровский, Е. А., Гришунина, Ю. Б. Задачи математической статис-
Б91 тики и их решение с использованием языка программирования Python: учеб. пособие [Текст] / Е. А. Буровский, Ю. Б. Гришунина; Нац. исслед. ун-т «Высшая школа экономики», Мос. ин-т электроники и математики им. А. Н. Тихонова. — М.: Изд. дом Высшей школы экономики, 2022. — 64 с. — 150 экз. — ISBN 978-5-7598-2682-8 (в обл.). — ISBN 978-5-7598-2483-1 (e-book).

Учебное пособие содержит теоретические сведения, касающиеся основных задач математической статистики: определения, формулировки теорем, необходимые формулы. Для каждой задачи приведена постановка, указан метод ее решения с подробными комментариями, перечислены функции и методы языка программирования Python, которые рекомендуется применять при выполнении вычислений, проведен анализ конкретных статистических данных.

Предназначено для студентов всех специальностей и направлений подготовки, изучающих дисциплину «Теория вероятностей и математическая статистика».

УДК 519.2
ББК 22.172

Опубликовано Издательским домом Высшей школы экономики
http://id.hse.ru

doi:10.17323/978-5-7598-2682-8

ISBN 978-5-7598-2682-8 (в обл.)
ISBN 978-5-7598-2483-1 (e-book)

Содержание

Введение

Данное учебное пособие по структуре и формату изложения материала основано на ранее опубликованном пособии «Основные задачи математической статистики и их решение с использованием приложения Microsoft Excel», написанном в 2013 году, когда для обработки данных широко применялись средства Microsoft Excel. Оба издания объединены общей концепцией — в них кратко изложены теоретические основы математической статистики и на конкретных примерах проиллюстрированы методы анализа данных с применением прикладного программного обеспечения. Хотя в настоящее время приложение Excel продолжает активно использоваться на практике, аналогичный функционал доступен также и при использовании средств большинства современных высокоуровневых систем: R, Matlab, Wolfram Mathematica и многих других. Отметим, что выбор программных средств при решении задач математической статистики не является принципиальным. Это, в первую очередь, вопрос удобства и личных предпочтений пользователя. Выбор экосистемы языка Python и соответствующих свободно распространяемых библиотек с открытым кодом представляется адекватным как при изучении теоретических основ статистической обработки данных, так и при, собственно, практическом ее использовании для решения актуальных задач анализа данных.

Задачи математической статистики можно условно разделить на три группы:

— непараметрические задачи;

— параметрические задачи;

— проверка статистических гипотез.

Непараметрические задачи имеют место в случаях, когда нет информации о виде распределения генеральной совокупности; тогда возникает проблема оценки функции распределения или плотности распределения.

Если известны вид распределения и область возможных значений параметров, и необходимо оценить неизвестные параметры распределения — это параметрические задачи; при этом можно поставить задачу точечного или интервального оценивания.

Проверка статистических гипотез — это проверка различных предположений о вероятностных закономерностях изучаемого явления или процесса на основе статистических данных, т.е. результатов наблюдений или экспериментов; эти задачи могут быть как непараметрическими (например, гипотезы о виде теоретического распределения, о независимости выборок), так и параметрическими (гипотезы о значениях параметров вероятностных моделей, о равенстве средних, дисперсий и т.д.).

Рассмотрим подробнее перечисленные задачи математической статистики на примере следующих заданий:

1. Смоделировать выборку объема n=30 из генеральной совокупности с заданной теоретической функцией распределения (в качестве примера выберем распределение Вейбулла или логнормальное распределение с заданными параметрами).
2. Построить график эмпирической функции распределения; построить гистограмму и полигон частот. Сравнить построенные графики с соответствующими графиками теоретической функции распределения и плотности распределения.
3. Построить точечную оценку неизвестного параметра заданного распределения:

а) методом моментов;

б) методом максимального правдоподобия.

Сравнить полученные оценки с истинным значением параметра.

4. Построить доверительный интервал надежности $1 - \gamma$ для неизвестного математического ожидания:

а) считая дисперсию известной;

б) считая дисперсию неизвестной.

Сравнить точность полученных интервалов.

5. Используя критерий Пирсона, на заданном уровне значимости γ проверить, согласуется ли гипотеза о виде теоретического распределения с представленной выборкой.

Основные цели выполнения заданий:

— усвоение терминологии, используемой в математической статистике;

— углубленная проработка определений, постановок задач математической статистики и методов их решения;

— закрепление знаний по следующим темам: случайные величины, способы их задания и числовые характеристики, предельные теоремы теории вероятностей и др.;

— получение опыта работы со статистическими данными, их обработки и анализа;

— приобретение умения интерпретировать полученные результаты;

— совершенствование навыков работы с релевантными библиотеками Python и экосистемы PyData.

В частности, в настоящем пособии используются библиотеки NumPy, SciPy, statsmodels и matplotlib.

Сведения о распределении Вейбулла и логнормальном распределении

Необходимые для выполнения заданий сведения о распределении Вейбулла и логнормальном распределении приведены в табл. 1.

Таблица 1

	Распределение Вейбулла $W(r,\lambda)$	Логнормальное распределение $LogN(\mu,\sigma)$
Плотность распределения	$\begin{cases}0, x \le 0\\ \frac{r}{\lambda^r} x^{r-1} e^{-\left(\frac{x}{\lambda}\right)^r}, x > 0\end{cases}$	$\begin{cases}0, x \le 0\\ \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(lnx-\mu)^2}{2\sigma^2}}, x > 0\end{cases}$
Функция распределения	$\begin{cases}0, x \le 0\\ 1 - e^{-\left(\frac{x}{\lambda}\right)^r}, x > 0\end{cases}$	$\begin{cases}0, x \le 0\\ \int_{-\infty}^{lnx} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(t-\mu)^2}{2\sigma^2}} dt, x > 0\end{cases}$
Математическое ожидание	$\lambda\Gamma\left(1+\frac{1}{r}\right)$, Γ — гамма-функция	$e^{\mu+\frac{\sigma^2}{2}}$
Дисперсия	$\lambda^2\left(\Gamma\left(1+\frac{2}{r}\right) - \Gamma^2\left(1+\frac{1}{r}\right)\right)$	$\left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}$

Распределения вероятностей в библиотеке SciPy

В библиотеке SciPy распределения вероятностей и статистические тесты содержатся в модуле stats. Для использования объектов, реализующих распределения вероятностей, их необходимо импортировать из пространства имен scipy.stats:

```
>>> from scipy.stats import norm
```

Здесь norm — объект, реализующий нормальное (гауссово) распределение. Полный список распределений доступен в

официальной документации, https://docs.scipy.org/doc/scipy/reference/stats.html.

Для вычисления плотности вероятностей заданного распределения используется метод `pdf` (pdf — probability density function), для вычисления функции распределения — метод `cdf` (cdf — cumulative density function). Таким образом, например, значение плотности стандартного нормального распределения *N(0,1)* в точке x дается вызовом `norm.pdf(x)`.

Распределения вероятностей библиотеки `scipy.stats` параметризуются в терминах семейств сдвиг-масштаб: методы `pdf`, `cdf` (и прочие) имеют следующие аргументы (которые рекомендуются передавать с использованием явных имен): параметры сдвига `loc` и масштаба `scale`. По умолчанию значения данных параметров равны: `loc = 0, scale = 1`. Явное задание значений данных параметров эквивалентно вызову с аргументом `(x – loc) / scale`. Например, для нормального распределения, $N(\mu,\sigma^2)$, параметр сдвига совпадает с математическим ожиданием μ, а параметр масштаба дает σ.

Для иллюстрации построим плотности нормального распределения с разными параметрами, см. рис. 1:

```
>>> import numpy as np
>>> import matplotlib.pyplot as plt
>>> xx = np.linspace(-2, 4, 101)
>>> plt.plot(xx, norm.pdf(xx, loc=1), label='loc=%s, scale=%s'%(1, 1))
>>> plt.plot(xx, norm.pdf(xx, loc=1, scale=0.5),
...          label='loc=%s, scale=%s'%(1, 0.5))
>>> plt.grid()
>>> plt.legend(loc='best')
```

Некоторые распределения имеют дополнительные параметры, которые называются shapes. Например, распределение Вейбулла (в библиотеке `scipy.stats` называемое `weibull_min`) имеет дополнительный параметр `c`. Параметр `c` дается аргументом `scale` методов объекта `weibull_min`.

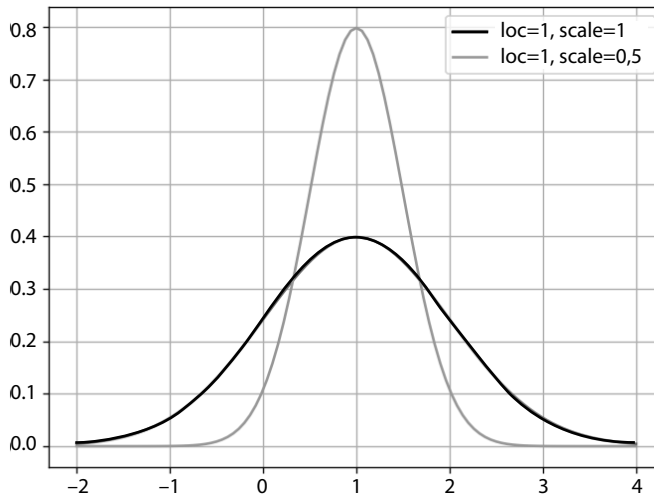

Рис. 1. Плотность нормального распределения для различных параметров масштаба

```
>>> import numpy as np
>>> from scipy.stats import weibull_min
>>> from numpy import exp
>>> r = 2; scale=1
>>> x = np.linspace(0, 3, 101)
>>> plt.plot(x, weibull_min.pdf(x, c=r, scale=scale))
>>> plt.plot(x, r * x**(r-1) * exp(-x**r) / scale,
...          lw=4, alpha=0.4)
```

Заметим, что параметризация в терминах семейства сдвиг-масштаб в некоторых случаях может отличаться от стандартной параметризации распределения. Например, для логнормального распределения (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.lognorm.html) параметризация сдвиг-масштаб отличается от стандартной параметризации (см. табл. 1) в терминах параметра μ, имеющего смысл математического ожидания величины $\exp(X)$. Для определения соответствия параметризаций заметим, что, взяв масштаб scale равным $exp(\mu)$, мы получим в точности плотность вероятности из табл. 1:

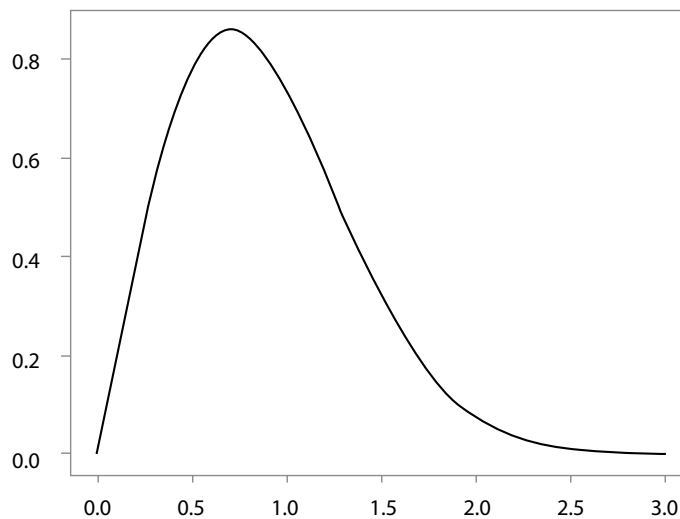

Рис. 2. Плотность распределения Вейбулла

```
>>> mu = 1.2
>>> scale = exp(mu)
>>> s = 2
```

```
>>> x = np.linspace(1, 1, 11)
>>> from numpy.testing import assert_allclose
>>> lpdf = exp(-(log(x) - mu)**2 / (2*s**2)) / x / s /
sqrt(2*pi)
>>> assert_allclose(lognorm.pdf(x, s=2, scale=scale),
...                 lpdf)
```

Более подробная информация о функционале библиотеки содержится в документации библиотеки (https://docs.scipy.org/doc/scipy/reference/tutorial/stats.html).

Задание 1.
Моделирование выборки из генеральной совокупности с заданной теоретической функцией распределения

Определение 1. Генеральной совокупностью называется вероятностное пространство $(\Omega, \mathcal{F}, P)$ *и определенная на этом пространстве случайная величина X.*

Случайную величину X, ее функцию распределения $F(x) = P(X < x)$, ее числовые характеристики и другие параметры будем называть теоретическими; все перечисленные элементы являются составными частями математической модели изучаемого явления или процесса. Таким образом, генеральная совокупность состоит из тех объектов, которые подлежат наблюдению и исследованию, и относительно них требуется сделать выводы при анализе конкретной проблемы.

Определение 2. Выборкой объема n называется последовательность n независимых одинаково распределенных случайных величин $\boldsymbol{X} = (X_1, \dots, X_n)$, *распределение каждой из которых совпадает с теоретическим распределением* $F(x)$.

Иными словами, выборка — это результат n последовательных независимых наблюдений над теоретической случайной величиной X.

Методика моделирования выборки основана на следующем *утверждении*:

Утверждение 1. Пусть X — непрерывная случайная величина, и ее функция распределения $F(x) = P(X < x)$ *монотонно возрастает. Тогда случайная величина* $Y = F(X)$ *имеет равномерное распределение на отрезке* $[0;1]$.

Доказательство. Напомним, что если случайная величина равномерно распределена на отрезке $[a; b]$, то ее функция распределения имеет следующий вид: $G(y) = \begin{cases} 0, y \le a \\ \frac{y-a}{b-a}, a < y \le b. \\ 1, y > b \end{cases}$

Очевидно, для отрезка [0; 1] $G(y) = \begin{cases} 0, y \le 0 \\ y, 0 < y \le 1. \\ 1, y > 1 \end{cases}$

Найдем функцию распределения случайной величины $Y = F(X)$.

При $y \le 0$ $G_Y(y) = P(Y < y) = P(F(X) < y) = 0$, поскольку $F(X)$ — это вероятность, и она не может принимать отрицательные значения. Аналогично, если $y > 1$, то $G_Y(y) = 1$, так как вероятность всегда ≤ 1.

Осталось рассмотреть случай $0 < y \le 1$. Заметим, что поскольку функция $F(x)$ непрерывна и монотонно возрастает, у нее существует обратная функция, которая также является монотонно возрастающей. Поэтому $G_Y(y) = P(Y < y) = P(F(X) < y) = P\left(F^{-1}\left(F(X)\right) < F^{-1}(y)\right) = P\left(X < F^{-1}(y)\right) = F\left(F^{-1}(y)\right) = y$, что и требовалось.

Из доказанного утверждения следует, что если $Y = F(X)$ — реализация случайной величины, имеющей равномерное распределение на отрезке [0; 1], то $X = F^{-1}(Y)$ — это реализация случайной величины, имеющей распределение $F(x)$. Поэтому, для того чтобы смоделировать выборку из генеральной совокупности с заданным теоретическим распределением, нужно сначала получить выборку $(Y_1, \dots, Y_n)$ из генеральной совокуп-

ности с равномерным распределением на отрезке [0; 1], а затем, подставляя полученные числа в формулу обратной функции, вычислить значения $(X_1, \dots, X_n)$, которые и будут реализациями случайной величины с заданным распределением $F(x)$.

Поскольку функции распределения случайных величин, имеющих распределение Вейбулла и логнормальное распределение, непрерывны и монотонно возрастают на интервале $(0; \infty)$, для моделирования соответствующих выборок можно использовать предложенный алгоритм.

Для распределения Вейбулла формула обратной функции легко выводится с помощью простых аналитических преобразований:

$Y_i = F(X_i) = 1 - e^{-\left(\frac{X_i}{\lambda}\right)^r}$, отсюда $X_i = \lambda\left(-ln(1 - Y_i)\right)^{\frac{1}{r}}$.

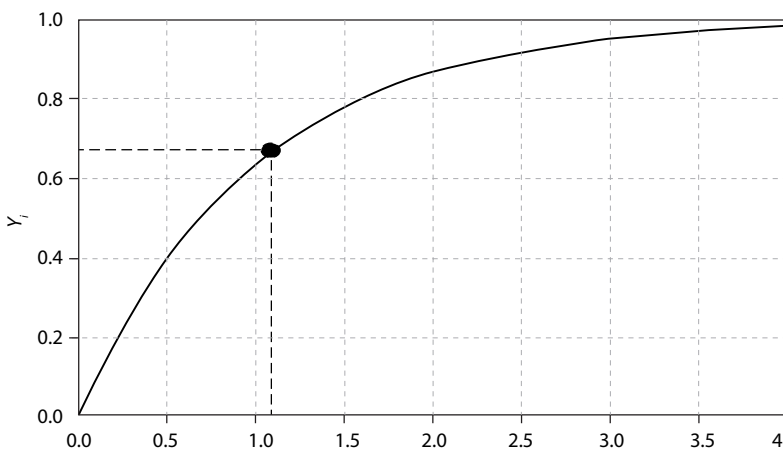

Рис. 3. Функция распределения Вейбулла

Рассмотрим реализацию предложенного алгоритма с использованием библиотек NumPy и SciPy. Для этого нам нужно создать выборку из равномерного распределения на отрезке [0; 1], а затем преобразовать значения согласно обратной функции распределения (см. рис. 3).

(Псевдо)случайные числа с равномерным распределением на отрезке [0; 1) генерируются с использованием библиотеки NumPy следующим образом. Сначала импортируем библиотеку и создадим генератор псевдослучайных чисел:

```
>>> import numpy as np
>>> rng = np.random.default_rng(seed=1234)
```

Здесь во второй строке создается объект (названный в данном примере `rng`) типа `np.random.Generator`, который реализует генерацию псевдослучайных последовательностей, «засеваемых» начальным значением, которое задается аргументом `seed`. Подчеркнем, методы библиотеки `numpy.random` генерируют именно псевдослучайные последовательности. Это означает, что последовательности на самом деле являются детерминированными (при идентичных начальных условиях последовательности будут идентичными), но с точки зрения статистических свойств слабо отличимыми от случайных.

Замечание 1. Аргумент `seed` можно опустить, но тогда при повторных вызовах, последовательности псевдослучайных чисел будут другими. Для воспроизводимости рекомендуется генератор «засевать» явно, как это сделано в примере выше. Более детальное описание функционала генерации псевдослучайных последовательностей представлено в документа-

ции библиотеки NumPy (https://numpy.org/doc/stable/reference/random/index.html).

Собственно генерация псевдослучайных последовательностей осуществляется вызовом соответствующего метода объекта `rng` (аргумент `size` задает желаемый размер выборки):

```
>>> rng.uniform(size=4)
array([0.17748 , 0.18344591, 0.42800468, 0.83810427])
```

Важно отметить, что генерируемые таким способом последовательности **не являются криптостойкими**.

Таким образом, генерация выборки из распределения Вейбулла с параметрами, например, r =1 и λ = 1 методом обратной функции может быть проведена следующим образом (ср. с формулой для обратной функции):

```
>>> Y = rng.uniform(size=10)
>>> X = -np.log(1 - Y).
```

Обобщение на случай других значений параметров не составляет труда.

Замечание 2. Если $r = 1$, то распределение Вейбулла — это экспоненциальное распределение с параметром $\frac{1}{\lambda}$ $\left(Exp\left(\frac{1}{\lambda}\right)\right)$.

Для алгоритма, реализующего моделирование выборки с произвольным распределением, генерация выборки с заданными параметрами производится вызовом метода `.rvs(...)`. Для примера сгенерируем выборку из распределения Вейбулла двумя способами: «вручную» и с использованием библиотечного метода:

```
>>> N = 30
# seed the generator for reproducibility
>>> rng = np.random.default_rng(1235)
>>> X_weib = weibull_min.rvs(c=1, size=30, random_state=rng)
# reseed the generator
>>> rng = np.random.default_rng(1235)
>>> Y = rng.uniform(size=30)
>>> X_inv = -np.log(1. - Y)
```

Заметим, что в этом примере мы явно «засеваем» генераторы, благодаря чему выборки совпадают:

```
>>> np.allclose(X_inv, X_weib, atol=1e-15)
True
```

Отметим также, что генерацию выборок можно осуществлять и без использования библиотеки SciPy, на «чистом» NumPy. Так, например, вызов scipy.stats.weibull_min.rvs (c = 1, size = 30) в точности эквивалентен вызову rng.weibull (a=1,size=30) (https://numpy.org/doc/stable/reference/random/generated/numpy.random.weibull.html).

Для удобства дальнейшей работы построим так называемый *вариационный ряд*, отсортировав выборку по возрастанию:

```
>>> X = X_weib.copy()
>>> X.sort()
>>> X
array([0.03137465, 0.04531734, 0.06544178, 0.12135278, 0.1791247 ,
       0.19538249, 0.20266212, 0.30067117, 0.31516295, 0.31756893,
       0.34691501, 0.34702768, 0.41694743, 0.43383668, 0.50926666,
```

```
0.53299031,  0.53517583,  0.55862447,  0.88995173,  0.89984881,
1.04866232,  1.07977963,  1.13848816,  1.17957629,  1.33374016,
1.33785406, 1.35656268, 1.7545202 , 1.82080282, 1.89182818])
```

Задание 2.
Эмпирическая функция распределения. Гистограмма и полигон частот

Эмпирическая функция распределения

Теоретическая функция распределения $F(x) = P(X < x)$ определяет вероятность события $\{X < x\}$. Согласно статистическому определению вероятности и закону больших чисел относительная частота появления события в n независимых испытаниях, т.е. доля тех испытаний, в которых данное событие произошло, практически не отличается от его вероятности, поэтому она является оценкой для вероятности события. Под оценкой понимается некоторая функция, зависящая от результатов наблюдений, значение которой мало отличается от истинного значения величины, которое требуется оценить. Поэтому оценкой для функции распределения является относительная частота события $\{X < x\}$ в n испытаниях, которая называется эмпирической функцией распределения и вычисляется как отношение числа испытаний, в которых произошло событие $\{X < x\}$, к общему числу испытаний n. По *определению 2* каждое выборочное значение X_i является реализацией теоретической случайной величины X, поэтому наступление события $\{X < x\}$ в i-м испытании означает, что $X_i \in \in (-\infty; x)$, а число испытаний, в которых произошло событие $\{X < x\}$, равно количеству выборочных значений, меньших x. Таким образом, эмпирическая функция распределения определяется следующим образом:

Определение 3. Эмпирической функцией распределения $\hat{F}_n(x)$ *называется функция* $\hat{F}_n(x)\frac{\nu(x)}{n}$, *где* $\nu(x)$ *— число точек вариационного ряда, меньших x, т.е.* $\nu(x) = \sum_{i=1}^{n} I\,(X_i < x)$, I *— индикатор,* $I(X_i < x) = \begin{cases} 0, \text{если } X_i \geq x \\ 1, \text{если } X_i < x \end{cases}$; n *— объем выборки.*

Из этого определения следует, что если в выборке нет повторяющихся значений, то эмпирическая функция распределения — это ступенчатая функция со скачками, равными по величине $\frac{1}{n}$, в точках вариационного ряда. Такая функция распределения соответствует распределению дискретной случайной величины, которая принимает каждое из значений $X_1, \ldots, X_n$ с вероятностью $\frac{1}{n}$. Отметим также, что эмпирическая функция распределения обладает всеми свойствами функции распределения.

Замечание 3. Если некоторое значение повторяется в выборке k раз, то скачок эмпирической функции распределения в соответствующей точке вариационного ряда равен $\frac{k}{n}$.

Построение эмпирической функции распределения

Проверим, что в выборке нет повторяющихся значений:

```
# check equal samples
>>> np.any(X[1:] == X[:-1])
False
```

Библиотека `statsmodels` (https://www.statsmodels.org/stable/index.html) предоставляет возможность сконструировать эмпирическую функцию распределения (ECDF — empirical cumulative distribution function) по выборке:

```
>>> from statsmodels.distributions import ECDF
>>> ecdf = ECDF(X)
```

Результирующий объект может быть вычислен в заданной точке x.

Продемонстрируем также альтернативный вариант построения эмпирической функции распределения «вручную», используя библиотеку `scipy.interpolate`:

```
>>> xx = np.r_[0, X]
>>> yy = np.repeat(1 / X.size, X.size)
>>> yy = np.r_[0, yy.cumsum()]
>>> from scipy.interpolate import interp1d
>>> ec = interp1d(xx, yy,
...               kind='previous', bounds_error=False)
```

И построим графики результатов вместе с исходным вариационным рядом (рис. 4):

```
>>> x = np.sort(np.r_[0, X, X-1e-10, X+1e-10])
>>> plt.plot(x, ec(x),
...          '-', label='ECDF, manual')
>>> plt.plot(x, ecdf(x),
...          '--', label='ECDF, statsmodels')
>>> plt.plot(X, yy[1:],
...          '.', label='data', ms=8)
>>> plt.grid(ls='--')
>>> plt.legend()
```

Замечание 4. Если в выборке есть повторяющиеся значения, то процедуру построения эмпирической функции распределения следует соответствующим способом модифицировать.

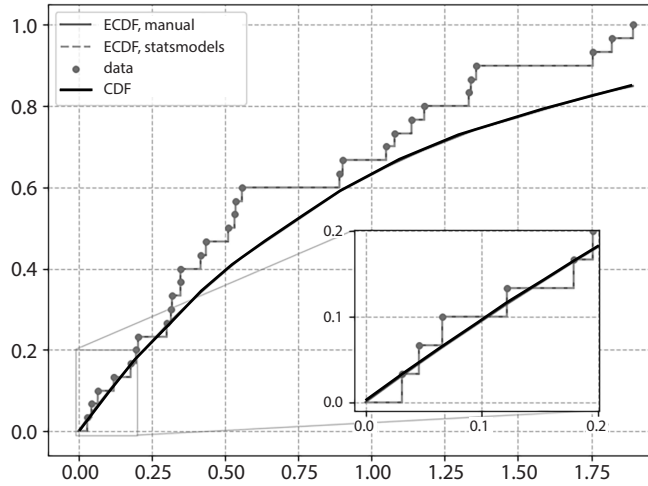

Рис. 4. Вариационный ряд, эмпирическая и теоретическая функции распределения для выборки из распределения Вейбулла

Замечание 5. Полученный график — это график эмпирической функции распределения на отрезке $[X_1^*; X_n^*]$; X_1^* и X_n^* — соответственно минимальное и максимальное выборочные значения. При $x \leq X_1^*$ $\hat{F}_n(x) = 0$, а при $x > X_n^*$ $\hat{F}_n(x) = 1$.

Гистограмма и полигон частот

Для наглядности представления результатов наблюдений и удобства восприятия иногда бывает полезно построить другие виды графиков статистического распределения, в частности, **гистограмму** и **полигон частот**.

Построение гистограммы и полигона основано на группировке статистических данных. Для этого отрезок $[a; b]$, содержащий все выборочные значения, т.е. $a \leq X_1^*$ и $b \geq X_n^*$, разбивается на m непересекающихся интервалов длины Δ:

$(z_0; z_1], (z_1; z_2], \dots, (z_{m-1}; z_m]$, где $z_0 = a, z_m = b, z_k - z_{k-1} = \Delta$, $k = 1, \dots, m$; рекомендуемое количество интервалов вычисляется по формуле Стерджесса $m = 1 + 1{,}41 \ln n$, а длина каждого интервала $\Delta = \frac{b-a}{m}$; в частности, для выборки объема n = = 30 рекомендуемое количество интервалов m = 6. Затем по вариационному ряду подсчитывается число выборочных значений, попавших в каждый интервал; *число наблюдений, попавших в k-й интервал* $(z_{k-1}; z_k]$, *называется эмпирической частотой и обозначается* n_k; *величина* $W_k = \frac{n_k}{n}$ *называется относительной эмпирической частотой.*

Определение 4. Гистограммой относительных частот называется функция $\hat{f}_n(x) = \begin{cases} 0, x \notin [a; b] \\ \frac{W_k}{\Delta}, x \in (z_{k-1}; z_k], k = 1, \dots, m \end{cases}$.

Геометрически гистограмму можно представить как фигуру на плоскости, состоящую из прямоугольников, основаниями которых (т.е. горизонтальными сторонами) являются интервалы $(z_0; z_1], (z_1; z_2], \dots, (z_{m-1}; z_m]$, а высоты (вертикальные стороны) равны соответственно $\frac{W_k}{\Delta}$. Площадь k-го прямоугольника равна $\frac{W_k}{\Delta}\Delta = W_k$, а сумма площадей всех прямоугольников, составляющих гистограмму, равна $\sum_{k=1}^{m} W_k =$ $= \frac{1}{n}\sum_{k=1}^{m} n_k = \frac{n}{n} = 1$. Отметим, что это свойство гистограммы аналогично свойству плотности распределения $\int_{-\infty}^{\infty} f(x)dx =$ $= 1$, так как геометрический смысл этого интеграла — площадь фигуры, ограниченной осью абсцисс и графиком плотности $f(x)$. Поэтому гистограмму разумно считать оценкой для теоретической плотности распределения.

Для построения полигона вычисляются середины интервалов группировки: $z_k^* = \frac{z_{k-1}+z_k}{2} = z_{k-1} + \frac{\Delta}{2}, k = 1, \dots, m$.

Определение 5. Полигоном относительных частот называется ломаная, соединяющая точки $\left(z_1^*; \frac{W_1}{\Delta}\right), \left(z_2^*; \frac{W_2}{\Delta}\right), \ldots, \left(z_m^*; \frac{W_m}{\Delta}\right)$.

Построение гистограммы и полигона

Построение гистограммы по вариационному ряду можно выполнить, используя методы библиотек matplotlib (`plt.hist`) либо NumPy (`np.histogram`), см. рис. 5. Основное отличие в том, что в первом случае результат сразу строится на графике, тогда как вызов `np.histogram` только возвращает вычисленные значения. В обоих случаях важно указать параметр нормировки `density=True`. В примере ниже мы используем автоматический выбор числа интервалов для заданной выборки (обе библиотеки предоставляют возможность более детального контроля разбиения). Отметим, что параметр `bins='auto'` необходимо задавать явно, так как значение по умолчанию использует фиксированное количество интервалов, что может оказаться неадекватно объему выборки.

```
# the histogram
>>> plt.hist(X, bins='auto', density=True,
...          alpha=0.3, label='histogram')
# the pdf itself
>>> xx = np.linspace(X.min(), X.max(), 201)
>>> plt.plot(xx, weibull_min.pdf(xx, c=1), '-', label='pdf')
# frequency polygon
>>> hist, bin_edges = np.histogram(X,
...                                density='True', bins='auto')
>>> mid = (bin_edges[1:] + bin_edges[:-1]) / 2
```

```
>>> plt.plot(mid, hist, 'o-', label='freq polygon', lw=3)
>>> plt.grid(True)
>>> plt.legend(loc='best')
```

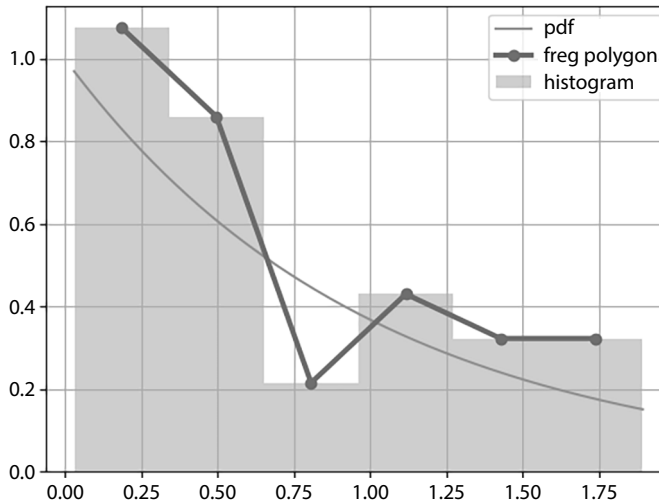

Рис. 5. Гистограмма и полигон частот для выборки из распределения Вейбулла

Задание 3.
Точечные оценки неизвестных параметров

Постановка задачи

Пусть в результате наблюдений получена выборка $X_1, \dots,$ X_n из генеральной совокупности с теоретическим распределением $F(x)$, и относительно функции $F(x)$ известно только, что она принадлежит определенному параметрическому семейству $F(x, \theta)$, где $\theta = (\theta_1, \dots, \theta_k)$ — вектор параметров, т.е. вид ее известен, но неизвестны параметры, определяющие это распределение. Естественно, возникает задача оценки этих неизвестных параметров по выборке.

Назовем оценкой (статистической оценкой) $\hat{\theta}$ некоторую функцию от выборочных значений $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$, предназначенную для статистического оценивания неизвестных параметров распределения. Если вычисление этой функции при подстановке в нее конкретных результатов наблюдений приводит к одному значению параметра (число при $k = 1$ или вектор при $k > 1$), то такая оценка называется точечной.

Задача состоит в выборе такой функции $\hat{\theta}$, которая в определенном смысле мало отличалась бы от истинного значения оцениваемого параметра. Желательными свойствами такой оценки являются:

- несмещенность;
- состоятельность;
- эффективность.

Определение 6. Оценка $\hat{\theta}$ *называется несмещенной, если ее математическое ожидание равно оцениваемому параметру:* $E\hat{\theta} = \theta$.

Выполнение этого свойства гарантирует, что оценка не будет давать систематического отклонения результата.

Определение 7. Оценка $\hat{\theta}$ *называется состоятельной, если она сходится по вероятности к оцениваемому параметру:* $\hat{\theta} \xrightarrow{P} \theta$, *т.е. для любого* $\varepsilon > 0$ $P\left(\left|\hat{\theta} - \theta\right| < \varepsilon\right) \to 1$ *при* $n \to \infty$.

Это свойство означает, что при достаточном объеме выборки оценка с вероятностью, близкой к 1, будет мало отличаться от истинного значения оцениваемого параметра.

Определение 8. Оценка $\hat{\theta}$ *называется эффективной, если она имеет минимальную дисперсию в определенном классе оценок.*

Поскольку дисперсия характеризует разброс значений случайной величины вокруг математического ожидания, для несмещенной эффективной оценки разброс значений $\hat{\theta}$ вокруг оцениваемого параметра θ будет наименьшим.

Существуют различные методы получения точечных оценок. Рассмотрим два из них — метод моментов и метод максимального правдоподобия.

Метод моментов

Метод моментов основан на том, что с ростом числа наблюдений эмпирическая функция распределения мало отличается от теоретической, поэтому мало отличаются и соответствующие числовые характеристики, что позволяет, в частности, считать приблизительно равными теоретические и эмпирические моменты одинаковых порядков.

Напомним, что моментом (начальным моментом) порядка N случайной величины X называется математическое ожидание N-й степени этой случайной величины: $\mu_N = EX^N$. Поскольку теоретическое распределение зависит от вектора параметров $\theta = (\theta_1, \dots, \theta_k)$, то и теоретические моменты также являются функциями этих параметров: для дискретной теоретической случайной величины X $\mu_N = \mu_N(\theta) = \sum_j x_j^N P_j(\theta)$, где $P_j(\theta)$ — закон распределения случайной величины X; для непрерывной теоретической случайной величины X $\mu_N = \mu_N(\theta) = \int_{-\infty}^{+\infty} x^N f(x, \theta) dx$, где $f(x, \theta)$ — плотность распределения случайной величины X.

Из *определения 3* следует, что эмпирическая функция распределения соответствует распределению дискретной случайной величины, которая принимает каждое из выборочных значений $X_1, \dots, X_n$ с вероятностью $\frac{1}{n}$, поэтому эмпирический момент порядка N вычисляется по формуле: $M_N = \frac{1}{n}\sum_{i=1}^{n} X_i^N$.

Метод моментов состоит в приравнивании теоретических и эмпирических моментов соответствующих порядков. Полученные при этом равенства образуют систему уравнений относительно параметров $\theta_1, \dots, \theta_k$; число уравнений в этой системе должно совпадать с числом неизвестных параметров:

$$\begin{cases} \mu_1(\theta_1, \dots, \theta_k) = M_1 \\ \mu_2(\theta_1, \dots, \theta_k) = M_2 \\ \dots \\ \mu_k(\theta_1, \dots, \theta_k) = M_k \end{cases}.$$

Решая эту систему, получаем точечную оценку $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$ для вектора неизвестных параметров $\theta = (\theta_1, \dots, \theta_k)$.

В частности, если k = 1, т.е. требуется оценить один параметр, достаточно выписать и решить одно уравнение относительно этого параметра. Приравнивая теоретический и эмпирический моменты первого порядка, получаем: $\mu_1(\theta) = M_1$. Заметим, что $\mu_1(\theta) = EX$ — это теоретическое математическое ожидание, а $M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_{\text{в}}$ — выборочное среднее, т.е. среднее арифметическое выборочных значений.

Построим с помощью метода моментов точечные оценки неизвестных параметров распределения Вейбулла и логнормального распределения.

Распределение Вейбулла $\boldsymbol{W(r, \lambda)}$

В задании 3 требуется построить точечную оценку параметра λ, считая параметр r известным, в нашем примере он равен 1. Приравняем теоретическое математическое ожидание и выборочное среднее: $\lambda\Gamma\left(1+\frac{1}{r}\right) = \overline{X}_{\text{в}}$. Решая это уравнение, получаем оценку параметра λ: $\hat{\lambda} = \frac{\overline{X}_{\text{в}}}{\Gamma\left(1+\frac{1}{r}\right)}$. При $r = 1$ $\hat{\lambda} = \overline{X}_{\text{в}}$.

Для вычисления $\hat{\lambda}$ воспользуемся библиотекой scipy.special, предоставляющей значения гамма-функции:

```
>>> from scipy.special import gamma    # NOT scipy.stats.gamma!
>>> scale  = X.sum() / X.size    # выборочное среднее
>>> scale /= gamma(2)
```

Для выборки $\boldsymbol{X}$ из распределения Вейбулла с параметрами c = 1, loc = 0 результат получается равным $\hat{\lambda}$ =1,07, что близко к истинному значению λ = 1.

Логнормальное распределение $LogN(\mu, \sigma)$

В задании 3 требуется построить точечную оценку параметра μ, считая параметр σ известным. Приравняем теоретическое математическое ожидание и выборочное среднее: $e^{\mu+\frac{\sigma^2}{2}} = \overline{X}_{\text{в}}$. Решая это уравнение, получаем оценку параметра μ: $\hat{\mu} = ln\left(\overline{X}_{\text{в}}\right) - \frac{\sigma^2}{2}$.

Метод максимального правдоподобия

При нахождении точечной оценки для неизвестных параметров $\theta = (\theta_1, \dots, \theta_k)$ распределения $F(x, \theta)$ естественно попытаться выбрать такое значение $\hat{\theta}$, которое лучше всего соответствовало бы полученной выборке $(X_1, \dots, X_n)$. Это означает, что при $\theta = \hat{\theta}$ вероятность (в дискретном случае) или плотность вероятности (в непрерывном случае) реализации данной выборки $L(\theta) = L(\theta, X_1, \dots, X_n)$ будет максимальной. Если теоретическая случайная величина X является дискретной, то вероятность реализации выборки $(X_1, \dots, X_n)$ — это вероятность совместного осуществления событий $\{X = X_i\}, i = 1, \dots, n$; если X — непрерывна, то соответствующая плотность вероятности — это совместная плотность распределения в точках $X_1, \dots, X_n$. По *определению 2* случайные величины X_i независимы и одинаково распределены, поэтому их совместное распределение $L(\theta, X_1, \dots, X_n)$ вычисляется как произведение вероятностей в дискретном случае или как произведение плотностей в непрерывном случае. Функция $L(\theta, X_1, \dots, X_n)$ была названа Фишером функцией правдоподобия.

Определение 9. Функция

$L(\theta, X_1, \dots, X_n) = P(X_1, \theta)P(X_2, \theta) \dots P(X_n, \theta) = \prod_{i=1}^{n} P(X_i, \theta)$,

где $P(X_i, \theta) = P(X = X_i)$*, в дискретном случае и*

$L(\theta, X_1, \dots, X_n) = f(X_1, \theta)f(X_2, \theta) \dots f(X_n, \theta) = \prod_{i=1}^{n} f(X_i, \theta)$,

где $f(x, \theta) = F'(x, \theta)$ *в непрерывном случае называется функцией правдоподобия.*

Определение 10. Оценкой максимального правдоподобия называется такое значение $\hat{\theta}$*, для которого* $L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta)$*, где* Θ *— замкнутая область допустимых значений параметров.*

Иногда на практике бывает удобнее пользоваться не самой функцией правдоподобия, а ее логарифмом. Функция $l(\theta) = lnL(\theta)$ называется логарифмической функцией правдоподобия; очевидно, точки максимума у функций $L(\theta)$ и $l(\theta)$ совпадают.

Если максимум функции $l(\theta)$ достигается внутри области Θ, то в точке максимума выполняются необходимые условия экстремума:

$\frac{\partial l(\theta)}{\partial \theta_j} = 0, j = 1, \dots, k.$

Полученные уравнения называются уравнениями правдоподобия. Решения системы уравнений правдоподобия могут быть точками максимума или минимума функции $l(\theta)$, а могут и не являться для нее точками экстремума, поэтому необходимо проверять, что полученное решение действительно является точкой максимума. Если система уравнений правдоподобия не имеет решений внутри области Θ, это означает, что максимум функции $l(\theta)$ достигается на границе области допустимых значений Θ.

Построим с помощью метода максимального правдоподобия точечные оценки неизвестных параметров распределения Вейбулла и логнормального распределения.

Распределение Вейбулла $\boldsymbol{W(r, \lambda)}$

В задании 3 требуется построить точечную оценку параметра λ, считая параметр r известным. Таким образом, вектор неизвестных параметров имеет размерность $k = 1$ и $\theta = \lambda$. По *определению 9* функция правдоподобия имеет вид:

$$L(\lambda, X_1, \dots, X_n) = \prod_{i=1}^{n} f(X_i, \lambda) = \frac{r^n}{\lambda^{rn}} \left(\prod_{i=1}^{n} X_i\right)^{r-1} e^{-\left(\frac{1}{\lambda}\right)^r \sum_{i=1}^{n} X_i^r};$$

тогда логарифмическая функция правдоподобия равна:
$l(\lambda) = n\, lnr - rn\, ln\lambda + (r-1)\sum_{i=1}^{n} lnX_i - \frac{1}{\lambda^r}\sum_{i=1}^{n} X_i^r$.

Дифференцируя по λ, получаем уравнение правдоподобия:
$-rn\frac{1}{\lambda} + r\frac{1}{\lambda^{r+1}}\sum_{i=1}^{n} X_i^r = 0$.

Легко проверить, что решение этого уравнения $\hat{\lambda} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^r\right)^{\frac{1}{r}}$ является точкой максимума для функции $l(\lambda)$, поэтому $\hat{\lambda} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^r\right)^{\frac{1}{r}}$ — оценка максимального правдоподобия.

В нашем примере $r = 1$, тогда $\hat{\lambda} = \overline{X}_{\text{в}}$, и значение оценки максимального правдоподобия совпадает со значением оценки, полученной методом моментов. В общем случае оценки, вообще говоря, не совпадают.

Для произвольного распределения точечные оценки максимального правдоподобия получаются использованием метода `.fit`:

```
>>> c, loc, scale = weibull_min.fit(X)
>>> print(c, loc, scale)
(0.8818760565106942, 0.003137203338172782, 0.9857005277979247)
```

Отметим, что по умолчанию данный метод подбирает оценки для всех параметров распределения, включая сдвиг и масштаб. Однако можно зафиксировать некоторые из переменных, передав методу .fit аргументы, начинающиеся с буквы f. Например, зафиксируем значения сдвига равным нулю и параметра с равным единице:

```
>>> c, loc,scale = weibull_min.fit(X, floc=0, fc=1)
>>> print(c, loc, scale)
(1, 0, 1.0735137411304754)
```

Видно, что результат для масштаба совпадает с вычисленным «руками».

Логнормальное распределение $\boldsymbol{LogN(\mu, \sigma)}$

В задании 3 требуется построить точечную оценку параметра μ, считая параметр σ известным. Поэтому вектор неизвестных параметров имеет размерность k =1 и $\theta = \mu$. По *определению 9* функция правдоподобия имеет вид:

$L(\mu, X_1, \dots, X_n) = \prod_{i=1}^{n} f(X_i, \mu) = \frac{1}{\prod_{i=1}^{n} X_i \sigma^n (2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(lnX_i - \mu)^2}$;

тогда логарифмическая функция правдоподобия равна:

$l(\mu) = -nln\sigma - \frac{n}{2} ln(2\pi) - \sum_{i=1}^{n} lnX_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(lnX_i - \mu)^2$.

Дифференцируя по μ, получаем уравнение правдоподобия:

$$\frac{1}{\sigma^2}\left(\sum_{i=1}^{n} lnX_i - n\mu\right) = 0.$$

Легко проверить, что решение этого уравнения $\hat{\mu} = \frac{\sum_{i=1}^{n} lnX_i}{n}$ является точкой максимума для функции $l(\mu)$, поэтому $\hat{\mu} = \frac{\sum_{i=1}^{n} lnX_i}{n}$ — оценка максимального правдоподобия.

Задание 4.
Доверительные интервалы

Постановка задачи

При рассмотрении методов получения точечных оценок было показано, что точечная оценка не совпадает с оцениваемым параметром; при малом объеме выборки такая оценка может значительно отличаться от истинного значения параметра. Поэтому разумно было бы указывать те допустимые границы, в которых может находиться неизвестный параметр θ при условии реализации выборки $(X_1, \dots, X_n)$, т.е. возникает задача интервального оценивания. Доверительный интервал — это статистическая оценка параметра вероятностного распределения, имеющая вид интервала, который с заданной вероятностью «накрывает» неизвестное значение параметра, а его границы являются функциями от результатов наблюдений.

Итак, пусть в результате наблюдений получена выборка $(X_1, \dots, X_n)$ из генеральной совокупности с теоретическим распределением $F(x, \theta)$, зависящим от числового параметра θ, $\theta \in \Theta \subseteq R$, значение которого неизвестно; α, $0 < \alpha < 1$ — фиксированное число.

Определение 11. Интервал $I(X_1, \dots, X_n) = \left(\hat{\theta}_1; \hat{\theta}_2\right)$ *с границами* $\hat{\theta}_1(X_1, \dots, X_n)$ *и* $\hat{\theta}_2(X_1, \dots, X_n)$, $\hat{\theta}_1 < \hat{\theta}_2$, *такой, что* $\inf_{\theta \in \Theta} P\left(\hat{\theta}_1 < \theta < \hat{\theta}_2\right) = 1 - \alpha$, *называется доверительным интервалом надежности* $1 - \alpha$ *для неизвестного параметра* θ.

Задача состоит в том, чтобы по выборке $(X_1, \dots, X_n)$ при заданном уровне надежности $1 - \alpha$ найти функции $\hat{\theta}_1(X_1, \dots, X_n)$ и $\hat{\theta}_2(X_1, \dots, X_n)$.

Из *определения 11* следует, что интервальные оценки позволяют установить точность и надежность точечных оценок.

Действительно, пусть $\hat{\theta}$ — точечная оценка неизвестного параметра θ. Точность этой оценки определяется отклонением $|\hat{\theta} - \theta|$; если $\varepsilon > 0$ и выполняется неравенство $|\hat{\theta} - \theta| < \varepsilon$, то чем меньше ε, тем точнее оценка. Таким образом, положительное число ε характеризует точность оценки.

Поскольку оценка $\hat{\theta} = \hat{\theta}_n(X_1, \dots, X_n)$ является функцией от случайных величин, нельзя наверняка утверждать, что оценка $\hat{\theta}$ удовлетворяет неравенству $|\hat{\theta} - \theta| < \varepsilon$; можно говорить только о вероятности, с которой это неравенство выполняется; эта вероятность называется надежностью (или доверительной вероятностью) оценки $\hat{\theta}$.

Пусть $P(|\hat{\theta} - \theta| < \varepsilon) = 1 - \alpha$. Неравенство $|\hat{\theta} - \theta| < \varepsilon$ равносильно неравенству $\hat{\theta} - \varepsilon < \theta < \hat{\theta} + \varepsilon$, поэтому $P(\hat{\theta} - \varepsilon < \theta < \hat{\theta} + \varepsilon) = 1 - \alpha$, т.е. интервал $(\hat{\theta} - \varepsilon; \hat{\theta} + \varepsilon)$ является доверительным интервалом надежности $1 - \alpha$ для неизвестного параметра θ. Таким образом, интервальная оценка определяет надежность точечной оценки.

Число α называется уровнем значимости. В математической статистике обычно задаются следующие значения уровня значимости: α = 0,1; 0,05; 0,01.

Доверительный интервал для неизвестного математического ожидания при известной дисперсии в случае нормального распределения генеральной совокупности

Пусть $(X_1, \dots, X_n)$ — выборка из генеральной совокупности с нормальным распределением $N(\theta, \sigma)$, θ — неизвестное математическое ожидание; дисперсия σ^2 предполагается известной. Построим доверительный интервал для θ при заданном уровне значимости α.

Точечную оценку $\hat{\theta}$ для неизвестного математического ожидания θ несложно получить, например, с помощью метода моментов, приравнивая теоретический и эмпирический моменты первого порядка: $\hat{\theta} = \overline{X}_{\text{в}}$. Теперь, чтобы построить доверительный интервал заданной надежности $1-\alpha$, надо найти такое число ε, чтобы $P\left(\left|\overline{X}_{\text{в}} - \theta\right| < \varepsilon\right) = 1-\alpha$.

По *определению 2* случайные величины $X_1, \ldots, X_n$ независимы, и распределение каждой из них совпадает с теоретическим распределением: $X_i \sim N(\theta, \sigma)$; соответственно, $EX_i = \theta$, $DX_i = \sigma^2$. Напомним, что сумма независимых нормально распределенных случайных величин снова имеет нормальное распределение, а линейное преобразование $aX + b$ при $a > 0$ не меняет вид распределения, поэтому случайная величина $\overline{X}_{\text{в}} - \theta = \frac{1}{n}\sum_{i=1}^{n} X_i - \theta$ также имеет нормальное распределение. Найдем параметры этого распределения. Используя свойства математического ожидания и дисперсии, получаем:

$$E\left(\overline{X}_{\text{в}} - \theta\right) = \frac{1}{n}\sum_{i=1}^{n} EX_i - \theta = \frac{n\theta}{n} - \theta = 0;$$

$$D\left(\overline{X}_{\text{в}} - \theta\right) = D\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} DX_i = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Таким образом, $\overline{X}_{\text{в}} - \theta \sim N\left(0, \frac{\sigma}{\sqrt{n}}\right)$.

Разделим эту случайную величину на корень из дисперсии (напомним, что такая процедура называется нормированием случайной величины; при этом вид распределения не меняется, а дисперсия нормированной случайной величины равна 1); тогда случайная величина $Y_0 = \frac{\sqrt{n}\left(\overline{X}_{\text{в}} - \theta\right)}{\sigma}$ имеет стандартное нормальное распределение: $Y_0 \sim N(0,1)$. Следовательно,

$$P\left(\left|\overline{X}_{\text{в}} - \theta\right| < \varepsilon\right) = P\left(\frac{\sqrt{n}\left|\overline{X}_{\text{в}} - \theta\right|}{\sigma} < \frac{\sqrt{n}\varepsilon}{\sigma}\right) =$$

$$= P\left(|Y_0| < \frac{\sqrt{n}\varepsilon}{\sigma}\right) = P\left(-\frac{\sqrt{n}\varepsilon}{\sigma} < Y_0 < \frac{\sqrt{n}\varepsilon}{\sigma}\right) =$$

$$= \int_{-\infty}^{\frac{\sqrt{n}\varepsilon}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx - \int_{-\infty}^{-\frac{\sqrt{n}\varepsilon}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx =$$

$$= 2\int_{-\infty}^{\frac{\sqrt{n}\varepsilon}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx - 1 = 2F_0\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) - 1 = 1 - \alpha,$$

где $F_0(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx$ — функция распределения стандартного нормального распределения $N(0,1)$.

Получаем, что $F_0\left(\frac{\sqrt{n}\varepsilon}{\sigma}\right) = 1 - \frac{\alpha}{2}$.

Напомним, что решение x_p уравнения $F(x_p) = p$, где $F(x)$ — некоторое заданное распределение, называется квантилью уровня вероятности p распределения F. Тогда $\frac{\sqrt{n}\varepsilon}{\sigma} = d_{1-\frac{\alpha}{2}}$, где $d_{1-\frac{\alpha}{2}}$ — квантиль уровня вероятности $1 - \frac{\alpha}{2}$ стандартного нормального распределения, т.е. решение уравнения $\int_{-\infty}^{d_{1-\frac{\alpha}{2}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = 1 - \frac{\alpha}{2}$. Отсюда следует, что $\varepsilon = \frac{\sigma d_{1-\frac{\alpha}{2}}}{\sqrt{n}}$, а интервал $\left(\overline{X}_{\text{в}} - \frac{\sigma d_{1-\frac{\alpha}{2}}}{\sqrt{n}}; \overline{X}_{\text{в}} + \frac{\sigma d_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right)$ является доверительным интервалом заданной надежности $1 - \alpha$ для неизвестного математического ожидания θ.

Таким образом, с заданной вероятностью $1 - \alpha$ выполняется неравенство:

$$\overline{X}_{\text{в}} - \frac{\sigma d_{1-\frac{\alpha}{2}}}{\sqrt{n}} < \theta < \overline{X}_{\text{в}} + \frac{\sigma d_{1-\frac{\alpha}{2}}}{\sqrt{n}}.$$

Точность полученной интервальной оценки равна $\varepsilon = \frac{\sigma d_{1-\frac{\alpha}{2}}}{\sqrt{n}}$, а границы имеют вид: $\hat{\theta}_1 = \overline{X}_{\text{в}} - \frac{\sigma d_{1-\frac{\alpha}{2}}}{\sqrt{n}}$; $\hat{\theta}_2 = \overline{X}_{\text{в}} + \frac{\sigma d_{1-\frac{\alpha}{2}}}{\sqrt{n}}$.

Доверительный интервал для неизвестного математического ожидания в случае произвольного теоретического распределения генеральной совокупности

Пусть теперь $(X_1, \dots, X_n)$ — выборка из генеральной совокупности с произвольным теоретическим распределением $F(x, \theta)$, отличным от нормального; θ — неизвестное математическое ожидание; дисперсия σ^2 предполагается известной.

В этом случае точечной оценкой $\hat{\theta}$ для неизвестного математического ожидания θ, как и при нормальном распределении генеральной совокупности, является выборочное среднее: $\hat{\theta} = \overline{X}_{\text{в}}$. Это следует, например, из метода моментов: независимо от вида теоретического распределения первый теоретический момент — это всегда математическое ожидание, а эмпирический — выборочное среднее. По *определению 2* случайные величины $X_1, \dots, X_n$ независимы и одинаково распределены, поэтому и $EX_i = \theta, DX_i = \sigma^2,\ i = 1, \dots, n$. Согласно центральной предельной теореме, если независимые одинаково распределенные случайные величины $X_1, \dots, X_n$ имеют конечную дисперсию, то каков бы ни был их закон распределения, сумма $\sum_{i=1}^{n} X_i$ при достаточно больших n $(n \geq 30)$ имеет распределение, близкое к нормальному. Поскольку линейное преобразование случайной величины X $aX + b$ при $a > 0$ не меняет вид распределения, при достаточно большом объеме вы-

борки можно считать, что случайная величина $\overline{X}_{\text{в}} - \theta = \frac{1}{n}\sum_{i=1}^{n} X_i - \theta$ имеет нормальное распределение с теми же параметрами, что и в случае нормального распределения генеральной совокупности, так как общие свойства математического ожидания и дисперсии не зависят от вида функции распределения, и эти числовые характеристики вычисляются аналогично.

Таким образом, снова $\overline{X}_{\text{в}} - \theta \sim N\left(0, \frac{\sigma}{\sqrt{n}}\right)$. Поэтому все рассуждения, проведенные при построении доверительного интервала для неизвестного математического ожидания в случае нормального распределения генеральной совокупности, остаются справедливыми и для генеральной совокупности с произвольным теоретическим распределением. Соответственно, и доверительный интервал заданной надежности $1 - \alpha$ для неизвестного математического ожидания имеет тот же самый вид: $\left(\overline{X}_{\text{в}} - \frac{\sigma d_{1-\frac{\alpha}{2}}}{\sqrt{n}}; \overline{X}_{\text{в}} + \frac{\sigma d_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right)$.

Если дисперсия σ^2 неизвестна, то можно использовать несмещенную оценку для дисперсии, полученную по выборке; напомним, что такая оценка называется исправленной выборочной дисперсией и вычисляется по формуле: $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_{\text{в}}\right)^2$. При достаточно большом объеме выборки оценку $\hat{\sigma}^2$ можно считать приблизительно равной истинному значению σ^2, так как с ростом числа наблюдений эмпирическая функция распределения мало отличается от теоретической, поэтому и соответствующие числовые характеристики отличаются незначительно. Заменяя σ на $\hat{\sigma}$, получаем доверительный интервал заданной надежности $1 - \alpha$ для не-

известного математического ожидания генеральной совокупности с произвольным теоретическим распределением при неизвестной дисперсии:
$\left(\overline{X}_{\text{в}} - \frac{\hat{\sigma} d_{1-\frac{\alpha}{2}}}{\sqrt{n}}; \overline{X}_{\text{в}} + \frac{\hat{\sigma} d_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right)$, где $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_{\text{в}}\right)^2}$.

Отметим, что замена σ на $\hat{\sigma}$ оправдана при достаточно большом объеме выборки; для малых выборок такая замена может привести к значительным ошибкам.

Вычисление границ доверительных интервалов

При известной дисперсии границы доверительного интервала для неизвестного математического ожидания вычисляются следующим образом: $\hat{\theta}_1 = \overline{X}_{\text{в}} - \frac{\sigma d_{1-\frac{\alpha}{2}}}{\sqrt{n}}$; $\hat{\theta}_2 = \overline{X}_{\text{в}} + \frac{\sigma d_{1-\frac{\alpha}{2}}}{\sqrt{n}}$. В наших примерах значение дисперсии явно не задано, и его необходимо вычислить по формулам, представленным в табл. 1.

Также необходимо вычислить квантиль нормального распределения. Для распределений, реализованных в библиотеке scipy.stats, это производится следующим образом. Для произвольной функции распределения $F(x)$ квантиль x_p на уровне p является решением уравнения $F(x_p) = p$. В scipy.stats функция распределения называется cdf (англ. cumulative density function), а ее дополнение — sf (англ. survival function), так что cdf(x) + sf(x) = 1.

Функция, обратная к cdf, — т.е. дающая решение уравнения выше, $x_p = F^{-1}(p)$ — обозначается ppf (англ. percent point function), обратная к sf обозначается isf (англ. inverse survival function).

Таким образом, для нахождения квантиля на уровне $1 - \alpha/2$ необходимо решить уравнение $\mathrm{cdf}(x) = 1 - a/2$, что эквивалентно $\mathrm{sf}(x) = a/2$, и решение дается вызовом $\mathrm{isf}(a/2)$.

Итак, для распределения Вейбулла $W(1,1)$, считая дисперсию известной, на уровне значимости 0,99 ($\alpha = 0{,}01$) получаем следующие доверительные интервалы:

```
>>> alpha = 0.01
>>> from scipy.stats import norm
>>> from scipy.special import gamma
>>> quantile = norm.isf(alpha/2.)
>>> c = 1
>>> sigma = sqrt(gamma(1 + 2./c) - gamma(1 + 1./c)**2)
# sample mean
>>> X_av = X.sum() / X.size
# confidence interval
>>> delta = quantile * sigma / sqrt(X.size)
>>> print("quantile =", quantile, " (alpha =", alpha, ")")
>>> print("W(1, 1) mean = ", weibull_min.mean(c=1))
>>> print("sample mean = ", X_av)
>>> print("conf interval : (%s, %s)" %
...       (X_av - delta, X_av + delta))
```

Получаем следующий результат вычисления:

```
quantile = 2.575829303548901  (alpha = 0.01 )
W(1, 1) mean =  1.0
sample mean =  1.073521178040133
conf interval : (0.6032412400946148, 1.5438011159856513)
```

Если дисперсия неизвестна, то границы доверительного интервала для неизвестного математического ожидания вычисляются по формулам: $\hat{\theta}_1 = \overline{X}_в - \frac{\hat{\sigma} d_{1-\frac{\alpha}{2}}}{\sqrt{n}}$; $\hat{\theta}_2 = \overline{X}_в + \frac{\hat{\sigma} d_{1-\frac{\alpha}{2}}}{\sqrt{n}}$, где $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_в\right)^2}$.

Для вычисления оценки стандартного отклонения используем библиотечную функцию, в которой указываем параметр ddof = 1 для получения несмещенной оценки. Отметим, что значения данного параметра по умолчанию в различных функциях библиотек NumPy и SciPy не являются согласованными, и всегда следует проверять данное значение по документации.

```
# ddof=1 для несмещенной оценки
>>> sigma = np.std(X, ddof=1)
>>> delta = quantile*sigma / sqrt(X.size)
>>> print("quantile =", quantile, " (alpha =", alpha, ")")
>>> print("W(1, 1) mean = ", weibull_min.mean(c=1))
>>> print("sample mean = ", X_av)
>>> print("conf interval : (%s, %s)" % (X_av - delta, X_av + delta))
```

В результате получаем оценки, которые несколько отличаются от оценок при известной дисперсии:

```
quantile = 2.575829303548901  (alpha = 0.01 )
W(1, 1) mean =  1.0
sample mean =  1.073521178040133
conf interval : (0.5598379714978815, 1.5872043845823844)
```

По представленным результатам можно сделать следующие выводы:

1) оценка $\hat{\sigma}$ незначительно отличается от истинного значения σ;

2) точность доверительного интервала при известной дисперсии выше, чем при неизвестной; это можно объяснить тем, что в случае известной дисперсии имеется больше информации о генеральной совокупности;

3) построенные доверительные интервалы накрывают истинное значение математического ожидания; напомним, что для распределения Вейбулла $W(1,1)$ $EX = 1$.

Задание 5.
Проверка статистических гипотез

Постановка задачи

Статистической гипотезой называют предположение о вероятностных закономерностях, которым подчиняется изучаемое случайное явление или процесс. Как правило, статистическая гипотеза — это предположение о виде неизвестного теоретического распределения и его свойствах или о значениях параметров известных распределений.

Гипотеза называется *простой*, если она содержит только одно предположение, т.е. определяет единственное распределение или единственную точку из области возможных значений параметров. *Сложной* называется гипотеза, которая состоит из конечного или бесконечного числа простых гипотез.

Одну из гипотез выделяют в качестве *основной (нулевой)* и обозначают H_0, а другую — в качестве *альтернативной (конкурирующей)* и обозначают H_1; конкурирующая гипотеза противоречит нулевой.

Выдвинутая основная гипотеза может быть правильной или неправильной, поэтому возникает необходимость ее проверки. Задача проверки статистических гипотез состоит в том, чтобы на основе выборки $(X_1, \dots, X_n)$ принять (т.е. считать справедливой) либо основную гипотезу H_0, либо конкурирующую H_1.

Общая схема проверки гипотез

Для проверки гипотезы формулируется правило, в соответствии с которым принимается или отклоняется основная гипотеза.

Это правило определяется выбором подходящей функции $K = K(X_1, \dots, X_n)$ от результатов наблюдений, которая служит мерой расхождения между опытными (выборочными) и гипотетическими (теоретическими) значениями. При этом предполагается, что функция K, зависящая от выборочных значений, является случайной величиной, распределение которой при правильной нулевой гипотезе точно или приближенно известно. Такая специально подобранная случайная величина K называется *критерием.* Значение критерия, вычисленное по результатам наблюдений, называется *наблюдаемым значением критерия*; обозначим его $K_{\text{набл}}$.

Все множество возможных значений критерия K разбивается на две непересекающиеся области: *допустимую* и *критическую.*

Допустимая область D — это совокупность значений критерия, при которых нулевая гипотеза принимается. Если наблюдаемое значение критерия $K_{\text{набл}}$ попадает в допустимую область, то считается, что результаты наблюдений не противоречат нулевой гипотезе H_0, и она принимается.

Критическая область S — это совокупность значений критерия, при которых нулевая гипотеза отвергается. Если наблюдаемое значение критерия $K_{\text{набл}}$ попадает в критическую область, то расхождение между гипотетическими и опытными данными считается значимым, основная гипотеза H_0 отвергается, и принимается конкурирующая гипотеза H_1.

Критическими точками $k_{\text{кр}}$ называются точки, отделяющие критическую область S от допустимой области D. Поскольку критерий K — это случайная величина, все его возможные значения принадлежат некоторому интервалу. Соответственно,

критическая область S и допустимая область D также являются интервалами, границы которых и называются критическими точками.

Значение критерия зависит от результатов наблюдений, которые являются случайными величинами, поэтому выводы, сделанные при проверке гипотез на основе статистических данных, могут оказаться ошибочными. При этом возможны ошибки двух родов.

Ошибка первого рода состоит в том, что отвергается правильная нулевая гипотеза. В этом случае нулевая гипотеза верна, но значение критерия попадет в критическую область, и принимается конкурирующая гипотеза H_1; вероятность ошибки первого рода α называется уровнем значимости критерия: $\alpha = P(K \in S|H_0)$.

Ошибка второго рода состоит в том, что принимается неправильная нулевая гипотеза. В этом случае нулевая гипотеза неверна, но значение критерия попадет в допустимую область, и нулевая гипотеза H_0 принимается; вероятность ошибки второго рода обозначим β: $\beta = P(K \in D|H_1)$.

Мощностью критерия называется вероятность принятия конкурирующей гипотезы H_1, если она верна; это происходит в случае, когда при правильной конкурирующей гипотезе значение критерия попадает в критическую область *S.* При этом $P(K \in S|H_1) = 1 - P(K \in D|H_1) = 1 - \beta$ — мощность критерия.

При построении допустимой области желательно было бы выбрать ее границы таким образом, чтобы при проверке гипотез как можно реже происходили ошибки как первого, так и второго рода. Но одновременно минимизировать вероятности α и β невозможно, потому что для уменьшения вероятности ошибки первого

рода α необходимо расширять границы допустимой области D, и, наоборот, для уменьшения вероятности ошибки второго рода β (т.е. увеличения мощности критерия) необходимо расширять границы критической области S. Поэтому обычно поступают следующим образом: фиксируют уровень значимости α, как более важный с практической точки зрения, а затем при заданном уровне α выбирают критерий, имеющий наибольшую мощность $1-\beta$.

Построение допустимой области заключается в выборе таких критических точек $k_{\text{кр}}^1$ и $k_{\text{кр}}^2$ (возможно, $k_{\text{кр}}^1 = -\infty$ или $k_{\text{кр}}^2 = +\infty$), чтобы при заданном уровне значимости α вероятность попадания критерия в допустимую область (т.е. вероятность принять нулевую гипотезу, если она верна) была равна $1-\alpha$: $P\left(k_{\text{кр}}^1 < K < k_{\text{кр}}^2 | H_0\right) = 1-\alpha$. Очевидно, выбор допустимой области не является однозначным, так как при заданном уровне α и известном распределении критерия K существует сколько угодно интервалов, удовлетворяющих этому условию. Поэтому критические точки по возможности выбираются таким образом, чтобы вероятность ошибки второго рода $\beta = P(k_{\text{кр}}^1 < K < k_{\text{кр}}^2 | H_1)$ была минимальной, т.е. выбирается наиболее мощный критерий. Данная задача, как правило, тоже не имеет однозначного решения, особенно в случае сложной конкурирующей гипотезы. Однако в ряде случаев удается найти так называемые равномерно наиболее мощные критерии, например, для параметрических гипотез конструктивный способ построения наиболее мощного критерия дает лемма Неймана — Пирсона [Ивченко, Медведев, 2010].

В зависимости от вида конкурирующей гипотезы можно построить правостороннюю, левостороннюю или двустороннюю критическую область.

Если $k_{\text{кр}}^1 = -\infty$, то критическая область S — это интервал $[k_{\text{кр}}; +\infty)$, границу которого находят из уравнения $P(K \geq k_{\text{кр}}) = \alpha$; такая критическая область называется правосторонней.

Если $k_{\text{кр}}^2 = +\infty$, то критическая область S — это интервал $(-\infty; k_{\text{кр}}]$, границу которого находят из уравнения $P(K \leq k_{\text{кр}}) = \alpha$; такая критическая область называется левосторонней.

Двусторонняя критическая область — это объединение двух интервалов: $S = (-\infty; k_{\text{кр}}^1] \cup [k_{\text{кр}}^2; +\infty)$. При этом $P(K \leq k_{\text{кр}}^1) + P(K \geq k_{\text{кр}}^2) = \alpha$; часто выбирается симметричная критическая область, для которой $P(K \leq k_{\text{кр}}^1) = P(K \geq k_{\text{кр}}^2) = \frac{\alpha}{2}$.

Таким образом, общая схема проверки гипотез сводится к следующим этапам:

1) задается уровень значимости α;
2) выбирается критерий для проверки нулевой гипотезы;
3) определяются границы критической области;
4) по выборочным данным вычисляется наблюдаемое значение критерия;
5) если значение $K_{\text{набл}}$ попадает в критическую область, то нулевая гипотеза H_0 отвергается и принимается конкурирующая гипотеза H_1; если $K_{\text{набл}}$ попадает в допустимую область, то нулевая гипотеза H_0 принимается.

Отметим, что даже если нулевая гипотеза H_0 принимается, это не является доказательством того, что она верна, потому что принятие гипотезы происходит на некотором фиксированном уровне надежности и основывается на случайных результатах наблюдений. Принятие гипотезы H_0 означает только, что на выбранном уровне надежности эта гипотеза не противоречит полученным выборочным данным.

Замечание 6. Альтернативный способ проверки гипотез связан с вычислением величины *P-value,* которую можно интерпретировать как условную вероятность получить более «нетипичное» или «экстремальное» значение критерия по сравнению с наблюдаемым при условии, что справедлива нулевая гипотеза H_0. Эта величина характеризует вероятность отвергнуть правильную нулевую гипотезу на основе имеющихся данных, соответственно, чем выше значение *P-value,* тем меньше оснований отклонить нулевую гипотезу, т.е. данная величина показывает степень соответствия полученных результатов нулевой гипотезе. Способ вычисления *P-value* определяется распределением критерия, который используется для проверки гипотезы, и видом критической области. Для правосторонней критической области *P-value* = $P(K \geq K_{\text{набл}} | H_0)$, для левосторонней *P-value* = $P(K < K_{\text{набл}} | H_0)$, для двусторонней *P-value* = $= 2min\big(P(K \geq K_{\text{набл}} | H_0), P(K < K_{\text{набл}} | H_0)\big)$. Решение о принятии или отклонении нулевой гипотезы принимается следующим образом: полученное значение *P-value* сравнивается с заданным уровнем значимости α; если *P-value* $< \alpha$, то нулевая гипотеза H_0 отвергается и принимается конкурирующая гипотеза H_1; если *P-value* $\geq \alpha$, то нулевая гипотеза H_0 принимается.

Проверка гипотезы о распределении генеральной совокупности. Критерий согласия Пирсона (критерий χ^2)

Иногда под конкурирующей гипотезой подразумевается то, что просто не выполнена основная. В этом случае задача проверки нулевой гипотезы ставится следующим образом: требуется про-

верить, согласуются ли результаты наблюдений с высказанным предположением. Соответствующие критерии для проверки таких гипотез называются критериями согласия.

Пусть имеется выборка объема n $(X_1, \dots, X_n)$ из генеральной совокупности с неизвестным теоретическим распределением. По выборочным данным при заданном уровне значимости α требуется проверить предположение, что теоретическое распределение имеет определенный вид $F(x, \theta)$, где $\theta = (\theta_1, \dots, \theta_k)$ — вектор неизвестных параметров распределения, т.е. гипотеза H_0 состоит в том, что функция распределения теоретической случайной величины X равна: $F(x, \theta) = P(X < x)$; $H_0 \sim F(x, \theta)$.

Для проверки данной гипотезы необходимо построить критерий, характеризующий меру расхождения между теоретической и эмпирической функциями распределения. Один из таких критериев был предложен Пирсоном.

Критерий Пирсона основан на сравнении теоретических и эмпирических частот. Для построения критерия множество возможных значений теоретической случайной величины X разбивается на m непересекающихся интервалов: $(-\infty; z_1)$, $[z_1; z_2), \dots, [z_{m-1}; +\infty)$; затем вычисляются теоретические и эмпирические частоты.

Напомним, что эмпирической частотой n_j называется число точек вариационного ряда, попавших в j-й интервал $\left[z_{j-1}; z_j\right)$. Под теоретической частотой понимается математическое ожидание числа наблюдений, которые должны попасть в j-й интервал в соответствии с теоретическим распределением $F(x, \theta)$. По *определению 2* случайные величины $X_1, \dots, X_n$ независимы, и распределение каждой из них совпадает с гипотетическим распре-

делением: $X_i \sim F(x, \theta)$. Обозначим $p_j = P\left(X_i \in \left[z_{j-1}; z_j\right)\right)$ — вероятность того, что *i*-е наблюдение попало в *j*-й интервал; поскольку случайные величины $X_1, \ldots, X_n$ независимы и одинаково распределены, эта вероятность не зависит от *i*. Поэтому выборку можно интерпретировать как результат *n* независимых испытаний, где в каждом испытании успех (попадание наблюдения в *j*-й интервал) происходит с вероятностью p_j, т.е. как схему независимых испытаний Бернулли. Напомним, что число успехов в схеме Бернулли имеет биномиальное распределение, а математическое ожидание числа успехов вычисляется как произведение числа испытаний на вероятность успеха в одном испытании. Таким образом, теоретическая частота равна np_j. Для вычисления вероятностей p_j воспользуемся известными свойствами функции распределения:

$$p_j = P\left(X_i \in \left[z_{j-1}; z_j\right)\right) = F\left(z_j, \theta\right) - F\left(z_{j-1}, \theta\right);$$

для первого интервала эта вероятность равна: $p_1 = P\left(X_i \in (-\infty; z_1)\right) = F(z_1, \theta)$,

а для последнего: $p_m = P\left(X_i \in [z_{m-1}; +\infty)\right) = 1 - F(z_{m-1}, \theta)$; очевидно, $\sum_{j=1}^{m} p_j = 1$.

Поскольку теоретическое распределение зависит от параметров, значения которых неизвестны, в формулы для вычисления p_j подставляют точечные оценки этих параметров $\hat{\theta} = \left(\hat{\theta}_1, \ldots, \hat{\theta}_k\right)$, найденные по результатам наблюдений.

Пирсон показал, что при $n \to \infty$, независимо от закона распределения генеральной совокупности, распределение случайной величины $K = \sum_{j=1}^{m} \frac{\left(n_j - np_j\right)^2}{np_j}$ сходится к распределению χ^2 с

$m - k - 1$ степенями свободы; m — количество интервалов разбиения, k — число неизвестных параметров гипотетического распределения.

Для проверки гипотезы H_0 построим правостороннюю критическую область; критическая точка является решением уравнения $P(K \geq k_{кр}) = \alpha$, где случайная величина K имеет χ^2-распределение с $m - k - 1$ степенями свободы, $K \sim \chi^2_{m-k-1}$.

По выборочным данным найдем наблюдаемое значение критерия $K_{набл}$ и сравним его с $k_{кр}$. Если $K_{набл}$ попадает в критическую область, т.е. $K_{набл} > k_{кр}$, то гипотеза H_0 отвергается. Это означает, что экспериментальные данные не согласуются с выдвинутым предположением о виде распределения генеральной совокупности. Если $K_{набл}$ попадает в допустимую область, т.е. $K_{набл} < k_{кр}$, то гипотеза H_0 принимается, т.е. гипотеза о виде теоретического распределения не противоречит результатам наблюдений.

Замечание 7. При построении критерия Пирсона рекомендуется выбирать интервалы разбиения таким образом, чтобы для каждого интервала теоретическая частота была не менее 10: $np_j \geq 10$.

Замечание 8. В данном случае конкурирующая гипотеза является сложной (состоит из бесконечного числа простых гипотез о виде теоретического распределения), соответственно, отсутствует аналитическое представление условного распределения критерия при условии выполнения конкурирующей гипотезы $P(K < x|H_1), x \in \mathbb{R}$, поэтому вычисление мощности критерия не представляется возможным.

Проверка гипотезы об экспоненциальном распределении генеральной совокупности

В качестве примера проверим гипотезу об экспоненциальном распределении генеральной совокупности на основе выборочных данных, полученных в задании 1.

По *замечанию 2* распределение Вейбулла $W(1,1)$ — это экспоненциальное распределение с параметром 1 $\left(Exp(1)\right)$. Поэтому выборка, смоделированная в задании 1, — это выборка из генеральной совокупности с экспоненциальным теоретическим распределением $Exp(1)$. Зададим уровень значимости $\alpha = 0{,}1$ и с помощью критерия Пирсона проверим, согласуются ли полученные выборочные данные с гипотезой об экспоненциальном распределении генеральной совокупности.

По замечанию 7 интервалы разбиения при построении критерия следует выбирать таким образом, чтобы $np_j \geq 10$. Поскольку объем выборки $n = 30$, то $p_j \geq \frac{1}{3}$, при этом должно выполняться условие $\sum_{j=1}^{m} p_j = 1$, тогда максимально возможное число интервалов разбиения $m = 3$, и для каждого интервала $p_j = \frac{1}{3}$. Найдем точки разбиения z_1 и z_2, определяющие границы этих интервалов. По формулам для вычисления p_j получаем: $p_1 = F(z_1, \theta) = \frac{1}{3}$; $p_2 = 1 - F(z_2, \theta) = \frac{1}{3}$. Отсюда, учитывая, что гипотетическое распределение F не зависит от неизвестных параметров, получаем $F(z_1, \theta) = \frac{1}{3}$; $F(z_2, \theta) = \frac{2}{3}$.

Для вычисления границ интервалов разбиения воспользуемся значениями обратной функции для проверяемого распределения:

```
>>> z1, z2 = weibull_min.ppf([1/3, 2/3], c=1)
>>> print(z1, z2)
(0.4054651081081643, 1.0986122886681096)
```

Теперь вычислим наблюдаемые частоты попадания в интервалы $(0; z_1)$, $[z_1; z_2)$, и $[z_2; +\infty)$.

```
>>> n1 = np.count_nonzero(X < z1)
>>> n2 = np.count_nonzero((z1 <= X) & (X < z2))
>>> n3 = np.count_nonzero(z2 <= X)
>>> obs_freq = np.array([n1, n2, n3])
>>> obs_freq
array([12, 10,  8])
```

Поскольку по построению ожидаемые частоты равны 10 (так как $np_j = \frac{1}{3} \cdot 30 = 10$), наблюдаемое значение критерия Пирсона

$$K_{\text{набл}} = \sum_{j=1}^{m} \frac{\left(n_j - np_j\right)^2}{np_j} = \sum_{j=1}^{m} \frac{\left(n_j - 10\right)^2}{10} = \frac{1}{10} \sum_{j=1}^{m} \left(n_j - 10\right)^2$$

оказывается равным 0,8:

```
# Pearson statistic
>>> exp_freq = 10
>>> K = np.sum((obs_freq - exp_freq)**2) / exp_freq
>>> print(K)
0.8
```

Для построения правосторонней критической области найдем критическую точку $k_{\text{кр}}$. Для этого надо решить уравнение $P\left(K \geq k_{\text{кр}}\right) = \alpha$, где уровень значимости задан $\alpha = 0{,}1$, а случайная величина K имеет распределение χ^2 с $m - k - 1 = 3 - 0 - 1 = 2$ степенями свободы: $k = 0$, поскольку гипотетическое распределение полностью задано и не зависит от неизвестных параметров.

Решение данного уравнения находится вызовом метода isf распределения хи-квадрат:

```
>>> alpha=0.1
>>> from scipy.stats import chi2
>>> k_c = chi2.isf(alpha, df=2)
>>> print(k_c)
4.605170185988092
```

Поскольку $K_{\text{набл}} < k_{\text{кр}}$, т.е. наблюдаемое значение критерия попало в допустимую область, гипотеза об экспоненциальном распределении генеральной совокупности принимается.

Проиллюстрируем графически данное утверждение (рис. 6):

```
>>> xx = np.linspace(0, 7, 101)
>>> plt.plot(xx, chi2.cdf(xx, df=2),
...          label=r"cdf  $\chi^2$(df=2)")
>>> alpha = 0.1
>>> plt.axhline(y=1-alpha, ls='--',
...             label=r"$\alpha = %s$"%alpha, color='C1')
>>> plt.plot([K, K, 0],
...          [0, chi2.cdf(K, df=2), chi2.cdf(K, df=2)], '-')
>>> plt.plot(K, chi2.cdf(K, df=2), 'o', ms=8, label='K_Pearson')
>>> plt.legend(loc='best')
>>> plt.xlabel('K')
>>> plt.ylabel('cdf  $\chi^2$')
>>> plt.grid()
```

Отметим альтернативный способ проведения вычислений: после того как рассчитаны наблюдаемые частоты, проверку гипотезы осуществляет библиотечная функция chisquare:

```
>>> from scipy.stats import chisquare
>>> statistic, pvalue = chisquare(obs_freq, [exp_freq]*3)
>>> print(statistic, pvalue)
(0.8, 0.6703200460356394)
```

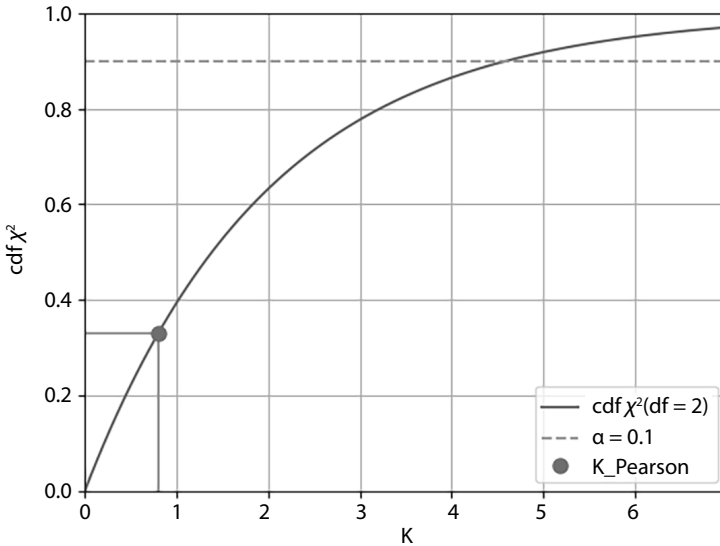

Рис. 6. Критерий Пирсона: критическая область и наблюдаемое значение критерия

Заметим, что статистика совпала с вычисленным «руками» значением, а величина *p-value* дается просто значением дополнения функции распределения χ^2, поскольку в данном примере рассматривается правосторонняя критическая область (ср. с рис. 6):

```
>>> chi2.sf(K, df=2)
0.6703200460356394
```

Заметим, что на рис. 6 взят уровень значимости 0,1. Для уровня 0,01 или 0,05 значения статистики Пирсона и *p-value* останутся неизменными, т.е. гипотеза также будет принята.

Литература

1. *Белько И.В., Свирид Г.П.* Теория вероятностей и математическая статистика. Примеры и задачи. Минск: Новое знание, 2002. 250 с.
2. *Бочаров П.П., Печинкин А.В.* Теория вероятностей. Математическая статистика. М.: Гардарика, 1998. 327 с.
3. *Гмурман В.Е.* Теория вероятностей и математическая статистика. М.: Высшая школа, 2000. 480 с.
4. *Ивченко Г.И., Медведев Ю.И.* Введение в математическую статистику. Учебник. М.: Изд-во ЛКИ, 2010.
5. *Коваленко И.Н., Филиппова А.А.* Теория вероятностей и математическая статистика. М.: Высшая школа, 1973. 368 с.
6. *Колемаев В.А., Староверов О.В., Турундаевский В.Б.* Теория вероятностей и математическая статистика. М.: Высшая школа, 1991. 400 с.
7. *Гришунина Ю.Б.* Задачи математической статистики и их решение с использованием приложения Microsoft Excel. М.: РИО МИЭМ НИУ ВШЭ, 2013.

Вопросы для повторения

1. Задачи математической статистики. Генеральная совокупность. Выборка. Вариационный ряд.
2. Эмпирическая функция распределения.
3. Гистограмма и полигон частот.
4. Точечные оценки. Свойства оценок (несмещенность, состоятельность, эффективность).
5. Выборочное среднее. Выборочная дисперсия. Исправленная выборочная дисперсия.
6. Метод моментов.
7. Найти с помощью метода моментов точечные оценки для: неизвестного параметра λ распределения Вейбулла $W(r, \lambda)$, считая параметр r известным; неизвестного параметра μ логнормального распределения $LogN(\mu, \sigma)$, считая параметр σ известным.
8. Метод максимального правдоподобия. Функция правдоподобия.
9. Найти с помощью метода максимального правдоподобия точечные оценки для: неизвестного параметра λ распределения Вейбулла $W(r, \lambda)$, считая параметр r известным; неизвестного параметра μ логнормального распределения $LogN(\mu, \sigma)$, считая параметр σ известным.
10. Доверительные интервалы. Точность и надежность доверительных интервалов.
11. Доверительный интервал для неизвестного математического ожидания при известной дисперсии (нормальное распределение).

12. Доверительный интервал для неизвестного математического ожидания при известной и неизвестной дисперсии (произвольное распределение).
13. Статистические гипотезы. Критерии. Ошибки первого и второго рода. Уровень значимости. Схема проверки статистических гипотез.
14. Критерий согласия Пирсона (χ^2). Проверка гипотезы о виде распределения генеральной совокупности.

Учебное издание

Буровский Евгений Андреевич,
Гришунина Юлия Борисовна

Задачи математической статистики и их решение с использованием языка программирования Python

Зав. книжной редакцией *Е.А. Бережнова*
Редактор *Н.М. Дмуховская*
Компьютерная верстка: *Ю.Н. Петрина*
Корректор *Н.М. Дмуховская*, *Н.В. Андрианова*

Дизайн обложки:

Все новости издательства — http://id.hse.ru

По вопросам закупки книг
обращайтесь в отдел реализации
Тел.: +7 495 772-95-90 доб. 15295, 15297
bookmarket@hse.ru

Подписано в печать 02.06.2022.
Формат 60×88/16. Гарнитура Cambria
Усл. печ. л. 3,9. Уч.-изд. л. 1,8. Тираж 150 экз.
Изд. № 2646. Заказ

Национальный исследовательский университет
«Высшая школа экономики»
101000, Москва, ул. Мясницкая, 20
Тел.: +7 495 772-95-90 доб. 15285

Отпечатано в «ООО Фотоэксперт»
109316, Москва, Волгоградский проспект, д. 42