Практическая работа № 5. Интервальное оценивание параметров распределения по выборке

5.1. О содержании и задачах практической работы

В настоящей практической работе обсуждаются четыре задачи.

1. Построение доверительных интервалов параметров генеральной совокупности по выборке из известного закона распределения с использованием стандартных библиотечных средств языка программирования Python.

2. Изучение свойств интервальных оценок параметров распределения.

3. Планирование эксперимента с целью получения оценки параметра с заданной точностью.

4. «Нетрадиционные» методы построения доверительных интервалов.

О первой задаче. В языке программирования Python есть несколько функций, которые строят некоторые доверительные интервалы непосредственно по выборке (или по заранее рассчитанным по выборке точечным оценкам числовых параметров). Однако все потребности ими не покрываются, что, впрочем, не является проблемой, поскольку написание программы, находящей границы доверительных интервалов, не составляет труда.

О второй задаче. Для того, чтобы корректно и эффективно использовать доверительные интервалы для анализа данных, требуется осознать вероятностную суть этого понятия, противоречивость между двумя характеристиками транслируемой им информации – достоверностью утверждения (доверительной вероятностью) и конкретностью (линейным размером интервала), четко понимать зависимость качества информации от объема анализируемых данных. Все вышеперечисленное требует уяснения, и выполнение нескольких специально разработанных заданий данной практической работы призвано этому поспособствовать.

О третьей задаче. Объем выборки оказывает определяющее влияние на точность оценок, и это необходимо учитывать при планировании экспериментов. Экспериментатор часто задается вопросом: сколько раз нужно провести эксперимент, чтобы отклонение точечной оценки от истинного значения параметра с приемлемой вероятностью не превышало некоторого значения?

О четвертой задаче. При традиционном подходе, если выборка задана, то для получения наиболее эффективных оценок надо использовать все до одного выборочные наблюдения. Исключение из расчета каждого наблюдения означает уменьшение на единицу числа степеней свободы со всеми вытекающими последствиями. А если наблюдений не так много? Тогда попробуем воспользоваться бутстреп-методом, который в отличие от традиционных методов предполагает многократную обработку различных частей одних и тех же данных. И еще очень важно, что для построения доверительных интервалов традиционным методом необходимо знать закон распределения генеральной совокупности, а он не всегда известен. Бутстреп-процедура не требует априорной информации о виде закона распределения изучаемой случайной величины и статистик, используемых для оценки ее параметров.

Бутстреп-метод состоит в многократном извлечении псевдовыборок из эмпирического распределения (имеющейся выборки). Для оценки параметров можно сформировать тысячи повторных бутстреп-выборок (обычно 500 – 10 000). Используя подвыборки, строят распределение любой статистики.

Известны разные способы построения доверительных интервалов на основе статистического ряда бутстрепа. Мы рассмотрим базовый, названный методом процентилей. Этот вариант бутстрепа в качестве границ центрального доверительного интервала использует процентили уровней $100\frac{\alpha}{2}$ и $100\left(1-\frac{\alpha}{2}\right)$ статистического ряда бутстрепа.

5.2. Математические понятия и утверждения: краткая информация и ссылки на источники

1. Понятие статистики; законы распределения некоторых статистик для нормальной генеральной совокупности $X \sim N(m,\sigma)$:

$U = \dfrac{\overline{X} - m}{\sigma / \sqrt{n}}$, $W = \dfrac{\overline{X} - m}{S / \sqrt{n}}$, $V_1 = \dfrac{S_0^2}{\sigma^2 / n}$, $V_2 = \dfrac{S^2}{\sigma^2 / (n-1)}$; закон распределения статистики $Z(p, P^*) = \dfrac{P^* - p}{\sqrt{pq/n}}$ для генеральной совокупности $X \sim B(n, p)$, распределенной по биномиальному закону (здесь $P^* = \dfrac{X}{n}$ – относительная частота успеха в n испытаниях) [1, с. 144 – 150; 2, с. 231 – 235].

2. Доверительный интервал; доверительная вероятность; уровень значимости [1, с. 166 – 177; 2, с. 237 – 245].

3. Доверительные интервалы для параметров нормально и биномиально распределенных генеральных совокупностей (таблица).

Параметр	Условия	Доверительный интервал	Номер формулы
m	$X \sim N(m,\sigma)$ σ известно	$\overline{X} - u_{\frac{1+\beta}{2}} \cdot \dfrac{\sigma}{\sqrt{n}} < m < \overline{X} + u_{\frac{1+\beta}{2}} \cdot \dfrac{\sigma}{\sqrt{n}}$	1
m	$X \sim N(m,\sigma)$ σ неизвестно	$\overline{X} - t_{\frac{1+\beta}{2}}(n-1) \cdot \dfrac{S}{\sqrt{n}} <$ $< m <$ $< \overline{X} + t_{\frac{1+\beta}{2}}(n-1) \cdot \dfrac{S}{\sqrt{n}}$	2
σ^2	$X \sim N(m,\sigma)$ m известно	$\dfrac{n \cdot S_0^2}{\chi^2_{\frac{1+\beta}{2}}(n)} < \sigma^2 < \dfrac{n \cdot S_0^2}{\chi^2_{\frac{1-\beta}{2}}(n)}$	3
σ^2	$X \sim N(m,\sigma)$ m неизвестно	$\dfrac{(n-1)S^2}{\chi^2_{\frac{1+\beta}{2}}(n-1)} < \sigma^2 < \dfrac{(n-1)S^2}{\chi^2_{\frac{1-\beta}{2}}(n-1)}$	4

Пара-метр	Условия	Доверительный интервал	Номер формулы
p	$X \sim B(n,p)$ $\sqrt{npq} \gg 1$	$\dfrac{P^* + \frac{1}{2}\frac{u^2_{\frac{1+\beta}{2}}}{n} - u_{\frac{1+\beta}{2}} \cdot \sqrt{\frac{P^*\left(1-P^*\right)}{n} + \frac{1}{4}\frac{u^2_{\frac{1+\beta}{2}}}{n^2}}}{1+\frac{u^2_{\frac{1+\beta}{2}}}{n}}$ $< p <$ $\dfrac{P^* + \frac{1}{2}\frac{u^2_{\frac{1+\beta}{2}}}{n} + u_{\frac{1+\beta}{2}} \cdot \sqrt{\frac{P^*\left(1-P^*\right)}{n} + \frac{1}{4}\frac{u^2_{\frac{1+\beta}{2}}}{n^2}}}{1+\frac{u^2_{\frac{1+\beta}{2}}}{n}}$	5
p	$X \sim B(n,p)$ $\sqrt{npq} \gg 1$	$P^* - u_{\frac{1+\beta}{2}} \cdot \sqrt{\frac{P^*\left(1-P^*\right)}{n}}$ $< p <$ $P^* + u_{\frac{1+\beta}{2}} \cdot \sqrt{\frac{P^*\left(1-P^*\right)}{n}}$	6

4. Построение доверительных интервалов бутстреп-методом.

Алгоритм построения 95 %-го доверительного интервала для параметра в простейшем варианте с помощью бутстреп-процедуры:

- выбираем n наблюдений из выборки (отбор осуществляется с возвратом, т.е. некоторые наблюдения могут быть выбраны несколько раз, а некоторые остаться невыбранными);
- вычисляем оценки параметра для полученного набора из n значений;
- повторяем последовательность действий, описанную в пп. 1 – 2, многократно (500 – 10 000 раз);

- отсортировываем полученный массив оценок по возрастанию;
- находим значения оценок, которые представляют собой 0,025 и 0,975 квантили. Два выбранных значения и будут границами 95 %-го доверительного интервала.

5.3. Библиотечные инструменты языка программирования Python

Загрузка основных модулей

```
import numpy as np
import scipy.stats as sts
import scipy.special as sc
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.stats.weightstats
import statsmodels.stats.proportion
%matplotlib inline
```

Построение доверительных интервалов в предположении нормального распределения генеральной совокупности

Функция `_zconfint_generic (mean, std_mean, alpha, alternative)` модуля `statsmodels.stats.weightstats`
Возвращает границы доверительного интервала (1).

Параметры: `mean` – выборочное среднее, `std_mean` = $\frac{\sigma}{\sqrt{n}}$; `alpha` – уровень значимости; `alternative` – вид доверительного интервала («`two-sided`» – двусторонний, по умолчанию, «`smaller`» – левосторонний, «`larger`» – правосторонний).

Функция `zconfint(x, alpha=0.05, alternative=«two-sided»)` модуля `statsmodels.api.stats`
Возвращает границы доверительного интервала (2).
Параметры: `x` – выборка; `alpha` – уровень значимости; `alternative` – вид доверительного интервала («`two-sided`» – двусторонний (по умолчанию), «`smaller`» – левосторонний, «`larger`» – правосторонний).

Построение доверительных интервалов в предположении биномиального распределения генеральной совокупности

Функция `proportion_confint(count=m, nobs=n, alpha, method='normal')` модуля `statsmodels.stats.proportion`
Возвращает границы доверительного интервала (5) или (6).
Параметры: `count` – число успехов; `nobs` – число испытаний; `alpha` – уровень значимости; `method` – вид доверительного интервала (если `'normal'`, строится интервал (6), если `'wilson'`, строится интервал (5)).

Функция `samplesize_confit_proportion(proportion, half_length, alpha, method='normal')` модуля `statsmodels.stats.proportion`
Возвращает минимальный объем выборки, необходимый для достижения желаемой точности при интервальном оценивании вероятности события.
Параметры: `proportion` – вероятность успеха; `half_length` – половина ширины требуемого интервала; `alpha` – уровень значимости (по умолчанию 0,05).

5.4. Примеры для совместного обсуждения

Пример 1. Чтобы попрактиковаться в построении доверительных интервалов, мы могли бы, как и в практических работах № 1 – 4, сгенерировать выборку нормально распределенной генеральной совокупности. Но в этом примере для разнообразия поработаем с реальными данными. На самом деле найти подходящие данные не так-то просто. В силу разных причин далеко не все результаты наблюдений оправдывают ожидания нормальности распределения (как проверять гипотезу о том, что выборка принадлежит нормально распределенной генеральной совокупности, мы обсудим позднее). На практике зачастую приходится снижать уровень требований к качеству исходного материала и работать с выборками, относительно которых мы предполагаем, что они получены из нормального распределения (или же предварительно выборки модифицировать). В этом примере воспользуемся данными Милликена по измерению заряда электрона, для которых гипотеза о нормальности

подтверждается. Данные представляют собой 58 значений величины q, полученных при определении величины заряда электрона $q \cdot 10^{-10}$ ед.:
4.740, 4.747, 4.749, 4.758, 4.761, 4.764, 4.764, 4.764, 4.765, 4.767, 4.768, 4.769, 4.769, 4.771, 4.771, 4.772, 4.772, 4.772, 4.774, 4.775, 4.775, 4.776, 4.777, 4.777, 4.778, 4.779, 4.779, 4.779, 4.781, 4.781, 4.782, 4.783, 4.783, 4.785, 4.785, 4.785, 4.788, 4.788, 4.789, 4.789, 4.790, 4.790, 4.790, 4.791, 4.791, 4.791, 4.792, 4.792, 4.795, 4.797, 4.799, 4.799, 4.801, 4.805, 4.806, 4.808, 4.809, 4.810.

Вначале построим доверительные интервалы для математического ожидания и дисперсии при доверительной вероятности $0,9$ по формулам (2) и (4).

Дополнительно найдем интервал для математического ожидания, воспользовавшись функцией `zconfint_generic()`.

Теперь решим обратную задачу: ответим на вопрос, с какой вероятностью можно утверждать, что абсолютное значение ошибки определения математического ожидания величины заряда по выборке значений q не превзошло $0,0001 \cdot 10^{-10}$ ед.? Вероятность неравенства $|m-\overline{x}| < \varepsilon$ $(\varepsilon = 0,0001)$ определяется распределением Стьюдента с $n-1$ степенью свободы (далее, F – его функция распределения):

$$P\{|m-\overline{x}| < \varepsilon\} = F\left(\frac{\varepsilon\sqrt{58}}{s}\right) - F\left(-\frac{\varepsilon\sqrt{58}}{s}\right).$$

Вариант выполнения всех расчетов приведен в приложении к практической работе № 5, пример 1.

Пример 2. Обсудим вопрос о влиянии на длину доверительного интервала объема выборки.

1) Начнем со случая, когда нет проблем с получением аналитической формулы, выражающей соответствующую зависимость. Пусть генеральная совокупность имеет нормальное распределение, причем среднее квадратичное отклонение известно. При этих условиях доверительный интервал для математического ожидания строится по формуле (1), его длина равна $2 \cdot u_{\frac{1+\beta}{2}} \cdot \frac{\sigma}{\sqrt{n}}$. Для определенности возьмем доверительную вероятность равной 0,95 и построим по этой формуле график зависимости длины интервала от объема выборки n.

2) Пусть теперь мы не знаем параметров генеральной совокупности и нас интересуют доверительные интервалы для дисперсии. Будем изучать вопрос о зависимости длины интервалов от объема выборки экспериментально. Сгенерируем выборки различных объемов n из генеральной совокупности, имеющей нормальное распределение с конкретными параметрами m,σ. Значения n будем брать в диапазоне $[10,1000]$ с шагом 10. Для каждой выборки построим по формуле (4) доверительный интервал для дисперсии при неизвестном математическом ожидании (доверительную вероятность примем равной 0,95). Результаты визуализируем, построив график экспериментальной зависимости верхней и нижней границ доверительных интервалов от объема выборки, а также график экспериментальной зависимости длины интервала от объема выборки.

Вариант выполнения этого пункта задания приведен в приложении к практической работе № 5, пример 2. Результаты показывают, что увеличение объема выборки при фиксированной доверительной вероятности ведет к уменьшению длины доверительных интервалов (т.е. при увеличении объема выборки мы получаем более конкретную информацию о диапазоне значений параметров).

Пример 3. Исследуем экспериментально вопрос о влиянии на длину доверительного интервала доверительной вероятности. Сгенерируем выборку объема n из генеральной совокупности, имеющей нормальное распределение с конкретными параметрами m, σ (значение параметров n, m, σ каждый выберет самостоятельно).

1) Построим в одной системе координат графики зависимости левой и правой границ доверительных интервалов для математического ожидания от значения доверительной вероятности β. Определим диапазон значений β, при которых доверительный интервал, построенный по сгенерированной выборке, не накрывает истинное значение математического ожидания.

2) Выполним задание, аналогичное п. 1, для дисперсии.

Что показывают полученные результаты?

5.5. Задания для самостоятельного выполнения

Задание 1. Экспериментально сопоставьте методы получения интервальных оценок: бутстреп и параметрический (традиционный). Сгенерируйте 1000 выборок из генеральной совокупности, имеющей нормальное распределение с конкретными параметрами m, σ (значение параметров m, σ и объем выборки выберите самостоятельно).

1) Для каждой выборки получите интервальные оценки для математического ожидания и дисперсии методом бутстреп (доверительную вероятность примите равной 0,95). Какая часть полученных интервалов накрыла соответствующие параметры?

2) Для тех же выборок получите интервальные оценки для математического ожидания и дисперсии традиционным методом (доверительную вероятность примите равной 0,95). Какая часть полученных интервалов накрыла соответствующие параметры?

Сопоставьте результаты пунктов 1 и 2 и прокомментируйте их.

Задание 2. Исследуйте вопрос о минимальном объеме выборки, необходимом для достижения заданной точности оценки при проведении экспериментов по схеме Бернулли. При выполнении задания исходите из предположения $X \sim B(n, p)$, $\sqrt{npq} \gg 1$ и используйте приближенный доверительный интервал

$$P^* - u_{\frac{1+\beta}{2}} \cdot \sqrt{\frac{P^*\left(1-P^*\right)}{n}} < p < P^* + u_{\frac{1+\beta}{2}} \cdot \sqrt{\frac{P^*\left(1-P^*\right)}{n}}$$

1) Сгенерируйте выборки различных объемов n из генеральной совокупности, имеющей индикаторное распределение с параметром p (значение параметра p возьмите равным отношению длины вашего имени к сумме длин имени и фамилии). Для каждой выборки найдите доверительный интервал для вероятности события p (доверительную вероятность принять равной 0,95). Результаты визуализируйте, построив график экспериментальной зависимости верхней и нижней границ доверительных интервалов от объема выборки. Экспериментально оцените минимальный объем выборки, позволяющий при заданной доверительной вероятности утверждать, что доля «успехов» в выборке отличается от значения параметра p не более чем на 2 %.

2) Теоретически оцените минимальный объем выборки, позволяющий при заданной доверительной вероятности утверждать, что доля «успехов» в выборке отличается от значения параметра p не более чем на 2 % (2, стр. 244-245).

3) Используя функцию samplesize_confint_proportion, оцените минимальный объем выборки, позволяющий при заданной доверительной вероятности утверждать, что доля «успехов» в выборке отличается от значения параметра p не более чем на 2 %.

Сравните результаты, полученные при выполнении п. 1, п. 2 и п. 3.