

Санкт-Петербургский Политехнический университет
Петра Великого
Институт Прикладной Математики и Механики
Кафедра «Прикладная Математика и Информатика»

Отчет
По лабораторной работе № 7
По Дисциплине «Математическая статистика»

Выполнил:
Студент Селянкин Федор
Группа 3630102/70301
Проверил:
к.ф. – м.н., доцент
Баженов Александр Николаевич

Санкт-Петербург
2020 г.

Содержание

Содержание	2
Постановка задачи	3
Теория.....	3
Метод максимального правдоподобия	3
Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат.....	5
Реализация.....	7
Результаты.....	8
Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат.....	8
Обсуждения	9
Литература	10
Список Иллюстраций:	
Рисунок 1 График частот для выборки мощностью 10 и график теоретических значений частот на таком же разбиении.....	8
Рисунок 2 График частот для выборки мощностью 10 и график теоретических значений частот на таком же разбиении.....	9
Список Таблиц:	
Таблица 1 Вычисление χ^2 при проверке гипотезы H_0 о нормальном законе распределения $N(\mu, \sigma)$ для выборки мощностью 100	8
Таблица 2 Вычисление χ^2 при проверке гипотезы H_0 о нормальном законе распределения $N(\mu, \sigma)$ для выборки мощностью 100	9

Постановка задачи

Задание 7: сгенерировать выборку объемом 10 и 100 элементов для нормального распределения $N(x, 0, 1)$. По сгенерированной выборке оценить параметры μ и σ нормального закона методом максимального правдоподобия. В качестве основной гипотезы H_0 будем считать, что сгенерированное распределение имеет вид $N(x, \hat{\mu}, \hat{\sigma})$. Проверить основную гипотезу, используя критерий согласия χ^2 . В качестве уровня значимости взять $\alpha = 0.05$. Привести таблицу вычислений χ^2 .

Построить гистограмму сгенерированных событий и теоретическое распределение на одном графике.

Если гипотеза не подтвердилась указать причины.

Теория

Метод максимального правдоподобия

Одним из универсальных методов оценивания является метод максимального правдоподобия, предложенный Р. Фишером (1921).

Пусть x_1, \dots, x_n – случайная выборка из генеральной совокупности с плотностью вероятности $f(x, \theta)$; $L(x_1, \dots, x_n, \theta)$ – функция правдоподобия (ФП), представляющая собой совместную плотность вероятности независимых с. в. x_1, \dots, x_n и рассматриваемая как функция неизвестного параметра θ :

$$L(x_1, \dots, x_n, \theta) = f(x_1, \theta)f(x_2, \theta) \dots f(x_n, \theta)$$

(1)

Определение. Оценкой максимального правдоподобия (о.м.п) будем называть такое значение $\hat{\theta}_{\text{мп}}$ из множества допустимых значений параметра θ , для которого ФП принимает наибольшее значение при заданных x_1, \dots, x_n .

(2)

$$\hat{\theta}_{\text{мп}} = \arg \max_{\theta} L(x_1, \dots, x_n, \theta)$$

Если ФП дважды дифференцируема, то её стационарные значения даются корнями уравнения

(3)

$$\frac{\partial L(x_1, \dots, x_n, \theta)}{\partial \theta} = 0$$

Достаточным условием того, чтобы некоторое стационарное значение $\tilde{\theta}$ было локальным максимумом, является неравенство

(4)

$$\frac{\partial^2 L}{\partial \theta^2}(x_1, \dots, x_n, \bar{\theta}) < 0$$

Определив точки локальных максимумов ФП (если их несколько), находят наибольший, который и даёт решение задачи (1).

Часто проще искать максимум логарифма ФП, так как он имеет максимум в одной точке с ФП:

$$\frac{\partial \ln L}{\partial \theta} = \frac{1}{L} \frac{\partial L}{\partial \theta}, \text{ если } L > 0,$$

и соответственно решать уравнение

(5)

$$\frac{\partial \ln L}{\partial \theta} = 0,$$

Которое называют *уравнением правдоподобия*.

В задаче оценивания векторного параметра $\theta = (\theta_1, \dots, \theta_m)$ аналогично (2) находится максимум ФП нескольких аргументов:

(6)

$$\hat{\theta}_{\text{МП}} = \arg \max_{\theta_1, \dots, \theta_m} L(x_1, \dots, x_n, \theta_1, \dots, \theta_m)$$

и в случае дифференцируемости ФП выписывается система уравнений правдоподобия

(7)

$$\frac{\partial L}{\partial \theta_k} = 0 \text{ или } \frac{\partial \ln L}{\partial \theta_k} = 0, k = 1, \dots, m$$

Пример. Оценивание м.о. m и дисперсии σ^2 нормального распределения $N(m, \sigma)$.

Составим функцию правдоподобия и получим последовательно:

$$L(x_1, \dots, x_n, m, \sigma) = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{(x_i - m)^2}{2\sigma^2} \right\} = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2 \right\},$$

$$\ln L = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2.$$

Получим следующие уравнения правдоподобия:

$$\begin{cases} \frac{\partial \ln L}{\partial m} = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n (x_i - \bar{m}) = \frac{n}{\hat{\sigma}^2} (\bar{x} - \bar{m}) = 0, \\ \frac{\partial \ln L}{\partial (\sigma^2)} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^n (x_i - \hat{m})^2 = \frac{n}{2(\hat{\sigma}^2)^2} \left[\frac{1}{n} \sum_{i=1}^n (x_i - \hat{m})^2 - \hat{\sigma}^2 \right] = 0, \end{cases}$$

откуда следует, что выборочное среднее \bar{x} – о.м.п. математического ожидания: $\hat{m}_{\text{МП}} = \bar{x}$, а выборочная дисперсия $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ – о.м.п. генеральной дисперсии: $\hat{\sigma}_{\text{МП}}^2 = s^2$.

Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

Исчерпывающей характеристикой изучаемой случайной величины является её закон распределения. Поэтому естественно стремление исследователей построить этот закон приближённо на основе статистических данных.

Сначала выдвигается гипотеза о виде закона распределения.

После того как выбран вид закона, возникает задача оценивания его параметров и проверки (тестирования) закона в целом.

Для проверки гипотезы о законе распределения применяются критерии согласия. Таких критериев существует много. Мы рассмотрим наиболее обоснованный и наиболее часто используемый в практике – критерий χ^2 (хи-квадрат), введённый К. Пирсоном (1900 г.) для случая, когда параметры распределения известны. Этот критерий был существенно уточнён Р. Фишером (1924 г.), когда параметры распределения оцениваются по выборке, используемой для проверки.

Мы ограничимся рассмотрением случая одномерного распределения.

Итак, выдвинута гипотеза H_0 о генеральном законе распределения с функцией распределения $F(x)$.

Рассматриваем случай, когда гипотетическая функция распределения $F(x)$ не содержит неизвестных параметров.

Разобьём генеральную совокупность, т.е. множество значений изучаемой случайной величины X на k непересекающихся подмножеств $\Delta_1, \dots, \Delta_k$.

Пусть $p_i = P(X \in \Delta_i), i = 1, \dots, k$.

Если генеральная совокупность – вся вещественная ось, то подмножества $\Delta_i = (a_{i-1}, a_i]$ – полуоткрытые промежутки ($i = 2, \dots, k-1$). Крайние промежутки будут полубесконечными: $\Delta_1 = (-\infty, a_1], \Delta_k = (a_{k-1}, +\infty)$. В этом случае $p_i = F(a_i) - F(a_{i-1})$; $a_0 = -\infty, a_k = +\infty$ ($i = 1, \dots, k$).

Пусть, далее, n_1, n_2, \dots, n_k – частоты попадания выборочных элементов в подмножества $\Delta_1, \dots, \Delta_k$ соответственно.

В случае справедливости гипотезы H_0 относительные частоты $\frac{n_i}{n}$ при большом n должны быть близки к вероятностям p_i ($i = 1, \dots, k$), поэтому за меру отклонения выборочного распределения от гипотетического с функцией $F(x)$ естественно выбрать величину

(8)

$$Z = \sum_{i=1}^n c_i \left(\frac{n_i}{n} - p_i \right)^2,$$

где c_i – какие-нибудь положительные числа (веса). К. Пирсоном в качестве весов выбраны числа $c_i = \frac{n}{p_i}$ ($i = 1, \dots, k$). Тогда получается статистика критерия хи-квадрат К. Пирсона

(9)

$$\chi^2 = \sum_{i=1}^k \frac{n}{p_i} \left(\frac{n_i}{n} - p_i \right)^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i},$$

Которая обозначена тем же символом, что и закон распределения хи-квадрат.

К. Пирсоном доказана теорема об асимптотическом поведении статистики χ^2 , указывающая путь её применения.

Теорема К. Пирсона. Статистика критерия χ^2 асимптотически распределена по закону χ^2 с $k - 1$ степенями свободы.

Это означает, что независимо от вида проверяемого распределения, т.е. функции $F(x)$, выборочная функция распределения статистики χ^2 при $n \rightarrow \infty$ стремится к функции распределения случайной величины с плотностью вероятности

(10)

$$f_{k-1}(x) = \begin{cases} 0, & x \leq 0 \\ \frac{1}{2^{\frac{k-1}{2}} \Gamma(\frac{k-1}{2})} x^{\frac{k-3}{2}} e^{-\frac{x}{2}}, & x > 0 \end{cases}$$

Для пояснения сущности метода χ^2 сделаем ряд замечаний.

Замечание 2.: Выбор подмножеств $\Delta_1, \dots, \Delta_k$ и их числа k в принципе ничем не регламентируется, так как $n \rightarrow \infty$. Но так как число n хотя и очень большое, но конечное, то k , должно быть с ним согласовано. Обычно его берут таким же, как и для построения гистограммы, т.е. можно руководствоваться формулой

(11)

$$k \approx 1.72 \sqrt[3]{n}$$

Или формулой Старджесса

(12)

$$k \approx 1 + 3.3 \lg n$$

При этом, если $\Delta_1, \Delta_2, \dots, \Delta_k$ – промежутки, то их длины удобно сделать равными за исключением крайних – полубесконечных.

Замечание 2. (о числе степеней свободы).

Числом степеней свободы функции (по старой терминологии) называется число её независимых аргументов. Аргументами статистики χ^2 являются частоты n_1, n_2, \dots, n_k . Эти частоты связаны одним равенством $n_1 + n_2 + \dots + n_k = n$, а в остальном независимы в силу независимости элементов выборки. Таким образом, функция χ^2 имеет $k - 1$ независимых аргументов: число частот минус одна связь. В силу теоремы Пирсона число степеней свободы статистики χ^2 отражается на виде асимптотической плотности $f_{k-1}(x)$.

На основе общей схемы проверки статистических гипотез сформулируем следующее правило.

Правило проверки гипотезы о законе распределения по методу χ^2 .

1. Выбираем значимости α .
2. По таблице [3, с 358] находим квантиль $\chi^2_{1-\alpha}(k-1)$ распределения хи-квадрат с $k-1$ степенями свободы порядка $1-\alpha$.
3. С помощью гипотетической функции распределения $F(x)$ вычисляем вероятности $p_i = P(X \in \Delta_i), i = 1, \dots, k$.
4. Находим частоты n_i попадания элементов выборки в подмножества $\Delta_i, i = 1, \dots, k$.
5. Вычисляем выборочное значение статистики критерия χ^2 :

$$\chi_B^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}.$$

6. Сравниваем χ_B^2 и квантиль $\chi^2_{1-\alpha}(k-1)$.
 - a. Если $\chi_B^2 < \chi^2_{1-\alpha}(k-1)$, то гипотеза H_0 на данном этапе проверки принимается.
 - b. Если $\chi_B^2 \geq \chi^2_{1-\alpha}(k-1)$, то гипотеза H_0 отвергается, выбирается одно из альтернативных распределений, и процедура проверки повторяется.

Замечание 3. Из формулы (9) видим, что веса $c_i = n/p_i$ пропорциональны n , т.е. с ростом n увеличиваются. Отсюда следует, что если выдвинутая гипотеза неверна, то относительные частоты n_i/n не будут близки к вероятностям p_i , и с ростом n величина χ_B^2 будет увеличиваться. При фиксированном уровне значимости α будет фиксированное пошаговое число – квантиль $\chi^2_{1-\alpha}(k-1)$, поэтому, увеличивая n , мы придём к неравенству $\chi_B^2 > \chi^2_{1-\alpha}(k-1)$, т.е. с увеличением объема выборки неверная гипотеза будет отвергнута.

Отсюда следует, что при сомнительной ситуации, когда $\chi_B^2 \approx \chi^2_{1-\alpha}(k-1)$, можно попытаться увеличить объем выборки (например, в 2 раза), чтобы требуемое неравенство было более чётким.

Замечание 4. Теория и практика применения критерия χ^2 указывают, что если для каких-либо подмножеств $\Delta_i (i = 1, \dots, k)$ условие $np_i \geq 5$ не выполняется, то следует объединить соседние подмножества (промежутки).

Это условие выдвигается требованием близости величин

$$(n_i - np_i)/\sqrt{np_i},$$

Квадраты которых являются слагаемыми χ^2 к нормальным $N(0,1)$. Тогда случайная величина в формуле (9) будет распределена по закону, близкому к хи-квадрат. Такая близость обеспечивается достаточной численностью элементов в подмножествах Δ_i .

Реализация

Лабораторная работа выполнена с помощью встроенных средств языка программирования Python в среде разработки PyCharm, с использованием дополнительных библиотек для отображения и расчетов. Исходный код лабораторной выложен на веб-сервисе GitHub [1].

Результаты

Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

Мощность выборки = 10

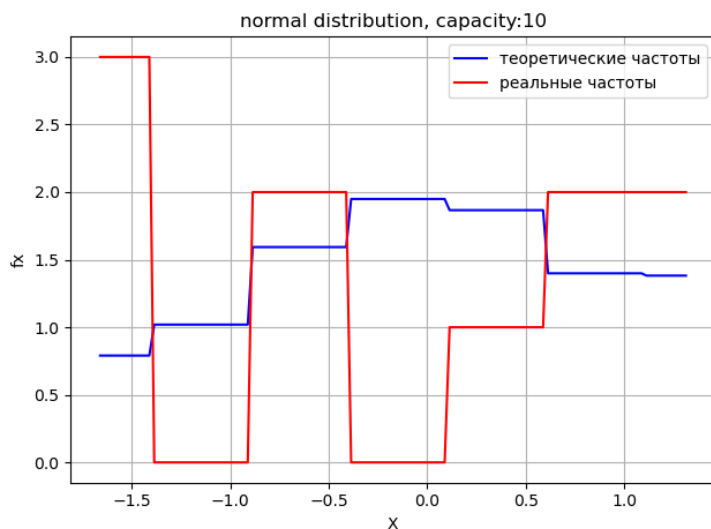


Рисунок 1 График частот для выборки мощностью 10 и график теоретических значений частот на таком же разбиении

Метод максимального правдоподобия:

$$\hat{\mu} \approx -0.16 \quad \hat{\sigma} \approx 1.69$$

Критерий согласия χ^2 :

Количество промежутков $k = 7$

i	Границы Δ_i	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
1	$-\infty, -1.41$	3	0.0790	0.79	2.21	6.18
2	$-1.41, -0.91$	0	0.1020	1.02	-1.02	1.02
3	$-0.91, -0.41$	2	0.1593	1.59	-0.41	0.10
4	$-0.41, 0.09$	0	0.1949	1.95	-1.95	1.95
5	$0.09, 0.59$	1	0.1866	1.87	-0.87	0.40
6	$0.59, 1.09$	2	0.1399	1.40	0.60	0.26
7	$1.09, \infty$	2	0.1382	1.38	0.62	0.28
\sum	-	10	1.0000	10.00	-0.00	$10.19 = \chi_B^2$

Таблица 1 Вычисление χ_B^2 при проверке гипотезы H_0 о нормальном законе распределения $N(x, \hat{\mu}, \hat{\sigma})$ для выборки мощностью 100

Мощность выборки = 100

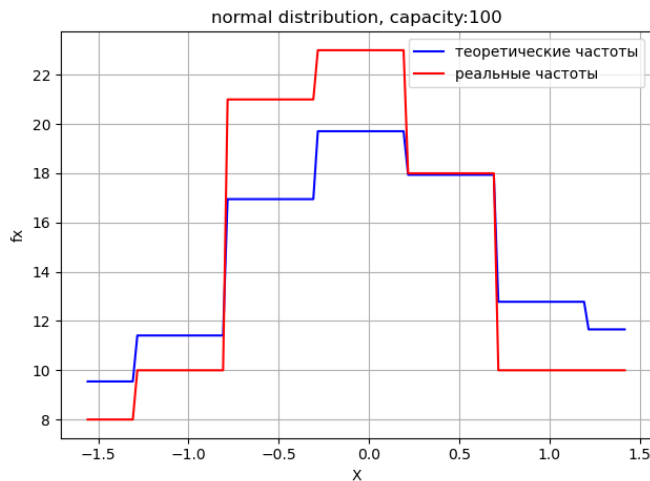


Рисунок 2 График частот для выборки мощностью 10 и график теоретических значений частот на таком же разбиении

Метод максимального правдоподобия:

$$\hat{\mu} \approx -0.06 \quad \hat{\sigma} \approx 0.95$$

Критерий согласия χ^2 :

Количество промежутков $k = 7$

i	Границы Δ_i	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
1	$-\infty, -1.31$	8	0.0955	9.55	-1.55	0.25
2	$-1.31, -0.81$	10	0.1141	11.41	-1.41	0.17
3	$-0.81, -0.31$	21	0.1695	16.95	4.05	0.97
4	$-0.31, 0.19$	23	0.1971	19.71	3.29	0.55
5	$0.19, 0.69$	18	0.1794	17.94	0.06	0.00
6	$0.69, 1.19$	10	0.1278	12.78	-2.78	0.61
7	$1.19, \infty$	10	0.1166	11.66	-1.66	0.24
Σ	-	100	1.0000	100.00	-0.00	$2.78 = \chi_B^2$

Таблица 2 Вычисление χ_B^2 при проверке гипотезы H_0 о нормальном законе распределения $N(x, \hat{\mu}, \hat{\sigma})$ для выборки мощностью 100

Обсуждения

Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат.

Выборка мощностью 10

Графики реальной и теоретически рассчитанной частот довольно похожи.

Проверка гипотезы H_0 о нормальном законе распределения $N(x, \hat{\mu}, \hat{\sigma})$ на уровне значимости $\alpha = 0.05$ не согласуется, так как $\chi_{1-\alpha}^2(k-1) = \chi_{0.95}^2(6) \approx 1.6354 < 10.19 = \chi_B^2$. Подходящий уровень значимости, при котором гипотеза будет согласовываться с выборкой $\alpha = 0.8$.

Выборка мощностью 100

Графики реальной и теоретически рассчитанной частот довольно сильно различаются.

Проверка гипотезы H_0 о нормальном законе распределения $N(x, \hat{\mu}, \hat{\sigma})$ на уровне значимости $\alpha = 0.05$ не согласуется, так как $\chi^2_{1-\alpha}(k-1) = \chi^2_{0.95}(6) \approx 1.6354 < 2.78 = \chi^2_B$. Подходящий уровень значимости, при котором гипотеза будет согласовываться с выборкой $\alpha = 0.1$.

Причиной несовпадения является недостаточное число разбиений. Как видно из таблицы [3, с 358] при увеличении числа степеней свободы увеличивается допустимое χ^2_B . Так же увеличивается допустимое χ^2_B при увеличении мощности выборки, так как при малых n связь $n_1 + n_2 + \dots + n_k = n$ перестаёт быть несущественной.

Литература

1. Ссылка на репозиторий GitHub
<https://github.com/SelyankinFyodor/math-statistics>
2. Вероятностные разделы математики. Учебник для бакалавров технических направлений.
//Под ред. Максимова Ю.Д. – Спб.: Иван Федоров, 2001. – 592 с.Максимов Ю.Д.
Математика. Теория и практика по математической статистике. Конспект-справочник по теории вероятностей : учеб. пособие Ю.Д. Максимов; под ред. В.И. Антонова. — Спб. : Изд-во Политехн. ун-та, 2009. — 395 с. (Математика в политехническом университете).
3. Максимов Ю.Д. Математика. Теория и практика по математической статистике. Конспект-справочник по теории вероятностей: учеб. пособие / Ю. Д. Максимов; под ред. В. И. Антонова. -СПб. : Изд-во Политехн. Ун-та, 2009. -395 с.(Математика в политехническом университете).
4. Квантили распределения хи-квадрат
https://ru.wikipedia.org/wiki/%D0%9A%D0%B2%D0%B0%D0%BD%D1%82%D0%B8%D0%BB%D0%B8_%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D1%8F_%D1%85%D0%B8-%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82