

Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Высшая школа прикладной математики и вычислительной физики

Математическая статистика

Отчёт по лабораторной работе №7-8

Работу

выполнил:

П. П. Филиппов

Группа:

5030102/10101

Преподаватель:

А. Н. Баженов

Санкт-Петербург
2024

Содержание

1. Постановка задачи	3
2. Теория	3
2.1. Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат	3
2.2. Проверка гипотезы однородности выборки	6
2.2.1. F-тест или критерий Фишера	6
2.2.2. Правило Стерджесса	7
3. Результаты	7
3.1. Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат	7
3.2. Проверка гипотезы о равенстве дисперсий двух нормальных генеральных совокупностей	10

1. Постановка задачи

1. Мощность распределения $n=20, 100$.
Распределения: нормальное, Стьюдента ($k=3$), равномерное.
Провести исследование по методике проверки гипотезы о законе распределения по методу χ^2
Результаты вычислений оформить в виде таблицы.
2. Из файла **wave-ampl.txt** выбрать случайно часть данных длиной 1024 отсчёта (всего в файле 800 таких массивов).
Необходимо разделить данные на содержательные подобласти: сигнал, фон, переходные процессы.
3. Мощность нормального распределения $N = 100$.
 - (a) Выбрать две выборки мощностью 20 и 40
 - (b) Выбрать две выборки мощностью 20 и 100

Провести исследование по методу теста Фишера для случаев 1 и 2.

2. Теория

2.1. 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 $\Delta_1, \Delta_2, \dots, \Delta_k$.

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 $p_i = F(a_i) - F(a_{i-1})$; $a_0 = -\infty, a_k = +\infty (i = 1, \dots, k)$

Отметим, что $\sum_{i=1}^k p_i = 1$. Будем предполагать, что все $p_i > 0 (i = 1, \dots, k)$

Пусть, далее, n_1, n_2, \dots, n_k — частоты попадания выборочных элементов в подмножества $\Delta_1, \Delta_2, \dots, \Delta_k$ 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

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

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

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

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

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

Для прояснения сущности метода χ^2 сделаем ряд замечаний.

Замечание 1 Выбор подмножеств $\Delta_1, \Delta_2, \dots, \Delta_k$ 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

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

или формулой Старджесса

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

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 уровень значимости α
2. Находим квантиль $\chi_{1-\alpha}^2(k-1)$ 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. Вычисляем выборочное значение статистики критерия χ^2 :

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

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

Замечание 3. Веса $c_i = \frac{n}{p_i}$ пропорциональны n , т.е. с ростом n увеличиваются. Отсюда следует, что если выдвинутая гипотеза неверна, то относительные частоты $\frac{n}{p_i}$ не будут близки к вероятностям p_i , и с ростом n величина χ_B^2 будет увеличиваться. При фиксированном уровне значимости α будет фиксировано пороговое число — квантиль $\chi_{1-\alpha}^2(k-1)$, поэтому, увеличивая n , мы придём к неравенству $\chi_B^2 > \chi_{1-\alpha}^2(k-1)$, т.е. с увеличением объёма выборки неверная гипотеза будет отвергнута.

Отсюда следует, что при сомнительной ситуации, когда $\chi_B^2 \approx \chi_{1-\alpha}^2(k-1)$, можно попытаться увеличить объём 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

$$\frac{n_i - np_i}{\sqrt{np_i}}$$

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

2.2. Проверка гипотезы однородности выборки

Если выборка близка к нормальной, можно проверить, насколько она однородна.

2.2.1. F-тест или критерий Фишера

F-тест или критерий Фишера (F-критерий, ϕ^* -критерий) — статистический критерий, тестовая статистика которого при выполнении нулевой гипотезы имеет распределение Фишера (F-распределение). В общем случае сравниваются две выборки А и В и проверяется равенство их дисперсий:

$$F = \frac{\varsigma_A^2}{\varsigma_D^2}$$

Нулевая гипотеза $H_0: \varsigma_A^2 = \varsigma_B^2$

Статистика теста так или иначе сводится к отношению выборочных дисперсий (сумм квадратов, деленных на «степени свободы»). Чтобы статистика имела распределение Фишера, необходимо, чтобы числитель и знаменатель были независимыми случайными величинами и соответствующие суммы квадратов имели распределение χ^2 . Для этого требуется, чтобы данные имели нормальное распределение.

Статистика критерия Фишера (1.6) имеет распределение Фишера с $n - 1$ и $m - 1$ степенями свободы $F(n - 1, m - 1)$. Обычно в числителе ставится большая из двух сравниваемых дисперсий.

Применим тест Фишера для выяснения однородности конкретной выборки. Для этого дисперсию выборки можно оценить двумя способами.

Во-первых, дисперсия, вычисленная для каждой группы — это оценка дисперсии совокупности. Поэтому дисперсию совокупности можно оценить на основании групповых дисперсий. Такая оценка не будет зависеть от различий групповых средних.

Во-вторых, разброс выборочных средних тоже позволяет оценить дисперсию совокупности. Понятно, что такая оценка дисперсии зависит от различий выборочных средних. В качестве оценки дисперсии совокупности возьмем среднее выборочных дисперсий. Эта оценка называется внутригрупповой дисперсией. Обозначим ее s_{in}^2

$$s_{in}^2 = \frac{1}{k} \sum_{i=1}^k s_i^2 = \frac{1}{k} \sum_{i=1}^k \frac{\sum_{j=1}^n (x_{ij} - \hat{X})^2}{k - 1}$$

где \hat{X} — среднее для части выборки, k — количество частей, на которое делим сходную выборку, n — количество элементов в подвыборке.

Также нужно вычислить межгрупповую дисперсию. Обозначается она s_{out}^2 .

Вычисление межгрупповой дисперсии происходит в несколько этапов:

1. Вычисление среднего значения для всех выбранных подвыборок $(\overline{X}_1, \dots, \overline{X}_k)$
2. Вычисление среднего этих средних: $\overline{X} = \frac{1}{k} \sum_{i=1}^k \overline{X}_i$
3. Вычисление межгрупповой дисперсии:

$$s_{out}^2 = k \cdot \frac{\sum_{j=1}^n (\overline{X}_i - \overline{X})^2}{k - 1}$$

Окончательно имеем для проверки гипотезы

$$F = \frac{s_{out}^2}{s_{in}^2}$$

Если значение $F = \frac{s_{out}^2}{s_{in}^2} \approx 1$, выборку можно считать однородной

2.2.2. Правило Стерджесса

Данное правило используется для определения оптимального количества интервалов, на которые разбивается наблюдаемый диапазон изменения случайной величины при изучении ее распределения. По этому правилу число интервалов считается по формуле:

$$n = 1 + \log_2 N$$

где n — число интервалов, N — общее число наблюдений.

3. Результаты

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

Пределы	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
$(-2.5687, -1.5801)$	3	0.0519	1.0385	1.9614	3.7044
$(-1.5801, -0.5916)$	7	0.2200	4.4002	2.5997	1.5359
$(-0.5916, 0.3969)$	4	0.3772	7.5447	-3.5447	1.6654
$(0.3969, 1.3854)$	5	0.2627	5.2552	-0.2552	0.0123
$(1.3854, 2.3740)$	1	0.0741	1.4831	-0.4831	0.1573

Таблица 3.1

Результаты проверки гипотезы нормального распределения $n = 20$

$\chi_B^2 = 7.0756$, $\chi_{1-\alpha}^2(k-1) = 9.4877$, гипотеза не отвергнута.

Пределы	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
$(-2.6787, -1.8684)$	5	0.0271	2.7153	2.2846	1.9222
$(-1.8684, -1.0582)$	12	0.1141	11.4130	0.5869	0.0301
$(-1.0582, -0.2479)$	22	0.2571	25.7101	-3.7101	0.5353
$(-0.2479, 0.5622)$	27	0.3109	31.0957	-4.0957	0.5394
$(0.5622, 1.3725)$	25	0.2020	20.2012	4.7987	1.1399
$(1.3725, 2.1827)$	8	0.0704	7.0423	0.9576	0.1302
$(2.1827, 2.9930)$	0	0.0131	1.3145	-1.3145	1.3145
$(2.9930, 3.8032)$	1	0.0013	0.1309	0.8690	5.7660

Таблица 3.2

Результаты проверки гипотезы нормального распределения $n = 100$

$\chi_B^2 = 11.3779$, $\chi_{1-\alpha}^2(k-1) = 14.0671$, гипотеза не отвергнута.

Пределы	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
$(-2.3720, -1.5223)$	2	0.0598	1.1975	0.8024	0.5377
$(-1.5223, -0.6726)$	3	0.1787	3.5750	-0.5750	0.0924
$(-0.6726, 0.1770)$	11	0.3102	6.2056	4.7943	3.7039
$(0.1770, 1.0266)$	2	0.2671	5.3427	-3.3427	2.0914
$(1.0266, 1.8763)$	2	0.1193	2.3869	-0.3869	0.0627

Таблица 3.3

Результаты проверки гипотезы распределения Стьюдента $n = 20$

$\chi_B^2 = 6.4882$, $\chi_{1-\alpha}^2(k-1) = 9.4877$, гипотеза не отвергнута.

Пределы	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
$(-3.7782, -2.8816)$	1	0.0063	0.6363	0.3636	0.2078
$(-2.8816, -1.9850)$	1	0.0294	2.9447	-1.9447	1.2843
$(-1.9850, -1.0885)$	7	0.1133	11.3338	-4.3338	1.6571
$(-1.0885, -0.1919)$	31	0.2748	27.4867	3.5132	0.4490
$(-0.1919, 0.7046)$	26	0.3256	32.5625	-6.5625	1.3225
$(0.7046, 1.6012)$	24	0.1783	17.8349	6.1650	2.1311
$(1.6012, 2.4978)$	9	0.0544	5.4420	3.5579	2.3260
$(2.4978, 3.3944)$	1	0.0123	1.2364	-0.2364	0.0452

Таблица 3.4

Результаты проверки гипотезы распределения Стьюдента $n = 100$

$\chi_B^2 = 9.4234$, $\chi_{1-\alpha}^2(k-1) = 14.0671$, гипотеза не отвергнута.

Пределы	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
(0.0029, 0.1950)	5	0.1920	3.8416	1.1583	0.3492
(0.1950, 0.3871)	1	0.1920	3.8416	-2.8416	2.1019
(0.3871, 0.5791)	0	0.1920	3.8416	-3.8416	3.8416
(0.5791, 0.7712)	6	0.1920	3.8416	2.1583	1.2125
(0.7712, 0.9633)	8	0.1920	3.8416	4.1583	4.5010

Таблица 3.5

Результаты проверки гипотезы равномерного распределения $n = 20$

$\chi_B^2 = 12.0065$, $\chi_{1-\alpha}^2(k-1) = 9.4877$, гипотеза отвергнута.

Пределы	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i - np_i)^2}{np_i}$
(0.0101, 0.1331)	13	0.1230	12.3026	0.6973	0.0395
(0.1331, 0.2561)	9	0.1230	12.3026	-3.3026	0.8865
(0.2561, 0.3792)	14	0.1230	12.3026	1.6973	0.2341
(0.3792, 0.5022)	12	0.1230	12.3026	-0.3026	0.0074
(0.5022, 0.6252)	9	0.1230	12.3026	-3.3026	0.8865
(0.6252, 0.7482)	17	0.1230	12.3026	4.6973	1.7935
(0.7482, 0.8713)	12	0.1230	12.3026	-0.3026	0.0074
(0.8713, 0.9943)	14	0.1230	12.3026	1.6973	0.2341

Таблица 3.6

Результаты проверки гипотезы равномерного распределения $n = 100$

$\chi_B^2 = 4.0895$, $\chi_{1-\alpha}^2(k-1) = 14.0671$, гипотеза не отвергнута.

3.2. Проверка гипотезы о равенстве дисперсий двух нормальных генеральных совокупностей

	s_X^2	s_Y^2	F_B	$F_{1-\frac{\alpha}{2}}(k_1, k_2)$	Результат
205 и 410	13590.3779	12179.6046	1.1158	1.2627	Не отвергается
205 и 1024	13590.3779	15280.5865	0.8893	1.2279	Не отвергается

Таблица 3.7

Результаты теста Фишера