значения критических точек для F распределения можно найти также и в интернете.

2.6. Объяснение критерия Фишера. Можно показать, что при выполнении гипотезы H_0 обе величины $\mathrm{SS}_{\mathrm{TR}}/(k-1)$ и $\mathrm{SS}_{\mathrm{E}}/(n-k)$ являются “хорошими” статистическими оценками теоретической дисперсии σ^2. Это, в частности означает, что их отношение близко к 1. Если же H_0 не верна, то $\mathrm{SS}_{\mathrm{TR}}/(k-1)$ с большой вероятностью становится больше $\mathrm{SS}_{\mathrm{E}}/(n-k)$, так как различия между группами становятся доминирующими. Таким образом, большие значения отношения F в (2) свидетельствуют о том, что гипотеза H_0 не верна. Критическая величина F, при котором “доверие” к H_0 становится слишком низким, равна $F_{\alpha;k-1,n-k}$. Степень достоверности принятия такого решения определяется уровнем α.

Замечание 4. Хотя приведенные выше сведения о критерии Фишера и распределении Фишера относятся к конкретным значениям $k = 4$, $n_1 = 65$, $n_2 = 56$, $n_3 = 52$ и $n_4 = 76$, все рассуждения, формулы и результаты остаются справедливыми и для произвольного количества групп и произвольного количества наблюдений в каждой группе.

3. Проверка гипотезы для исследования Льюиса

Вернемся к исследованию Льюиса.

Нашей задачей является проверка гипотезы о том, что профилактический прием витамина С **не** снижает риск простуды и **не** уменьшает длительность заболевания (см. замечание 1).

К сожалению необходимые для этого данные — длительность j-ого заболевания в i-ой группе X_{ij} — в отчете Льюиса и его коллег не приведены. Поэтому проверить указанную гипотезу с помощью критерия Фишера не представляется возможным. Более внимательный анализ отчета Льюиса все же показывает, что в нем содержится немного больше информации, чем в таблице 2. В оригинальной таблице Льюиса средние длительности простуд приведены для каждой группы в виде $\overline{X}_i \pm \widehat{\sigma}_i$, то есть с точностью до одного *среднеквадратического отклонения* $\widehat{\sigma}_i$, которое определяется как корень кадратный из выборочной дисперсии $\widehat{\sigma}_i^2$:

$$\widehat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_i\right)^2, \qquad 1 \le i \le 4. \tag{5}$$

Заметим, что сумма величин $(n_i - 1)\widehat{\sigma}_i^2$ по $i = 1, 2, 3, 4$ обозначалась выше символом $\mathrm{SS_{TO}}$.

Таким образом мы фактически обладаем следующей информацией:

Т а б л и ц а 4
Эмпирические характеристики

	Группа 1	Группа 2	Группа 3	Группа 4
n_i	65	56	52	76
$\overline{X}_i$	7.1	6.5	6.7	5.9
$\widehat{\sigma}_i$	0.46	0.39	0.53	0.40
$\widehat{\sigma}_i^2$	0.21	0.15	0.28	0.16

В этой таблице n_i — это количество простудных заболеваний в i-ой группе, $\overline{X}_i$ — средняя длительность заболеваний, $\widehat{\sigma}_i^2$ — эмпирическая дисперсия, вычисленная по формуле (5).

Все же и этой информации недостаточно, чтобы вычислить суммы квадратов $\mathrm{SS_{TO}}$, $\mathrm{SS_{TR}}$ и $\mathrm{SS_E}$. Чтобы решить поставленную задачу проверки гипотезы, мы поступим следующим образом: сначала для каждой группы i смоделируем на компьютере n_i наблюдений с распределением $\mathcal{N}\left(\overline{X}_i, \widehat{\sigma}_i^2\right)$. Затем используем полученные данные для проверки гипотезы.

Замечание 5. Для того, чтобы обоснованно пользоваться критерием Фишера, необходимо быть уверенным в истинности предположений 1–5. Предположение 4 требует, чтобы дисперсии были одинаковыми. В таблице 4 наибольшая дисперсия отвечает группе 3 ($\sigma_3^2 = 0.281$), а наименьшая — группе 2 ($\sigma_2^2 = 0.152$). Их отношение приблизительно равно 1.85. Хотя дисперсии отличаются почти в 2 раза, можно допустить что различие объясняется статистической изменчивостью данных. Гипотезу о равенстве дисперсий нескольких независимых гауссовских выборок можно проверить с помощью *критерия Хартли*, который в нашем случае не отвергает ее на уровне доверия 0.95. В дальнейшем считаем, что это является достаточным основанием справедливости предположения 4.

3.1. Моделирование. Необходимые данные сгенерируем с помощью команды `СЛЧИС()` программы `Excel`, которая генерирует значение случайной величины $\mathbf{U}$, имеющей равномерное распределение на отрезке $(0,1)$. Ниже приведен алгоритм генерации одного значения случайной величины $\mathbf{G}_{a,b^2}$, имеющей распределение $\mathcal{N}\left(a, b^2\right)$:

1. сгенерировать 12 значений $U_1, \ldots, U_{12}$ случайной величины $\mathbf{U}$;
2. вычислить $G_{0,1} = b(U_1 + \cdots + U_{12} - 6 + a)$.

Результаты моделирования представлены в таблице 5, вид которой полностью идентичен таблице 3. Для генерации i-го столбца таблицы 5 указанная процедура применялась для параметров $a = \overline{X}_i$ и $b^2 = \widehat{\sigma}_i^2$, взятых из таблицы 4. Результаты моделирования были бы более адекватными, если бы оно проводилось не для эмпирических значений среднего $\overline{X}_i$ и дисперсии $\widehat{\sigma}_i^2$, а для их теоретических значений μ_i и σ_i^2, но последние нам не известны.

Чтобы различать данные эксперимента Льюиса и сгенерированные данные, мы обозначаем последние через Y_{ij}.

Результаты всех вычислений с данными $\{Y_{ij}\}$ также будут содержать букву Y. Например, $\overline{Y}_1$, $\overline{Y}_2$, $\overline{Y}_3$ и $\overline{Y}_4$ обозначают средние по группам, а $\overline{Y}$ — общее среднее для всей таблицы.

3.2. Проверка гипотезы. Для подсчетов удобно пользоваться следующими обозначениями и формулами (хотя мы по-прежнему используем значение $k = 4$, формулы ниже справедливы и для другого количества групп):

$n_1, \dots, n_k$ количества наблюдений в группах,
$T_1, \dots, T_k$ суммы наблюдений по группам,

$$n = n_1 + \dots + n_k,$$
$$T = T_1 + \dots + T_k,$$
$$c \stackrel{\mathrm{def}}{=} \frac{T^2}{n}.$$

Тогда

$$\mathrm{SS_{TO}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}^2 - c, \tag{6}$$

$$\mathrm{SS_{TR}} = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - c, \tag{7}$$

$$\mathrm{SS_E} = \mathrm{SS_{TO}} - \mathrm{SS_{TR}}\,. \tag{8}$$

Для данных из таблицы 5 имеем:

$$n_1 = 65, \qquad n_2 = 56, \qquad n_3 = 52, \qquad n_4 = 76,$$
$$T_1 = 466.18, \quad T_2 = 362.44, \quad T_3 = 351.45, \quad T_4 = 451.20$$
$$n = 249, \qquad T = 1631.27, \qquad c = 10686.91,$$
$$\mathrm{SS_{TO}} = 105.38, \qquad \mathrm{SS_{TR}} = 56.32, \qquad \mathrm{SS_E} = 49.05.$$

Т а б л и ц а 5

Сгенерированные данные

j	$\mathbf{Y_{1j}}$	$\mathbf{Y_{2j}}$	$\mathbf{Y_{3j}}$	$\mathbf{Y_{4j}}$	j	$\mathbf{Y_{1j}}$	$\mathbf{Y_{2j}}$	$\mathbf{Y_{3j}}$	$\mathbf{Y_{4j}}$
1.	7,72	6,56	6,21	6,40	39.	7,92	6,35	6,92	5,79
2.	7,76	6,19	6,83	6,07	40.	7,29	6,62	8,01	6,32
3.	7,36	6,11	7,75	6,06	41.	7,98	5,98	7,07	5,45
4.	7,41	6,95	6,18	6,37	42.	6,94	6,45	7,16	6,46
5.	7,28	6,84	6,87	6,04	43.	7,64	6,73	7,53	6,16
6.	7,03	5,98	6,43	6,90	44.	7,10	6,55	6,65	5,83
7.	7,51	6,57	6,86	5,81	45.	6,59	6,26	6,57	5,51
8.	7,69	6,78	6,82	5,80	46.	7,23	6,55	5,55	5,49
9.	7,50	7,53	7,00	5,56	47.	6,38	6,47	7,36	5,85
10.	7,12	6,24	6,03	5,15	48.	7,34	5,90	7,27	6,89
11.	7,71	5,87	6,75	5,74	49.	6,74	6,09	5,86	6,41
12.	7,00	6,57	7,15	6,48	50.	6,97	6,54	6,50	6,40
13.	7,80	6,09	7,09	6,70	51.	6,94	6,47	6,55	5,64
14.	6,10	6,12	6,82	5,72	52.	7,47	6,83	6,89	5,35
15.	7,63	7,18	7,22	5,45	53.	6,69	7,31		6,25
16.	7,11	7,07	6,44	5,96	54.	6,77	6,76		6,38
17.	7,55	5,97	6,85	5,68	55.	7,34	6,24		5,72
18.	7,88	5,94	6,60	6,16	56.	6,78	5,69		5,46
19.	7,66	6,36	6,36	5,48	57.	7,16			5,65
20.	6,75	7,50	7,91	5,72	58.	6,47			6,36
21.	6,51	7,16	7,05	5,75	59.	7,12			6,38
22.	7,06	6,32	6,83	5,83	60.	7,12			5,95
23.	6,74	6,22	5,52	5,90	61.	7,73			6,86
24.	7,27	6,44	6,54	5,62	62.	7,08			5,99
25.	7,70	6,89	6,55	5,82	63.	6,87			5,08
26.	6,35	6,41	6,96	6,12	64.	7,12			5,63
27.	7,62	6,02	6,93	5,79	65.	6,91			5,44
28.	7,12	6,50	6,85	5,33	66.				6,04
29.	7,09	7,06	6,03	5,59	67.				5,53
30.	7,38	5,54	6,19	5,65	68.				6,47
31.	7,27	6,28	6,84	6,67	69.				5,94
32.	7,09	6,90	6,04	5,60	70.				5,60
33.	7,08	6,46	7,22	6,43	71.				6,13
34.	6,83	6,35	7,32	5,59	72.				6,18
35.	7,02	6,26	6,72	5,89	73.				6,80
36.	7,23	6,68	6,91	5,74	74.				5,48
37.	6,44	6,06	6,22	5,66	75.				6,34
38.	7,12	6,68	6,67	5,91	76.				5,85

Для проверки гипотезы

$\mathbf{H_0}$: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

о том, что прием витамина C не оказывает влияния на продолжительность простуд, выберем достаточно высокий уровень доверия $\alpha = 0.99$.

Значение критической точки $F_{0.99;3,245}$ в таблицах отсутствует. Однако ее значение находится между $F_{0.99;3,120}$ и $F_{0.99;3,\infty}$, а эти квантили в таблицах имеются. Таким образом, $3.78 < F_{0.99;3,245} < 3.95$. По формуле (2) вычисляем значение тестовой статистики:

$$F = 93.77.$$

Поскольку $F \geq F_{0.99;3,245}$, то согласно критерию Фишера §2.5, гипотеза H_0 отвергается в пользу альтернативы, которая состоит в том, что не все теоретические средние равны между собой. Это свидетельствует о том, что в эксперименте присутствует хотя-бы один фактор, который делает результаты разными в разных группах.

Более сильным выводом после проведенного дисперсионного анализа является утверждение о том, профилактический прием витамина C уменьшант длительность простуд.

УПРАЖНЕНИЯ

Упражнение 1. Доказать, что случайные величины ε_{ij} в (1) независимы, причем $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Упражнение 2. Интегрируя плотность распределения Фишера, доказать формулы (3) и (4).

Упражнение 3. Объяснить почему реализации нормальной случайной величины можно генерировать с помощью способа, описанного выше в §3.1.

Упражнение 4. Доказать равенства (6)–(8).