Оглавление

1. Основные принципы планирования эксперимента

Под статистическим планированием эксперимента понимается организация экспериментального исследования, которая позволит собрать необходимые данные, применить для их анализа статистические методы и сделать правильные и объективные выводы. Без статистического подхода к планированию эксперимента не обойтись.

Если данные эксперимента содержат ошибки, то статистические методы являются единственным объективным подходом к их анализу. Таким образом, в любой экспериментальной задаче два аспекта: планирование эксперимента и статистический анализ данных, причем эти два аспекта тесно взаимосвязаны, так как метод анализа непосредственно зависит от использованного плана.

В основе планирования эксперимента лежат два принципа – репликация и рандомизация. Под репликацией понимают повторение основного эксперимента. Рандомизация – краеугольный камень, на котором основано применение статистических методов в планировании эксперимента. Рандомизация означает, что распределение экспериментального материала и порядок, в котором должны проводиться отдельные опыты или прогоны, устанавливаются случайным образом.

При использовании статистического подхода к планированию экспериментов и анализу данных необходимо, чтобы все участники эксперимента еще до его начала ясно понимали, что именно предстоит исследовать и каким образом нужно собирать данные. Можно рекомендовать следующую схему:

1. Признание факта существования задачи и ее формулировка.
2. Выбор факторов и уровней.
3. Выбор переменной отклика (зависимой переменной).
4. Выбор плана эксперимента.
5. Проведение эксперимента
6. Анализ данных.
7. Выводы и рекомендации.

2. Дисперсионный анализ

Инициатором применения статистических методов в планировании экспериментов является Рональд А. Фишер. В течение нескольких лет он был ответственным за статистическую обработку данных в Лондоне. Фишер разработал и впервые применил дисперсионный анализ в качестве важнейшего метода статистического анализа в планировании экспериментов.

Методы планирования эксперимента впервые начали использовать в сельскохозяйственных и биологических науках. Современные методы планирования экспериментов сегодня широко применяются во всех областях исследований: агрономии, медицине, биологии, прикладных, естественных и общественных науках и др.

Дисперсионный анализ – статистический метод, позволяющий анализировать влияние различных факторов (категориальных, группирующих, независимых переменных), обозначаемых латинскими буквами *A*,*B*,*C* и т. д., на результаты эксперимента (зависимые переменные). Для проведения дисперсионного анализа необходимо, чтобы независимая переменная была категориальной, а зависимая – метрической. Например, факторами, влияющими на содержание микроэлементов в пробе, могут быть: *A* – метод геохимического анализа, *B* – территория, *C* – среда съёмки (почва, снег, зола, накипь).

В этом случае говорят о применении 3-х факторного дисперсионного анализа для исследования влияния 3-х факторов (*A* – метод геохимического анализа с 2-мя уровнями; *B* – территория с 3-мя уровнями и *C* – среда съемки с 4-мя уровнями) на содержание микроэлементов в пробе.

Суть дисперсионного анализа (*analysis of variance* – сокращенно *ANOVA*) заключается в разложении дисперсии измеряемого признака на независимые слагаемые, каждое из которых характеризует влияние того или иного фактора или их взаимодействия. Последующее сравнение таких слагаемых позволяет оценить значимость каждого изучаемого фактора, а также их комбинации.

Анализ основан на расчете ***F-статистики*** (статистика Фише-ра), которая представляет собой отношение двух *дисперсий*: межгрупповой и внутригрупповой. ***F***-тест в однофакторном дисперсионном анализе определяет, значимо ли различаются средние нескольких независимых выборок.

Задача дисперсионного анализа состоит в том, чтобы из общей вариативности признака вычленить вариативность иного рода:

а) вариативность обусловленную действием каждой из исследуемых независимых переменных;

б) вариативность, обусловленную взаимодействием исследуемых независимых переменных;

в) случайную вариативность, обусловленную всеми другими неизвестными переменными.

Чем в большей степени вариативность признака обусловлена исследуемыми переменными (факторами) или их взаимодействием, тем выше **эмпирические значения критерия**.

Нулевая гипотеза в дисперсионном анализе будет гласить, что средние величины исследуемого результативного признака во всех градациях одинаковы.

Альтернативная гипотеза будет утверждать, что средние величины результативного признака в разных градациях исследуемого фактора различны.

Ограничения метода однофакторного дисперсионного анализа для несвязанных выборок:

1. Однофакторный дисперсионный анализ требует не менее трех градаций фактора и не менее двух испытуемых в каждой градации.

2. Результативный признак должен быть нормально распределен в исследуемой выборке.

Правда, обычно не указывается, идет ли речь о распределении признака во всей обследованной выборке или в той ее части, которая составляет дисперсионный комплекс.

3. Однофакторный дисперсионный анализ (при одинаковом числе испытаний на всех уровнях)

Пусть на количественный нормально распределенный признак X воздействует фактор F, который имеет p постоянных уровней F_1, F_2,…,F_p. На каждом уровне произведено по q испытаний. Результаты наблюдений – числа x_{ij}, где i-номер испытания (i=1,2,…,q), j – номер уровня фактора (j=1,2,…,p) записываются в виде таблицы (таблица 1)

Таблица 1

Номер испытания	Уровни фактора			
i	F_1	F_2	…	F_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
…	…	…	…	…
q	x_{q1}	x_{q2}	…	x_{qp}
Групповая средняя $\overline{x_{\text{гр}}}$	$\overline{x_{\text{гр}1}}$	$\overline{x_{\text{гр}2}}$	…	$\overline{x_{\text{гр}\,p}}$

Ставится задача: на уровне значимости α проверить нулевую гипотезу о равенстве групповых средних при допущении, что групповые генеральные дисперсии хотя и независимы, но одинаковы. Для решения этой задачи вводятся:

- общая сумма квадратов отклонений наблюдаемых значений признака от общей средней

$$S_{\text{общ}} = \sum_{j=1}^{p}\sum_{i=1}^{q}\left(x_{ij} - \bar{x}\right)^2 \qquad (1)$$

- факторная сумма квадратов отклонений групповых средних от общей средней (характеризует рассеяние «между группами»)

$$S_{\text{факт}} = q\sum_{j=1}^{p}(\overline{x_{\text{гр}\,J}} - \bar{x})^2 \qquad (2)$$

- остаточная сумма квадратов отклонений наблюдаемых значений группы от всей групповой средней (характеризует рассеяние «внутри групп»)

$$S_{\text{ост}} = \sum_{i=1}^{q}(x_{i1} - \overline{x_{\text{гр}1}})^2 + \cdots + \sum_{i=1}^{q}(x_{ip} - \overline{x_{\text{гр}\,p}})^2 \qquad (3)$$

Практически остаточную сумму находят по формуле: $S_{ост}=S_{общ} - S_{факт}$ (4)

Для вычисления общей и факторной сумм более удобны следующие формулы:

$$S_{\text{общ}} = \sum_{j=1}^{p} P_j - \frac{(\sum_{j=1}^{p} R_j)^2}{pq}$$

$$S_{\text{факт}} = \frac{\sum_{j=1}^{p} R_j^2}{q} - \frac{(\sum_{j=1}^{p} R_j)^2}{pq} \qquad (5)$$

где$P_j = \sum_{i=1}^{q} x_{ij}^2$– сумма квадратов наблюдаемых значений признака на уровне F_j

$R_j = \sum_{i=1}^{q} x_{ij}$ - сумма наблюдаемых значений признака на уровне F_j

Если наблюдаемые значения признака – сравнительно большие числа, то для упрощения вычислений вычитают из каждого наблюдаемого значения одно и то же число C, примерно равное общей средней. Если уменьшенные значения $y_{ij} = x_{ij} - C$

$S_{\text{общ}} = \sum_{j=1}^{p} Q_j - \frac{(\sum_{j=1}^{p} T_j)^2}{pq}$, $S_{\text{факт}} = \frac{\sum_{j=1}^{p} T_j^2}{q} - \frac{(\sum_{j=1}^{p} T_j)^2}{pq}$

где$Q_j = \sum_{i=1}^{q} y_{ij}^2$– сумма квадратов уменьшенных наблюдаемых значений признака на уровне F_j

$T_j = \sum_{i=1}^{q} y_{ij}$ - сумма уменьшенных значений признака на уровне F_j.

Разделив уже вычисленные факторную и остаточную суммы на соответствующее число степеней свободы, находят факторную и остаточную дисперсии:

$S_{\text{факт}}^2 = \frac{S_{\text{факт}}}{p-1}$, $S_{\text{ост}}^2 = \frac{S_{\text{ост}}}{p(q-1)}$, $F_{\text{набл}} = \frac{S_{\text{факт}}^2}{S_{\text{ост}}^2}$, $F_{\text{кр}} = \left(\alpha; p-1; p(q-1)\right)$

Наконец, сравнивая, фактурную и остаточную дисперсии по критерию Фишера-Снедекора (приложение 1).

Если $F_{набл} < F_{кр}$ – различие групповых средних незначимое.

Если $F_{набл} > F_{кр}$ – различие групповых средних значимое.

Замечание 1. Если факторная дисперсия окажется меньше остаточной, то уже отсюда непосредственно следует справедливость нулевой гипотезы о равенстве групповых средних, поэтому дальнейшие вычисления (сравнение дисперсий с помощью критерия F) излишни.

Замечание 2. Если наблюдаемые значения x_{ij} – десятичные дроби с kзнаками после запятой, то целесообразно перейти к целым числам $y_{ij} = 10^k x_{ij} - C$, где C–примерно среднее значение чисел $10^k x_{ij}$. При этом факторная и остаточная дисперсия увеличится в 10^{2k} раз, однако их отношение не изменится.

Рассмотрим задачу: При уровне значимости α=0,05 методом дисперсионного анализа проверить нулевую гипотезу о влиянии фактора на качество объекта на основании пяти измерений для трех уровней фактора

Номер измерения	Φ_1	Φ_2	Φ_3
1.	18	24	36
2.	28	36	12
3.	12	28	22
4.	14	40	45
5.	32	16	40

Решение: сформулируем гипотезы H_0 – фактор влияет на качество объекта незначительно, тогда H_1 – фактор оказывает влияние на качество объекта.

Вычислим вспомогательные величины:

$P_j = \sum x_{ij}^2$ -сумма квадратов наблюдаемых значений на уровне Φ_j(j=1,2,3)

$R_j = \sum x_{ij}$ - сумма наблюдаемых значений на уровне Φ_j(j=1,2,3)

Результаты занесем в таблицу:

Номер измерения	Φ_1	Φ_2	Φ_3	
1.	18	24	36	
2.	28	36	12	
3.	12	28	22	
4.	14	40	45	
5.	32	16	40	
				сумма
R	104	144	155	403
P	2472	4512	5549	12533
R^2	10816	20736	24025	55577

Тогда $S_{общ}$=12533- (1/3*5)*403^2=1705,7

$S_{факт.}$=(1/5)*55577-(1/3*5)*403^2=288,1

$S_{ост}=S_{общ} - S_{факт}$ =1704,7-288,1=1417,6

Найдем факторную дисперсию $S^2_{факт}=S_{факт}/3-1$=144,05

Найдем остаточную дисперсию $S^2_{ост} =S_{ост}/3*(5-1)$=118,13

Сравним факторную и остаточную дисперсию по критерию Фишера: найдем наблюдаемое значение критерия $F_{набл}=S^2_{факт}/S^2_{ост}$=1,22

Найдем критическую точку при уровне значимости α=0,05 и числам степеней свободы k_1=3-1=2 k_2=3*(5-1)=12 по таблице определили, что $F_{кр}$=3,88 делаем вывод, т.к. $F_{набл} < F_{кр}$ - нет оснований отвергать нулевую гипотезу (фактор влияет незначительно).

Задачи для решения на практическом занятии:

3.1.Произведено по четыре испытания на каждом из трех уровней фактора F. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты представлены в таблице

Номер испытания	Уровни фактора		
	F_1	F_2	F_3
1.	38	20	21
2.	36	24	22
3.	35	26	31
4.	31	30	34
$\overline{x\_\text{гр}}j$	35	25	27

Указание: для упрощения расчета из каждого наблюдаемого значения x_{ij}общую среднюю $\bar{x} = 29$, то есть перейти к уменьшаемым величинам: $y_{ij} = x_{ij} - 29$.

3.2.Произведено по восемь испытаний на каждом из шести уровней фактора. Методом дисперсионного анализа при уровне значимости 0,01 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты представлены в таблице

Номер испытания	Уровни фактора					
	F_1	F_2	F_3	F_4	F_5	F_6
1.	100	92	74	68	64	69
2.	101	102	87	80	83	71
3.	126	104	88	83	83	80
4.	128	115	93	87	84	80
5.	133	119	94	96	90	81
6.	141	122	101	97	96	82
7.	147	128	102	106	101	86
8.	148	146	105	127	111	99
$\overline{x\_\text{гр}}j$	128	116	93	93	89	81

При решении задачи использовать указание задачи 3.1.

3.3.Имеются результаты определения средней высоты сосны напробных площадях в разных условиях местопроизрастания (таблица)

Вариант опыта (ТУМ)	Средняя высота на пробных площадках, м
Лишайниковый	18,5; 17,5; 18; 18
Брусничный	18,5; 18; 18; 20
Черничный	19,5; 20; 20,5; 19,5
Кисличный	20; 20,5; 22; 21

Пробные площади были заложены таким образом, чтобы исключить влияние прочих факторов на результативный признак (одинаковый средний возраст, подзона тайги и т.д.). В ходе исследования предстоит установить, влияет ли тип условий место произрастания (ТУМ) на рост насаждений сосны.

3.4.Получены данные о плодовитости мышей при облучении рентгеновскими лучами:

Группы	Число мышат от отдельных самок			
Контроль	10	12	11	10
Доза 100 р.	8	10	7	9
Доза 200 р.	7	9	6	4

Влияет ли облучение на плодовитость мышей?

3.5.Проверьте влияет ли возраст на частоту распространенности изолированной систолической артериальной гипертензии в различных регионах России (в %):

Регионы	Возраст (лет)			
	50	60	70	80
1	24	47	66	73
2	23	45	60	70
3	21	43	65	72
4	25	42	65	71
5	23	46	65	73

3.6.Исследовать влияние породы животных на уровень их иммунитета. Животные трех пород в возрасте 31 месяц искусственно заражали одинаковым количеством личинок *Boophilus micropolus* и через 20 дней подсчитывали число самок клещей:

Номер животного	Порода животного		
	Африкандер-герефорд	Шортгорны	Герефорды
1	20	50	100
2	40	170	400
3	70	210	570
4	120	450	840
5	240	610	1200

3.7.Проверьте влияет ли уровень холестерина в крови на смертность от ишемической болезни сердца в различных регионах России (на 10 000 населения):

Регионы	Сывороточный холестерин, ммоль/л			
	4	5	6	7
1	9	12	17	28
2	8	13	16	27
3	9	13	18	27
4	7	14	17	26
5	8	14	17	29

3.8. Три различные группы из шести испытуемых получили списки из десяти слов. Первой группе слова предъявлялись с низкой скоростью -1 слово в 5 секунд, второй группе со средней скоростью - 1 слово в 2 секунды, и третьей группе с большой скоростью - 1 слово в секунду. Было предсказано, что показатели воспроизведения будут зависеть от скорости предъявления слов. Результаты представлены в таблице.Проверить предположение.

№ испытуемого	Группа 1: низкая скорость	Группа 2: средняя скорость	Группа 3: высокая скорость
1	8	7	4
2	7	8	5
3	9	5	3
4	5	4	6
5	6	6	2
6	8	7	4

3.9.Группа из 5 испытуемых была обследована с помощью трех экспериментальных заданий, направленных на изучение интеллектуальной, настойчивости. Каждому испытуемому индивидуально предъявлялись последовательно три одинаковые анаграммы: четырехбуквенная, пятибуквенная и шестибуквенная. Можно ли считать, что фактор длины анаграммы влияет на длительность попыток ее решения?

Код испытуемого	Условие 1. Четырехбуквенная анаграмма	Условие 2. Пятибуквенная анаграмма	Условие 3. Шестибуквенная анаграмма
1	5	235	7
2	7	604	20
3	2	93	5
4	2	171	8
5	35	141	7

3.10.Обработать данные вегетационного опыта с водными культурами по изучению действия соотношения N : P_2O_5: K_2O при питании рассады томатов на урожай плодов (таблица). В каждом варианте по 4 повторностей. Первый вариант контроль с соотношением N : P_2O_5: K_2O = 1:1:1. Второй вариант имеет соотношение N : P_2O_5: K_2O = 1:2:1, третий вариант 1:2:2, четвертый 2:1:1, пятый 2:2:1.

Ранний урожай плодов (г на сосуд)

Варианты	Урожай			
1	454	470	430	500
2	502	550	490	507
3	601	670	550	607
4	407	412	475	402
5	418	470	460	412

3.11. Изучали живой вес ягнят-одинцов при рождении (в кг), ношенных разное число дней:

Длительность беременности	Живой вес ягнят									
145	3,8	2,9	3,3	3,6	3,8	3,7	4,8	5,1	3,4	3,3
146	3,7	2,9	3,3	3,6	3,9	3,7	4,7	5,0	3,4	3,2
147	3,9	4,1	4,4	5,0	3,0	2,9	4,0	3,2	4,2	4,3
148	4,0	5,2	4,3	2,9	4,1	3,9	3,2	3,9	4,1	1,0
149	4,0	5,3	4,2	3,0	4,0	3,9	4,2	3,3	4,0	4,1
150	4,1	4,3	5,4	3,1	4,0	4,0	4,3	3,9	4,0	4,1
151	4,3	4,2	5,5	4,2	4,1	4,1	4,4	3,5	4,1	3,6
152	4,3	3,6	4,4	5,5	4,0	4,1	4,5	4,1	4,2	4,3
153	4,4	4,7	3,9	4,6	5,7	4,3	4,8	4,9	4,7	4,7

Примените метод дисперсионного анализа для выяснения влияния длительности плодоношения на живой вес ягнят.

3.12. Оценить влияние технологии чистовой обработки (три вида технологий) на точность изготовления детали. Проводятся по 4 замера (при каждом виде технологии) отклонения размера детали от номинала в мкм.Принять $\alpha = 0,05$.

Номер замера	Вид технологии		
	1	2	3
1.	1	2	3
2.	2	1	2
3.	2	3	2
4.	1	2	3

3.13. Требуется оценить влияние квалификации наладчиков (фактор *A*) на рассеяние диаметров шариков. Замеры отклонения диаметра от номинала для каждого из пяти наладчиков проводились по 6 раз:

№	A1	A2	A3	A4	A5
1	1,2	0,6	0,9	1,7	1
2	1,1	1,1	0,6	1,4	1,4
3	1	0,8	0,8	1,3	1,1
4	1,3	0,7	1	1,6	0,9
5	1,1	0,7	1	1,2	1,2
6	0,8	0,9	1,1	1,3	1,5

Проверяется нулевая гипотеза о равенстве математических ожиданий отклонения для всех пяти наладчиков, то есть предполагается, что квалификация наладчика не влияет на точность изготовления шариков.

4. Однофакторный дисперсионный анализ (при неодинаковом числе испытаний на всех уровнях)

Если число испытаний на уровне $F_1=q_1$, $F_2=q_2$, ..., $F_p=q_p$, то общую сумму квадратов отклонений вычисляют как и в случае с одинаковым числом испытаний на всех уровнях.

Факторную сумму квадратов отклонений находят по формуле:

$$S_{\text{факт}} = \frac{R_1^2}{q_1} + \frac{R_2^2}{q_2} + \cdots + \frac{R_p^2}{q_p} - \frac{(\sum_{j=1}^{p} R_j)^2}{n} \qquad (6)$$

где $n=q_1+q_2+\ldots+q_p$ – общее число испытаний.

Остальные вычисления производят, как и в предыдущем случае с одинаковым числом испытаний.

$$S_{\text{общ}} = \sum_{j=1}^{p} P_j - \frac{(\sum_{j=1}^{p} R_j)^2}{n} \qquad (7)$$

Разделив уже вычисленные факторную и остаточную суммы на соответствующее число степеней свободы, находят факторную и остаточную дисперсии:

$$S^2_{\text{факт}} = \frac{S_{\text{факт}}}{p-1}, \quad S^2_{\text{ост}} = \frac{S_{\text{ост}}}{n-p}, \quad F_{\text{набл}} = \frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}, \; F_{\text{кр}} = (\alpha; p-1; n-1)$$

Наконец, сравнивая, фактурную и остаточную дисперсии по критерию Фишера-Снедекора (приложение 1).

Если $F_{набл}<F_{кр}$ – различие групповых средних незначимое.

Если $F_{набл}>F_{кр}$ – различие групповых средних значимое.

Рассмотрим задачу: Произведено 13 испытаний, из них 4 – на первом уровне фактора, 4 – на втором, 3 – на третьем и 2 – на четвертом. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальной совокупностей с одинаковыми дисперсиями. Результаты испытаний приведены в таблице.

Номер испытания	Уровни фактора			
	F_1	F_2	F_3	F_4
1.	1,38	1,41	1,32	1,31
2.	1,38	1,42	1,33	1,33
3.	1,42	1,44	1,34	-
4.	1,42	1,45	-	-

Решение: сформулируем гипотезы H_0 – групповые средние различаются незначительно, тогда H_1 – групповые средние различаются значимо.

Вычислим вспомогательные величины:

$P_j = \sum x_{ij}^2$ -сумма квадратов наблюдаемых значений на уровне F_j(j=1,2,3,4)

$R_j = \sum x_{ij}$ - сумма наблюдаемых значений на уровне F_j(j=1,2,3,4)

Результаты занесем в таблицу:

Номер испытания	F_1	F_2	F_3	F_4	
1.	1,38	1,41	1,32	1,31	
2.	1,38	1,42	1,33	1,33	
3.	1,42	1,44	1,34	-	
4.	1,42	1,45	-	-	
					сумма
R	5,6	5,72	3,99	2,64	17,95
P	7,8416	8,1806	5,3069	3,485	24,8141
R^2	31,36	32,7184	15,9201	6,9696	86,9681

Вспомогательная таблица для расчета P_j :

Номер испытания	F_1^2	F_2^2	F_3^2	F_4^2
1.	1,9044	1,9881	1,7424	1,7161
2.	1,9044	2,0164	1,7689	1,7689
3.	2,0164	2,0736	1,7956	0
4.	2,0164	2,1025	0	0

Тогда $S_{общ}$=24,8141- $\frac{17,95^2}{13}$= 0,029292

$S_{факт}$= $\frac{31,36}{4}+\frac{32,7184}{4}+\frac{15,9201}{3}+\frac{6,9696}{2}-\frac{17,95^2}{13}=0,026292$

$S_{ост}=S_{общ}$ - $S_{факт}$ =0,029292-0,026292=0,0030

Найдем факторную дисперсию $S^2_{факт}=S_{факт}/4-1=0,00876$

Найдем остаточную дисперсию $S^2_{ост}=S_{ост}/13-4=0,00033$

Сравним факторную и остаточную дисперсию по критерию Фишера: найдем наблюдаемое значение критерия $F_{набл}=S^2_{факт}/S^2_{ост}=26,29$

Найдем критическую точку при уровне значимости α=0,05 и числам степеней свободы k_1=4-1=3; k_2=13-4=9 по таблице определили, что $F_{кр}$=3,86 делаем вывод, т.к. $F_{набл} > F_{кр}$ - нулевую гипотезу о равенстве групповых средних отвергаем, другими словами, групповые средние различаются значимо.

В таблицах Excel для проведения однофакторного равномерного или неравномерного дисперсионного анализа организуются вычисления по приведенным формулам с использованием функций СРЗНАЧ, СУММ, СУММКВ. Критическое значение критерия Фишера вычисляется при помощи функции FРАСПОБР. Также может быть использована процедура «Однофакторный дисперсионный анализ» из пакета анализа

Рассмотрим пример №1 (лабораторный практикум). Проверить, существенны ли различия содержания загрязняющего вещества на трех уровнях (глубинах взятия проб)

№	Уровни замеров		
	1	**2**	**3**
1	1,17	2,28	1,80
2	1,52	2,46	2,38
3	1,90	0,88	2,62
4	1,76	2,03	2,91
5	1,54	1,22	1,60
6	0,63	2,29	2,83
7	2,30	1,80	2,13
8	1,32	1,79	2,06
9	0,94	1,61	2,23
10	1,15	2,30	3,06
11	0,75	2,60	1,86
12	2,49	1,76	1,92
13	2,14	2,14	2,16
14	1,62	2,73	2,27
15	1,40		

Решение. Откроем таблицы *Excel* и внесем данные из таблицы.

Поскольку число измерений на разных уровнях неодинаково, требуется выполнить неравномерный дисперсионный анализ.

Вычислим в строке 17 объемы выборок: введем в ячейке B17

Формулу =СЧЁТ(B2:B16) и «растянем» результат в ячейки C17:D17. При этом вычисляется число непустых ячеек в каждом столбце. Общее число измерений n вычислим, просуммировав результаты в ячейке E17 (функция СУММ).

В строке 18 вычислим величины P_i при помощи функции СУММКВ и в ячейке E18 их сумму. В строке 19 вычислим величины R_i при помощи функции СУММ и в ячейке E19 – их сумму.

В строке 20 вычислим величины $\frac{R_i^2}{q_i}$ и в ячейке E20 – их сумму.

В ячейке F2 вычислим значение Собщ, введя формулу
=E18-E19^2/E17.

В ячейке G2 – значении $C_{факт}$:
=E20-E19^2/E17.

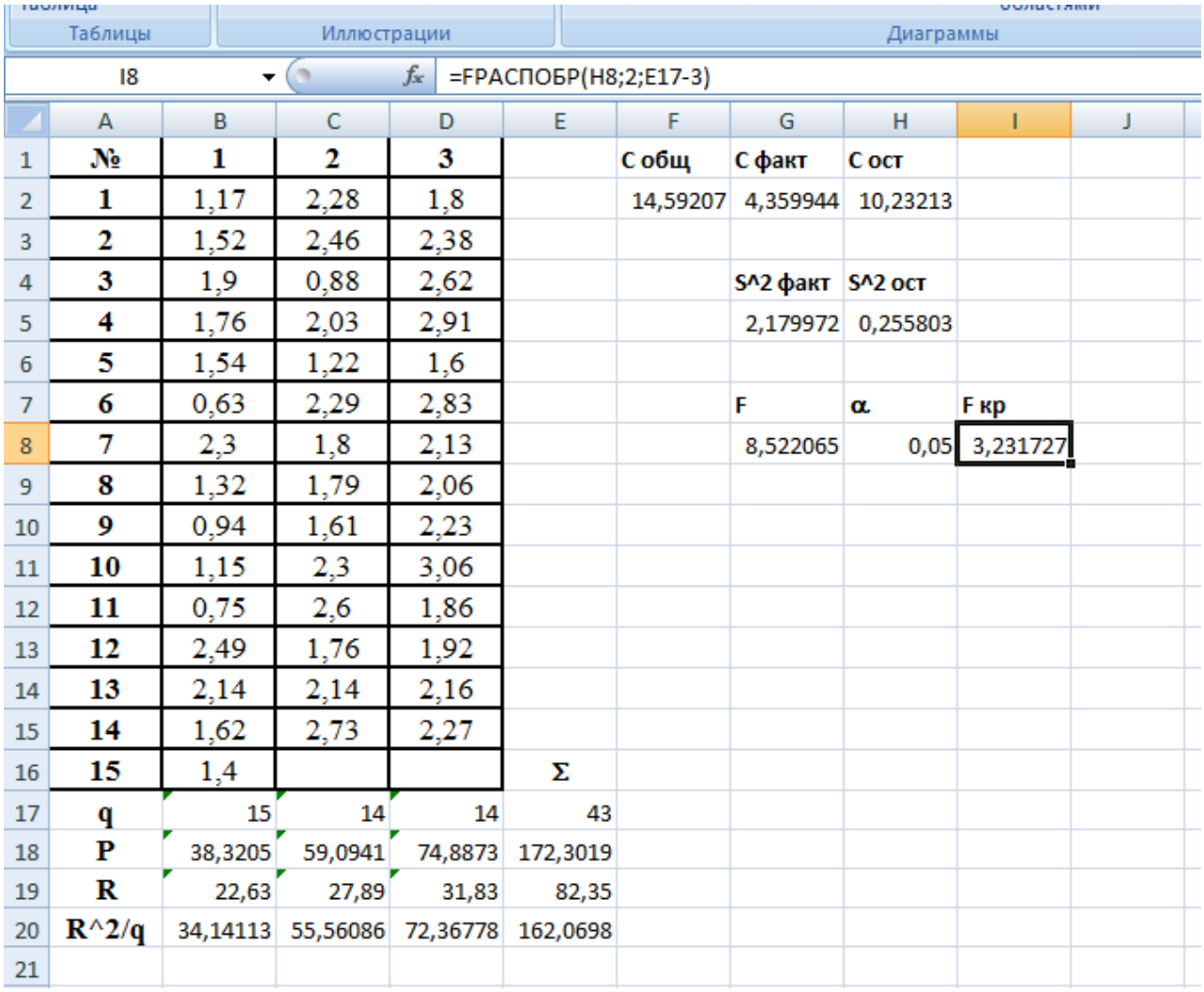
И в ячейке H2 – значение $C_{ост}$::
=F2-G2.

Далее, в ячейке G5 вычисляем значение $S^2_{факт}$, учитывая, что
p=3, и в ячейке H5 – значение $S^2_{ост}$, введя формулу =H2/(E17-3).

Значение F-статистики вычислим в ячейке G8. Для вычисления критического значения выберем уровень значимости и внесем его в ячейку H8. Критическое значение F-критерия в ячейке I8 вычисляем (с учетом того, что p=3), введя формулу

=FРАСПОБР(H8;2;E17-3).

Итог вычислений выглядит следующим образом:



	A	B	C	D	E	F	G	H	I	J
1	№	1	2	3		С общ	С факт	С ост		
2	1	1,17	2,28	1,8		14,59207	4,359944	10,23213		
3	2	1,52	2,46	2,38						
4	3	1,9	0,88	2,62			S^2 факт	S^2 ост		
5	4	1,76	2,03	2,91			2,179972	0,255803		
6	5	1,54	1,22	1,6						
7	6	0,63	2,29	2,83			F	α	F кр	
8	7	2,3	1,8	2,13			8,522065	0,05	3,231727	
9	8	1,32	1,79	2,06						
10	9	0,94	1,61	2,23						
11	10	1,15	2,3	3,06						
12	11	0,75	2,6	1,86						
13	12	2,49	1,76	1,92						
14	13	2,14	2,14	2,16						
15	14	1,62	2,73	2,27						
16	15	1,4			Σ					
17	q	15	14	14	43					
18	P	38,3205	59,0941	74,8873	172,3019					
19	R	22,63	27,89	31,83	82,35					
20	R^2/q	34,14113	55,56086	72,36778	162,0698					
21										

Поскольку $F>F_{кр}$, делаем вывод, что различия на разных уровнях существенные.

Следует отметить, однако, что уверенности в обоснованности применения параметрического дисперсионного анализа у нас нет, поскольку нет оснований считать данные в выборках нормально распределенными, а объем выборок не позволяет проверить гипотезу о соответствии данных нормальному закону при помощи критерия χ^2.

Однофакторный непараметрический дисперсионный анализ

Однофакторный непараметрический дисперсионный анализ производится при помощи критерия Краскала-Уоллиса (в русскоязычной литературе

его также называют критерием Краскела-Уоллеса, Крускала-Уоллеса). Этот критерий является многовыборочным обобщением критерия Уилкоксона (или Манна-Уитни).

Для применения критерия Краскала-Уоллиса следует проранжировать совмещенную выборку (из всех измерений при различных уровнях фактора); обозначим ранг *i*-го элемента выборки на j-м уровне фактора d_i^j. Далее находят суммы рангов $R_1,\ldots,R_p$ для каждого уровня фактора:

$R_i = \sum_{j=1}^{q_i} d_i^j$

При отсутствии связанных рангов статистика критерия Краскала-Уоллиса имеет вид:

$$H = \frac{12}{n\cdot(n+1)}\left(\frac{R_1^2}{q_1} + \cdots + \frac{R_p^2}{q_p}\right) - 3(n+1)$$

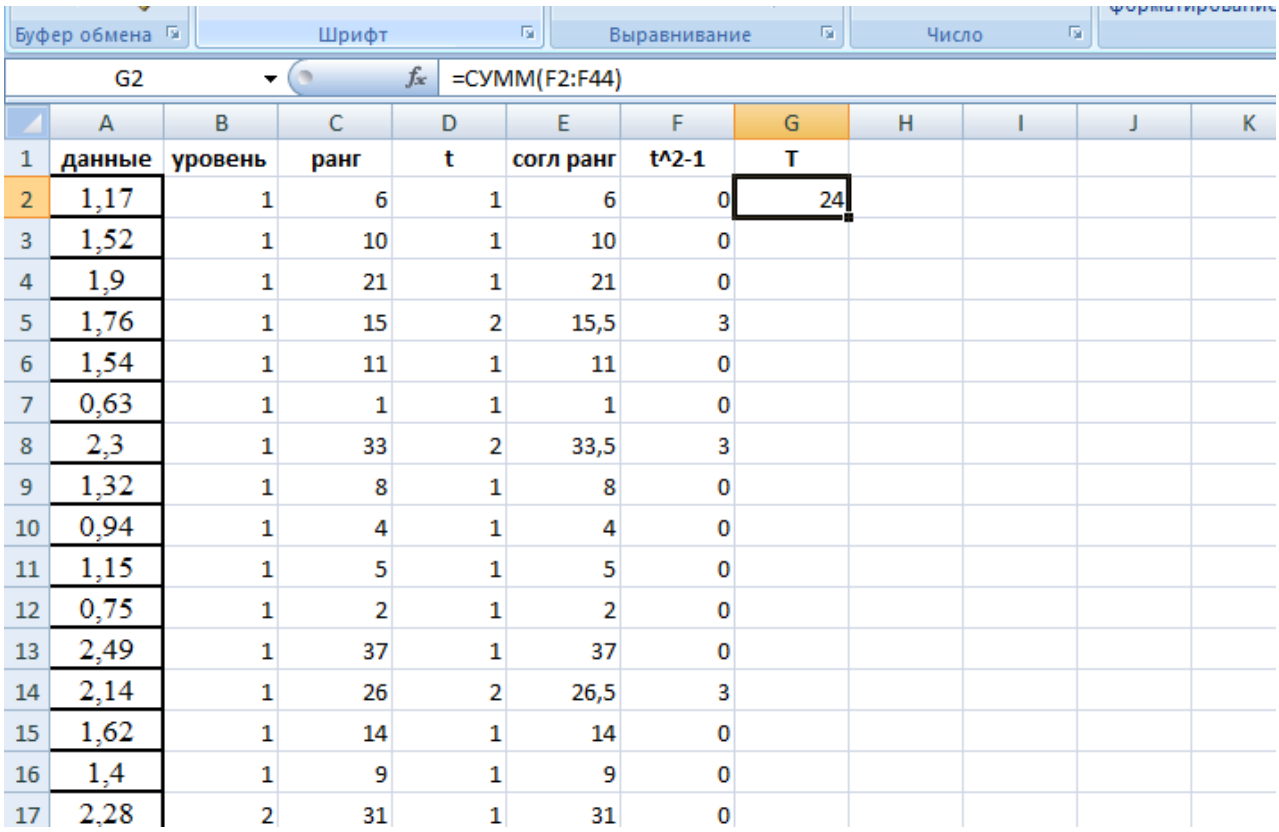
При наличии связанных рангов используют модифицированную статистику:

$H^* = \frac{H}{1-\frac{1}{n^3-n}\sum_{i=1}^{k}(t_i^3-t_i)}$, где t_i – число элементов в *i*-й связке, k –число связок.

Нулевая гипотеза (об отсутствии влияния фактора на признак) отклоняется, если рассчитанное значение критерия превышает критическое Hα для заданного уровня значимости. Для малых выборок ($p\leq5$, $q_i\leq8$) критические значения критерия Краскала-Уоллиса определяются χ^2 по таблицам. При достаточно большом объеме выборки критическое значение определяется исходя из распределения с p-1 степенями свободы. При организации вычислений в Excel для этого применяют функцию ХИ2ОБР(α;p-1).

Решить пример 1 при помощи критерия Краскала-Уоллиса.

Решение. На новом листе Excel внесем в столбец А совмещенную выборку данных из таблицы примера 1, в столбце В для каждого данного укажем, к какому уровню факторного признака оно относится. Проведем предварительные расчеты как для критерия Вилкоксона: в столбце С вычислим ранги данных в совмещенной выборке при помощи функции РАНГ (введем формулу в ячейку С2, зафиксируем используемый массив, растянем); в столбце D вычислим длины связок t при помощи функции СЧЁТЕСЛИ. Поскольку имеются связки неединичной длины, далее в столбце Е вычислим согласованные ранги (согласованный ранг=ранг+(t-1)/2); в столбце F – величины t^2-1, которые затем сложим в ячейке G2 (обозначив величину, например, Т). Результат вычислений имеет вид:



Буфер обмена | Шрифт | Выравнивание | Число

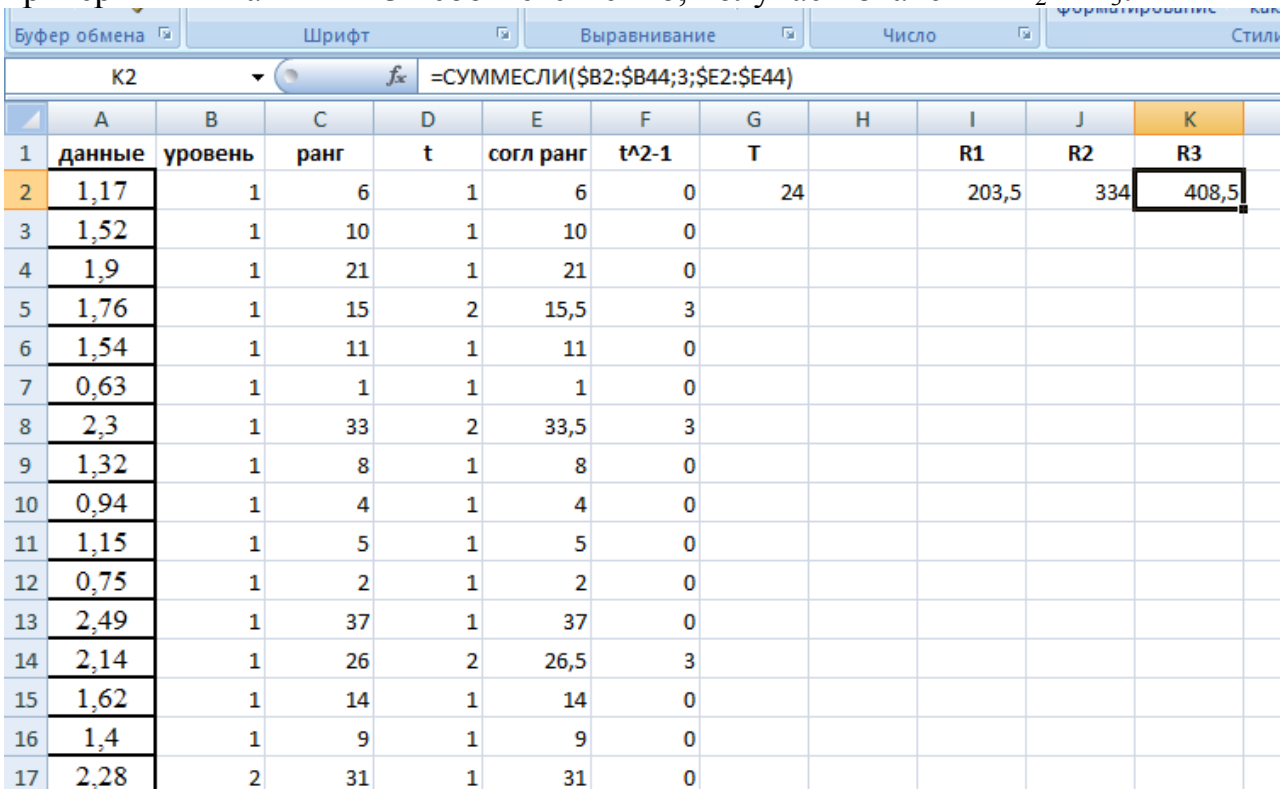
G2 | =СУММ(F2:F44)

	A	B	C	D	E	F	G	H	I	J	K
1	данные	уровень	ранг	t	согл ранг	t^2-1	T				
2	1,17	1	6	1	6	0	24				
3	1,52	1	10	1	10	0					
4	1,9	1	21	1	21	0					
5	1,76	1	15	2	15,5	3					
6	1,54	1	11	1	11	0					
7	0,63	1	1	1	1	0					
8	2,3	1	33	2	33,5	3					
9	1,32	1	8	1	8	0					
10	0,94	1	4	1	4	0					
11	1,15	1	5	1	5	0					
12	0,75	1	2	1	2	0					
13	2,49	1	37	1	37	0					
14	2,14	1	26	2	26,5	3					
15	1,62	1	14	1	14	0					
16	1,4	1	9	1	9	0					
17	2,28	2	31	1	31	0					

Вычислим в ячейке I2 величину R_1 – сумму согласованных рангов данных, относящихся к 1 уровню факторного признака:

=СУММ ЕСЛИ ($B2:$B44;1;$E2:$E44)

В этой формуле массивы зафиксированы для того, чтобы можно было скопировать ее в ячейки J2 и K2. Сделав это и заменив в формулах значение критерия «1» на «2» и «3» соответственно, получаем значения R_2 и R_3:



Буфер обмена | Шрифт | Выравнивание | Число | Стили

K2 | =СУММЕСЛИ($B2:$B44;3;$E2:$E44)

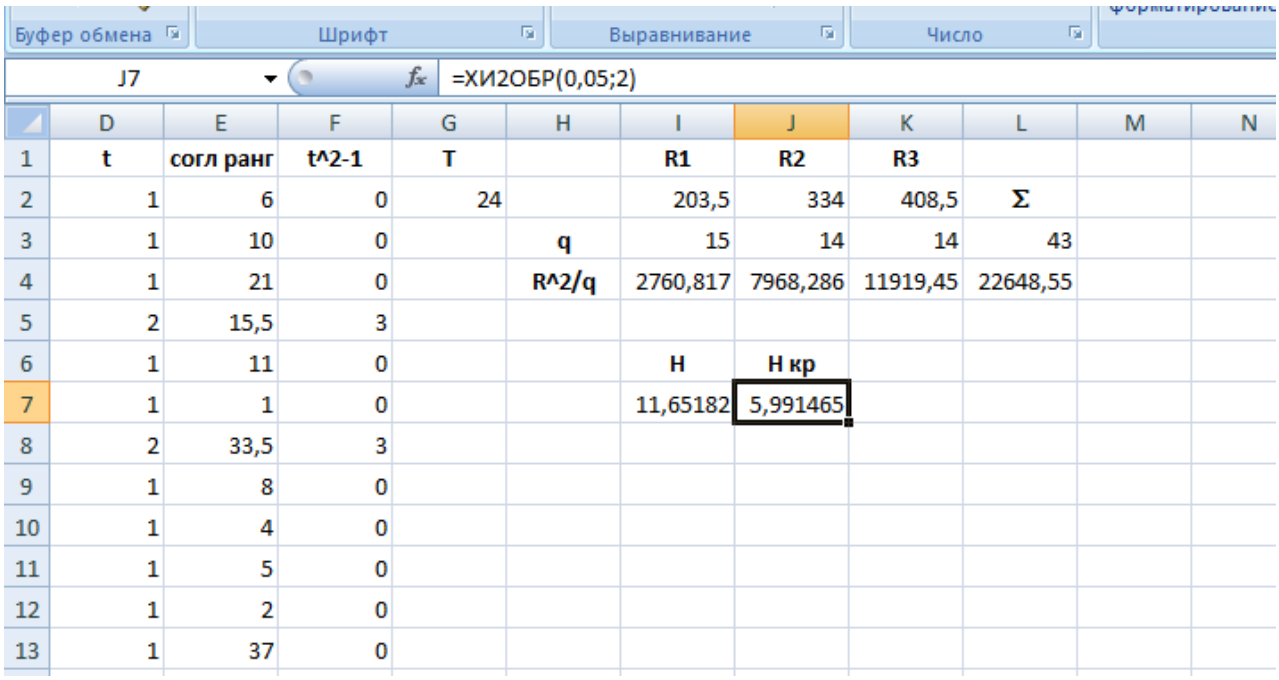
	A	B	C	D	E	F	G	H	I	J	K
1	данные	уровень	ранг	t	согл ранг	t^2-1	T		R1	R2	R3
2	1,17	1	6	1	6	0	24		203,5	334	408,5
3	1,52	1	10	1	10	0					
4	1,9	1	21	1	21	0					
5	1,76	1	15	2	15,5	3					
6	1,54	1	11	1	11	0					
7	0,63	1	1	1	1	0					
8	2,3	1	33	2	33,5	3					
9	1,32	1	8	1	8	0					
10	0,94	1	4	1	4	0					
11	1,15	1	5	1	5	0					
12	0,75	1	2	1	2	0					
13	2,49	1	37	1	37	0					
14	2,14	1	26	2	26,5	3					
15	1,62	1	14	1	14	0					
16	1,4	1	9	1	9	0					
17	2,28	2	31	1	31	0					

Значение исправленной статистики Краскала-Уоллиса можно вычислить теперь по приведенной выше формуле, но она достаточно громоздка, поэтому для облегчения вычислений проведем вспомогательные построения: укажем объемы выборок q_1=15, q_2=14, q_3=14; вычислим их сумму, получив объем совмещенной выборки n; вычислим величины $\frac{R_i^2}{q_i}$ (это удобно сделать, введя формулу один раз и скопировав в соседние ячейки) и их сумму.

Теперь вычислим значение исправленной статистики Краскала- Уоллиса, введя в ячейке I7 формулу:

=(12*L4/(L3*(L3+1))-3*(L3+1))/(1-G2/(L3^3-L3)).

Поскольку объемы выборок превышают 8, в таблицах критическое значение критерия найти нельзя; вычислим его при помощи функции ХИ2ОБР:



	D	E	F	G	H	I	J	K	L	M	N
1	t	согл ранг	t^2-1	T		R1	R2	R3			
2	1	6	0	24		203,5	334	408,5	Σ		
3	1	10	0		q	15	14	14	43		
4	1	21	0		R^2/q	2760,817	7968,286	11919,45	22648,55		
5	2	15,5	3								
6	1	11	0			H	H кр				
7	1	1	0			11,65182	5,991465				
8	2	33,5	3								
9	1	8	0								
10	1	4	0								
11	1	5	0								
12	1	2	0								
13	1	37	0								

Поскольку рассчитанное значение критерия превышает критическое для уровня значимости 0,05, значит с доверительной вероятностью 0,95 можно сделать вывод о влиянии фактора на исследуемый признак, то есть различия в загрязнении на разных уровнях существенны.

Задание для практической (лабораторной) работы:

При выполнении работы необходимо решить следующие задачи.

1) Проверить, существенны ли различия численности персонала, занятого научными исследованиями и разработками, по категориям по субъектам Российской Федерации в 2017 и 2016 году (первая и вторая колонка данных) при помощи дисперсионного анализа и критерия Краскала-Уоллиса

2) Выбрать подходящий метод и проверить, существенны ли различия численности персонала, занятого научными исследованиями и разработками в России в 2017, 2016, 2015 годах

Центральный федеральный округ	**2017г**	**2016г**	**2015г**
Белгородская область	1655	1717	1749
Брянская область	688	630	805
Владимирская область	5365	5421	5697
Воронежская область	10654	10334	10600
Ивановская область	574	618	634
Калужская область	9275	9963	10170
Костромская область	114	121	129
Курская область	2719	2846	2891
Липецкая область	530	616	700
Московская область[1)]	86579	87706	85864
Орловская область	837	878	915
Рязанская область	2461	2718	3100
Смоленская область	903	761	714
Тамбовская область	1125	1165	1594
Тверская область	3971	4430	4596
Тульская область	4142	4237	4154
Ярославская область	6354	6404	6319
г. Москва[1)]	224517	231728	239509

Северо-Западный федеральный округ	**2017г**	**2016г**	**2015г**
Республика Карелия	1168	1207	1202
Республика Коми	1655	1909	1981
Архангельская область	1021	1094	1107
в том числе Ненецкий автономный округ	22	59	62
Архангельская область без АО	999	1035	1045
Вологодская область	464	509	541
Калининградская область	1788	2057	2128
Ленинградская область	7265	7247	7229
Мурманская область	2138	2265	2342
Новгородская область	1739	1602	1638
Псковская область	236	278	818
г. Санкт-Петербург	77051	76950	79076

Южный федеральный округ	**2017г**	**2016г**	**2015г**
Республика Адыгея	283	282	279
Республика Калмыкия	158	184	175
Республика Крым[3)]	2113	2096	1676
Краснодарский край	6916	7532	9265
Астраханская область	653	692	933
Волгоградская область	3869	4026	3958
Ростовская область	11846	12102	12556
г. Севастополь [3)]	1084	1097	1288
Северо-Кавказский федеральный округ	**2017г**	**2016г**	**2015г**
Республика Дагестан	1693	1997	1689

Республика Ингушетия	244	326	346
Кабардино-Балкарская Республика	1050	1122	894
Карачаево-Черкесская Республика	589	581	586
Республика Северная Осетия - Алания	547	612	654
Чеченская Республика	480	480	561
Ставропольский край	2634	2537	2791

Приволжский федеральный округ	**2017г**	**2016г**	**2015г**
Республика Башкортостан	7743	8008	8262
Республика Марий Эл	187	281	203
Республика Мордовия	831	927	990
Республика Татарстан	12323	12189	12708
Удмуртская Республика	1959	1800	1603
Чувашская Республика	1555	1487	1296
Пермский край	10328	10304	11005
Кировская область	1776	1672	1729
Нижегородская область	40404	41427	39961
Оренбургская область	1387	1404	950
Пензенская область	4817	4690	5790
Самарская область	10844	9615	12700
Саратовская область	5684	5364	5245
Ульяновская область	5047	5136	5237

Уральский федеральный округ	**2017г**	**2016г**	**2015г**
Курганская область	629	683	671
Свердловская область	21212	22180	21900
Тюменская область	8260	8789	8811
в том числе:			
Ханты-Мансийский автономный округ - Югра	1568	1831	1978
Ямало-Ненецкий автономный округ	116	112	109
Тюменская область без АО	6576	6846	6724
Челябинская область	15167	14785	15114
Сибирский федеральный округ	**2017г**	**2016г**	**2015г**
Республика Алтай	125	132	138
Республика Бурятия	1144	1191	1266
Республика Тыва	385	388	384
Республика Хакасия	247	237	220
Алтайский край	2486	2719	3154
Забайкальский край	504	478	495
Красноярский край	7234	7632	7543
Иркутская область	4292	4409	4671
Кемеровская область	1361	1551	1491
Новосибирская область	22256	21843	21621
Омская область	4651	4779	4714
Томская область	9301	9922	9448

Дальневосточный федеральный округ	2017г	2016г	2015г
Республика Саха (Якутия)	2147	2279	2250
Камчатский край	921	1093	1133
Приморский край	5700	5655	5809
Хабаровский край	1717	1813	2043
Амурская область	536	667	692
Магаданская область	611	664	636
Сахалинская область	807	827	888

Данные взяты с официального сайта Росстата http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/science_and_innovations/science/#

Задачи для решения на практическом занятии:

4.1. В соответствии с данными эксперимента, приведенными в табл. 1, требуется при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних.

Таблица 1.

Данные эксперимента			
Номер испытания	Уровни фактора *Fj*		
i	F_1	F_2	F_3
1	37	60	69
2	47	86	100
3	40	67	98
4	60	92	–
5	–	95	–
6	–	98	–
Среднее уровней	46	83	89

4.2.Произведено 14 испытаний, из них 5 – на первом уровне фактора, 3 – на втором, 2 – на третьем , 3 – на четвертом и 1 – на пятом. Методом дисперсионного анализа при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Полагается, что выборки извлечены из нормальных совокупностей с одинаковыми дисперсиями. Результаты испытаний приведены в табл. 2.

Таблица 2

Номер испытания	Уровни фактора				
	F_1	F_2	F_3	F_4	F_5
1	7,3	5,4	6,4	7,9	7,1
2	7,6	7,1	8,1	9,5	-
3	8,3	7,4	-	9,6	-
4	8,3	-	-	-	-
5	8,4	-	-	-	-