Практическое занятие № 9

для зарубежных студентов 1-го курса лечебного факультета

Тема: Компьютерный анализ медицинских данных. Дисперсионный анализ.

Установление значимости различий средних арифметических, измерение степени влияния факторов и их градаций на варьирующий (результативный) признак наиболее эффективно достигаются путем применения дисперсионного анализа. Впервые основа дисперсионного анализа была разработана известным английским статистиком Р. Фишером в 1925 году.

ДИСПЕРСИОННЫЙ АНАЛИЗ – это метод в статистической математике, направленный на поиск зависимостей в экспериментальных данных путём исследования значимости различий вариабельности признака в исследуемой совокупности. В литературе также встречается обозначение ANOVA (от англ. ANalysis Of VAriance). Он базируется на определении степени рассеяния (дисперсии) оцениваемых признаков в нескольких группах. Это позволяет измерить силу влияния отдельных факторов на значения показателей.

Известно, что величина отдельных признаков представляет собой результат воздействия разнообразных факторов, различных по силе влияния. Одни факторы имеют значительно большую силу влияния, другие - меньшую. Причем, как правило, факторы сами воздействуют друг на друга, сочетая свое влияние, иногда усиливают действие друг друга, иногда, наоборот, погашают это действие. Преимуществом дисперсионного анализа является то, что он дает возможность изучить и сравнить роль каждого из них.

В отличие от дисперсионного анализа другие общепринятые в медицинских исследованиях статистические методы обработки, как правило, предусматривают проведение по парных сравнений, что приводит к огромному объему расчетов и часто не дает полной оценки.

Сущность дисперсионного анализа заключается в изучении статистического влияния одного или нескольких факторов на результативный признак.

ФАКТОР - это влияние, воздействие или состояние, которое отражается на размерах и разнообразии результативного признака.

РЕЗУЛЬТАТИВНЫЙ ПРИЗНАК - это элементарное свойство объектов, изучаемое как результат влияния факторов: организованных в исследовании (основных - **x**) и всех остальных, неорганизованных, не учтенных в данном исследовании (случайных - **z**).

ГРАДАЦИИ ФАКТОРА - это степень его воздействия, в том числе отсутствие воздействия (нулевое значение) в контрольной группе или состояние объектов изучения (пол, возраст и др.).

ДИСПЕРСИОННЫЙ КОМПЛЕКС - это совокупность градаций изучаемых данных (групп объектов наблюдения, выбранных исследователем) с вычисленными значениями относительных или средних величин по каждой градации.

При изучении количественных признаков в градации дисперсионного комплекса заносятся числовые результаты измерения изучаемого признака у каждого отдельного объекта. При изучении качественных признаков в градации комплекса заносится число объектов с наличием признака и общее число объектов.

Дисперсионные комплексы, составленные по принципу случайного отбора, называются РАНДОМИЗИРОВАННЫМИ.

Статистическое влияние - это отражение в разнообразии результативного признака того разнообразия фактора (его градаций), которое организовано в исследовании.

Сумма основных и случайных факторов составит общие факторы (**y**). Результативный признак изучается как результат воздействия факторов, организованных в исследовании (**x**)и неорганизованных(**z**).

Общее влияние как раз и определяет влияние всех организованных и неорганизованных (случайных) факторов, определивших такое развитие признака, которое наблюдалось в дисперсионном комплексе. Общее влияние служит базой для определения доли влияний - факториальных и случайных.

Факториальное влияние -это простое или комбинированное статистическое влияние изучаемых (учтенных) факторов.

Случайное влияние - это действие тех факторов, которые не учтены в дисперсионном комплексе и составляют общий фон, на котором действуют учитываемые факторы.

Таким образом, дисперсионный анализ исследует важнейшее свойство совокупности

– разнообразие (вариабельность, дисперсию) признака. Для этого выделяется три вида разнообразия: межгрупповое, внутригрупповое и общее .
Межгрупповое разнообразие зависит от влияния исследуемого фактора, по которому выделяется каждая группа. Иными словами, межгрупповое разнообразие- это различие средних в каждой группе.

Внутригрупповое разнообразие зависит от силы влияния каких-то неучтенных случайных факторов. Общее разнообразие складывается из межгруппового и внутригруппового.

В основе дисперсионного анализа лежит предположение о том, что одни переменные могут рассматриваться как причины (факторы, независимые переменные): $f_1, \ldots, f_k$, а другие как следствия (зависимые переменные). Независимые переменные называют иногда регулируемыми факторами именно потому, что в эксперименте исследователь имеет возможность варьировать ими и анализировать получающийся результат.

Основной целью дисперсионного анализа является исследование значимости различия между группами с помощью сравнения дисперсий. Разделение общей дисперсии на несколько источников позволяет сравнить дисперсию, вызванную различием между группами, с дисперсией, вызванной внутригрупповой изменчивостью. При истинности нулевой гипотезы (о равенстве средних в нескольких группах наблюдений, выбранных из генеральной совокупности), оценка дисперсии, связанной с внутригрупповой

изменчивостью, должна быть близкой к оценке межгрупповой дисперсии. Если выполняется сравнение средних в двух выборках, дисперсионный анализ даст тот же результат, что и обычный t-критерий Стьюдента. Однако, помимо этого, он позволяет оценить степень такого влияния, а также может использоваться при сопоставлении более чем 2-х групп.

Сущность дисперсионного анализа заключается в расчленении общей дисперсии (***D*** или ***SS***) изучаемого признака на отдельные компоненты, обусловленные влиянием конкретных факторов, и проверке гипотез о значимости влияния этих факторов на исследуемый признак. Обозначение ***SS*** - это сокращение от фразы "суммы квадратов" (Англ. sum of squares). Оно чаще всего используется в зарубежных источниках.

Сравнивая компоненты дисперсии друг с другом посредством F-критерия Фишера, можно определить, какая доля общей вариативности результативного признака обусловлена действием регулируемых факторов.

Критерий Фишера экспериментальных (эмпирических) данных (***F* Эмп.**) вычисляется как отношение среднего квадрата дисперсии, обусловленной изучаемым фактором, к среднему квадрату случайной дисперсии:

$$F_{Эмп.} = \frac{MS_{Факт.}}{MS_{случ.}}$$

,

где: ***F****Эмп.* – критерий Фишера, вычисленный в исследуемой совокупности,

MS*Факт.* –средний квадрат дисперсии, обусловленной изучаемым фактором,

MS*Случ.* – средний квадрат дисперсии, обусловленной случайными факторами.

Для оценки достоверности полученного результата вычисленный на экспериментальных данных критерий Фишера сравнивают с его критическим значением (***F****Крит.*) для принятого уровня вероятности (***p*)** и степеней свободы (***df*)**.

С целью уменьшения объема вычислений в программе Excel может применяться надстройка «Анализ данных» и ее модуль «Однофакторный дисперсионный анализ».

!!Выполнить задание.!!

Пример задачи на выявление влияния изучаемого фактора с помощью дисперсионного анализа.

Условие задачи: 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. Результаты измерений представлены в таблице.

Результаты запоминания слов, предъявляемых испытуемым

№ 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
Суммы	43	37	24
Средние	7,17	6,17	4
Общая сумма	104		

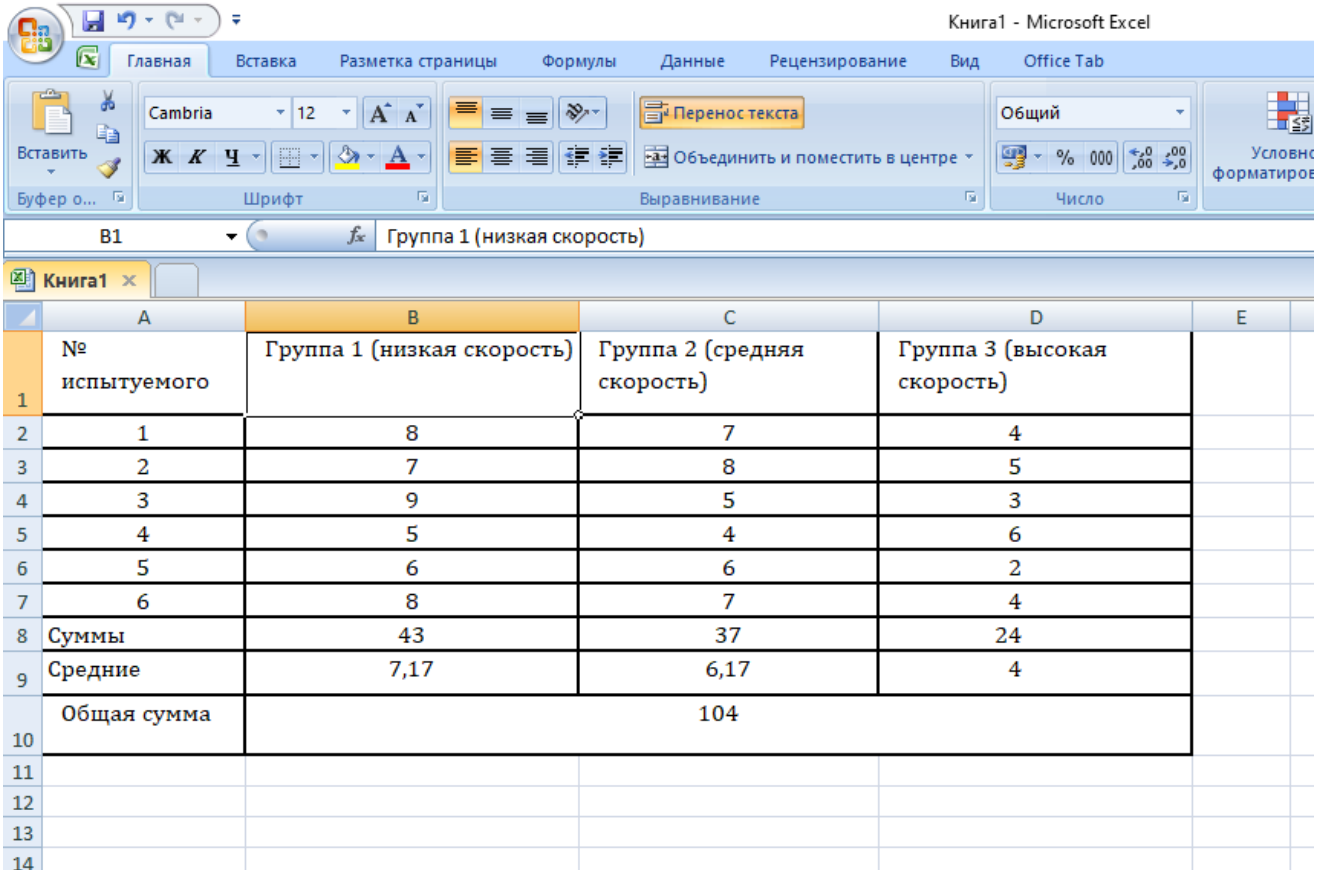
Предполагается, что показатели запоминания и воспроизведения слов будут зависеть от скорости их предъявления. Поэтому перед выполнением статистического наблюдения были выдвинуты гипотезы:

- Основная (***H₀***): различия в объеме воспроизведения слов *между* группами являются не более выраженными, чем случайные различия *внутри* каждой группы.
- Альтернативная (***H₁***): Различия в объеме воспроизведения слов *между* группами являются более выраженными, чем случайные различия *внутри* каждой группы.

Задание: подтвердить одну из указанных выше гипотез.

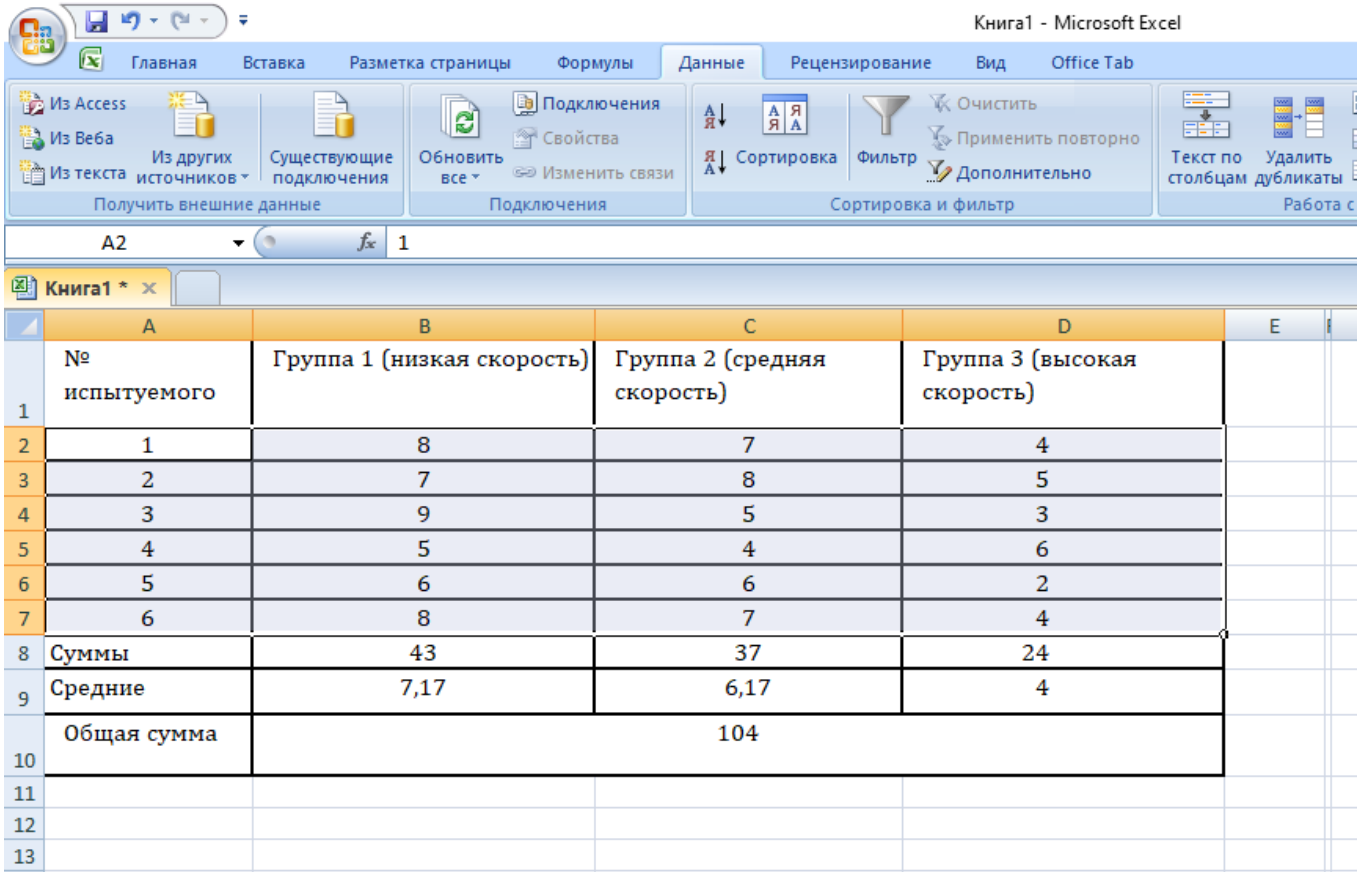
Решение:

1) **необходимо запустить программу Excel**
2) **создать новый лист**
3) **на этом листе ввести.**



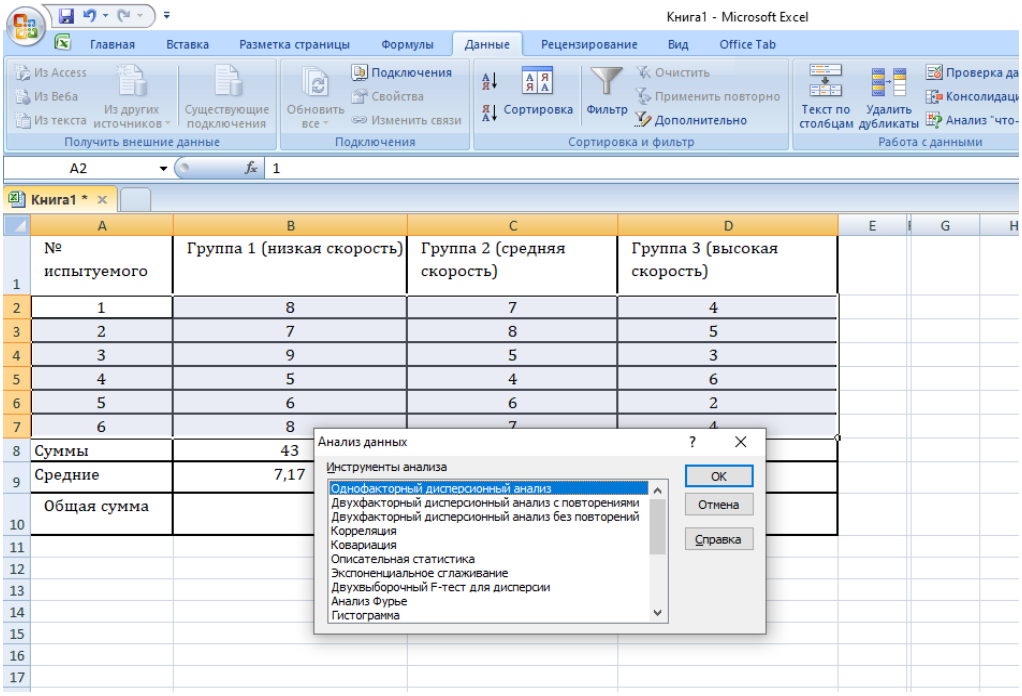
	A	B	C	D	E
1	№ испытуемого	Группа 1 (низкая скорость)	Группа 2 (средняя скорость)	Группа 3 (высокая скорость)	
2	1	8	7	4	
3	2	7	8	5	
4	3	9	5	3	
5	4	5	4	6	
6	5	6	6	2	
7	6	8	7	4	
8	Суммы	43	37	24	
9	Средние	7,17	6,17	4	
10	Общая сумма	104			
11					
12					
13					
14					

4. выделить ячейки с A2 по D7

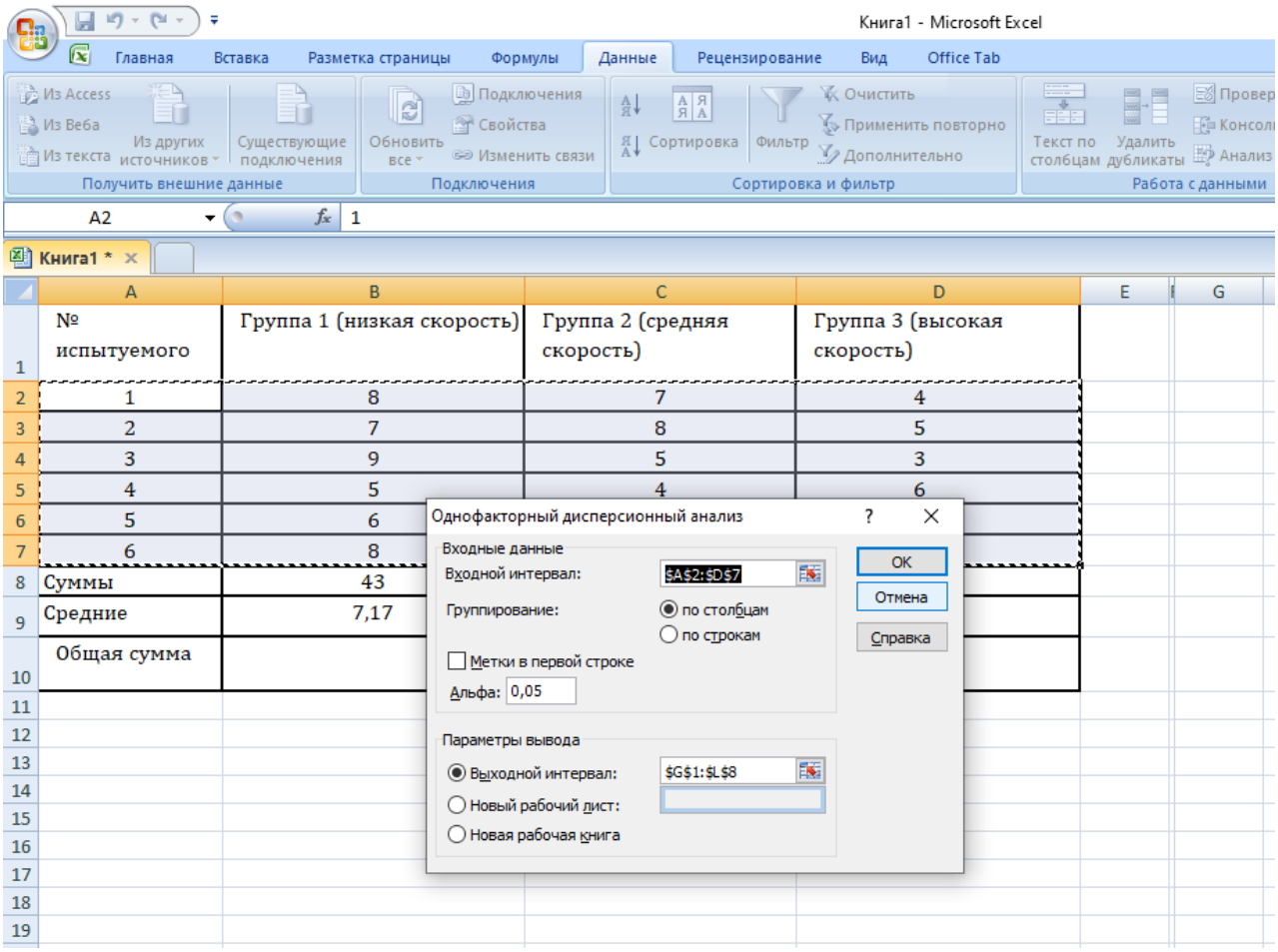


	A	B	C	D	E
1	№ испытуемого	Группа 1 (низкая скорость)	Группа 2 (средняя скорость)	Группа 3 (высокая скорость)	
2	1	8	7	4	
3	2	7	8	5	
4	3	9	5	3	
5	4	5	4	6	
6	5	6	6	2	
7	6	8	7	4	
8	Суммы	43	37	24	
9	Средние	7,17	6,17	4	
10	Общая сумма	104			
11					
12					
13					

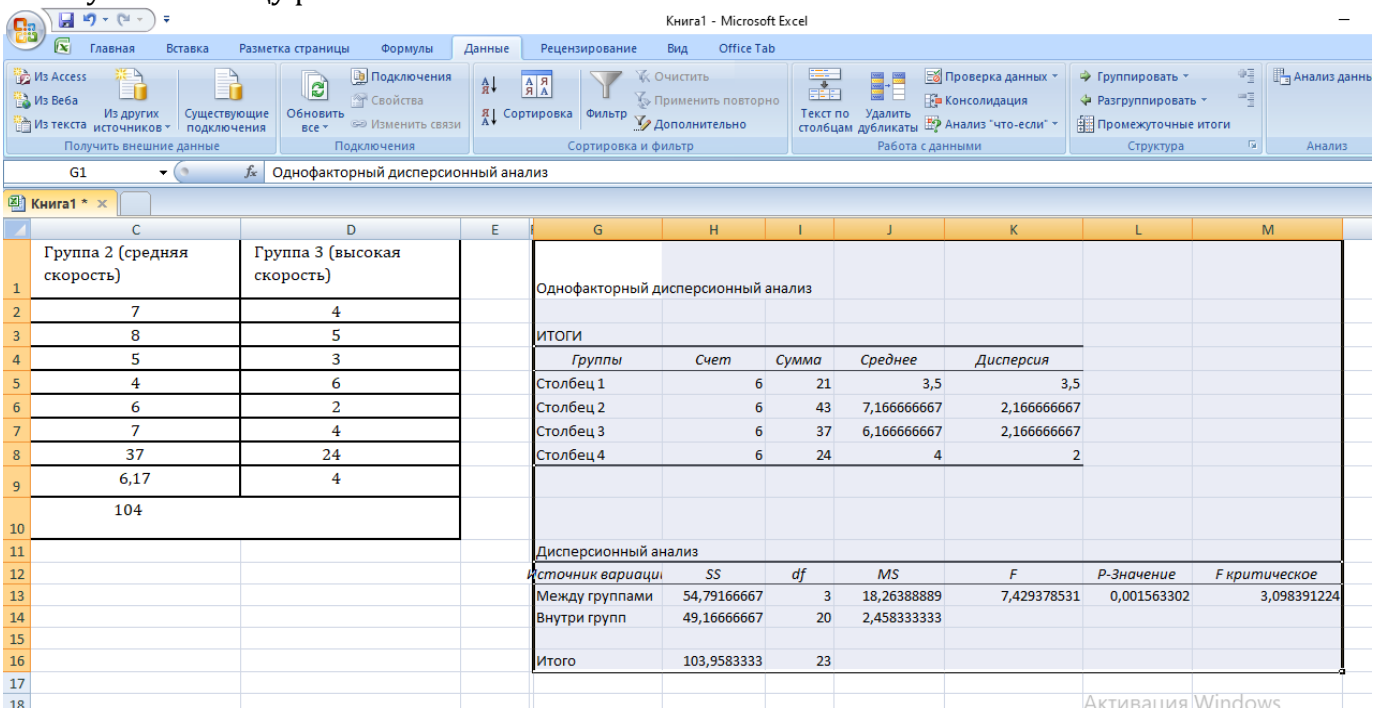
5. Меню Данные ***Анализ данных*** ***Однофакторный дисперсионный анализ*** ***ОК***



6. Выбрать выходной диапазон. Нажать Ок.



7. Получим таблицу расчетов



8. Сравним $F_{Эмп.} < F_{Крит.}$ Если $F_{Эмп.} < F_{Крит.}$, то нулевая гипотеза принимается, в противном случае принимается альтернативная гипотеза. Для нашего примера (7,43>3,09), следовательно, принимается альтернативная гипотеза - влияние существует.

Вывод: различия в объеме воспроизведения слов между группами являются более выраженными, чем случайные различия внутри каждой

группы (p<0,05). Таким образом, скорость предъявления слов влияет на объем их воспроизведения.

Самостоятельная работа.

Задание. С помощью модуля «Однофакторный дисперсионный анализ» программы Excel оцените влияние одного из факторов на изучаемый признак, сформулируйте вывод.

Задача 1.В эксперименте на животных измерено время пробежки мышей по лабиринту на фоне различной концентрации препарата, стимулирующего нервную систему. Результаты измерений в секундах указаны в таблице.

Результаты измерения времени пробежки мышей по лабиринту (сек.)

№ животного	Группа 1 (низкая концентрация)	Группа 2 (средняя концентрация)	Группа 3 (высокая концентрация)
1	8	7	4
2	7	8	5
3	9	5	3
4	5	4	6
5	6	6	2
6	8	4	3
7	7	7	4
8	8	6	2
9	9	7	4
10	8	7	3

Необходимо подтвердить влияние стимулирующего вещества с помощью дисперсионного анализа в программе Excel надстройкой «Анализ данных»модулем «Однофакторный дисперсионный анализ».

Задача 2. На предприятии проведено изучение уровня травматизма с учетом фактора стажа работы сотрудников 5-и участков с близкими условиями труда, получены следующие данные.

Уровень травматизма на 100 работающих

Участок	Стаж работы			
	до 5 лет	6-10 лет	11-15 лет	16 лет и более
1	11	8	6	4
2	12	9	7	7
3	10	6	6	5
4	10	9	7	7
5	13	8	5	3

Необходимо оценить влияние стажа работы на уровень травматизма с помощью дисперсионного анализа в программе Excel надстройкой «Анализ данных» модулем «Однофакторный дисперсионный анализ».

Рекомендуемая литература:

1. Гельман В.Я. Медицинская информатика. Практикум. СПб: Питер, 2001. – 420 с.
2. Гмурман В.Е.Теория вероятностей и математическая статистика : Учебное пособие – 12-е изд., перераб. и доп. – М. : Юрайт, 2011. – 478 с. :ил.
3. Информатика. Книга 2. Основы медицинской информатики : учебник / В.И. Чернов, И. Э. Есауленко, М В. Фролов и др. – М. : Дрофа, 2009. – 205, [3] с. : ил.
4. Применение методов статистического анализа для изучения общественного здоровья и здравоохранения [Электронный ресурс]: учебное пособие для практических занятий / под ред. В.З. Кучеренко. - 4-е изд., перераб. и доп. -М. : ГЭОТАР-Медиа, 2011. - 256 с. – Режим доступа: http://www.studmedlib.ru