Пример выполнения контрольной работы по математической статистике для восточников

1. По **первому** признаку для городов каждого штата (OH, TX,CA) вычислить средние значения, стандартные отклонения, ошибки среднего, медианы, минимальные и максимальные значения.

2. Проверить гипотезы однородности для штатов TX и CA по критериям Фишера (равенство дисперсий) и Стьюдента (равенство средних). Ответить на вопрос: является значимым различие средних показателей первого признака в указанных штатах.

3. Для трех штатов одновременно при помощи однофакторного дисперсионного анализа проверить гипотезу одинаковости средних. Проверить результат при помощи критерия Краскела-Уоллеса.

4. Вычислить коэффициент корреляции между первым и вторым признаками. При помощи статистики Стьюдента проверить значимость отклонения от нуля коэффициента корреляции.

город	штат	темпер.(X)	площадь (Y)
COLUMBUS	OH	73,2	190,9
CLIVELAN	OH	71,9	77
CINCINNA	OH	76,4	77,2
TOLEDO	OH	72,1	80,6
AKRON	OH	71,9	62,2
HOUSTON	TX	83,5	539,9
DALLAS	TX	85,9	342,4
SAN ANTO	TX	85	333
EL PASO	TX	82,3	245,4
AUSTIN	TX	84,5	217,8
FORT WOR	TX	85,3	281,1
ARLINGTO	TX	85,3	93
CORPUS C	TX	84,1	135
LOS ANGE	CA	74,3	469,3
SAN DIEG	CA	71	324
SAN JOSE	CA	69,5	171,3
SAN FRAN	CA	59,1	46,7
LONG BEA	CA	73,1	50
SACRAMEN	CA	75,7	96,3
FRESNO	CA	81,9	99,1
OAKLAND	CA	62,1	56,1
SANTA AN	CA	72,6	27,1
ANAHEIM	CA	72,6	44,3
RIVERSID	CA	77,9	77,7
STOCKTON	CA	77,7	52,6

1. **Описательная статистика.** Внесем данные в электронные таблицы. Для сравнения трех штатов по первому признаку (среднеиюльская температура по Фаренгейту) используем функции описательной статистики. Например, данные первого признака городов штата OH помещены в пяти ячейках C2:C6. При помощи функций вычисляем

СРЗНАЧ(A2:A6) среднее значение,

СТАНДОТКЛОН(A2:A6) стандартное отклонение,

МЕДИАНА(A2:A6) медиану,

МИН(A2:A6) и МАКС(A2:A6) минимальное и максимальное значение.

Аналогично поступаем с случае двух других выборок. Для вычисления ошибки среднего нужно стандартное отклонение поделить на корень из объема выборки. Нетрудно заметить, что объемами выборок, соответствующими штатам OH, TX и CA, являются числа 5, 8 и 12. При вычислении корня лучше использовать функцию "КОРЕНЬ", например, КОРЕНЬ(5) или КОРЕНЬ(12).

Результаты записываем в виде: среднее$\pm$ ошибка, минимум - медиана - максимум. Например, для температуры имеем: OH $(73,1 \pm 0,9; 71,9 - 72,1 - 76,4)$, TX $(84,5 \pm 0,4; 82,3 - 84,75 - 85,9)$, CA $(72,3 \pm 1,9; 59,1 - 72,85 - 81,9)$.

2. **Проверка однородности двух выборок.** Из предыдущих результатов видим, что теплее всего в штате TX, а в CA больше всего разброс относительно среднего значения. Эти различия могут оказаться случайными. Для того чтобы проверить это, используем статистические критерии.

В электронных таблицах в меню СЕРВИС используется раздел АНАЛИЗ ДАННЫХ. Если такого пункта нет, то в НАДСТРОЙКЕ нужно отметить строку ПАКЕТ АНАЛИЗА. В EXCEL2007 для этого имеется кнопка ПАРАМЕТРЫ EXCEL на странице открытия и сохранения файлов.

Критерий Фишера. Сначала проверяем гипотезу о равенстве дисперсий среднеиюльских температур $H_0 : \sigma_1^2 = \sigma_2^2$ для штатов TX и CA при наличии соответственных данных в ячейках $B2 : B9$ и $C2 : C13$. Путь: СЕРВИС - АНАЛИЗ ДАННЫХ - Двухвыборочный F-тест для дисперсий. Далее в окнах ИНТЕРВАЛ ПЕРЕМЕННОЙ 1 и ИНТЕРВАЛ ПЕРЕМЕННОЙ 2 записываем соответственно B2:B9 и C2:C13. Отмечаем ВЫХОДНОЙ ИНТЕРВАЛ, а в открывшемся поле указываем любую пустую ячейку, в которой будет записан результат счета (Табл.1).

Таблица 1

Двухвыборочный F-тест для дисперсии		
	Переменная 1	Переменная 2
Среднее	84,49	72,29
Дисперсия	1,36	41,53
Наблюдения	8	12
df	7	11
F	0,033	
P(F<=f) одностороннее	7,40636E-05	
F критическое одностороннее	0,28	

В табл.1 значимость $p = P(F <= f) = 7,4 \cdot 10^{-5} < 0,05$ свидетельствует о существенном различии по дисперсиям, поэтому для проверки гипотезы о равенстве средних нужно использовать критерий, действующий в предположении, что дисперсии различны. В случае, когда значимость $p = P(F <= f) > 0,05$ используется критерий, действующий в предположении, что дисперсии одинаковы.

Критерий Стьюдента. Проверяем гипотезу о равенстве средних июльских температур $H_0 : \mu_1 = \mu_2$ для штатов TX и CA при наличии соответственных данных в ячейках $B2 : B9$ и $C2 : C13$. Путь: СЕРВИС - АНАЛИЗ ДАННЫХ - Двухвыборочный t-тест с различными дисперсиями. Далее в окнах ИНТЕРВАЛ ПЕРЕМЕННОЙ 1 и ИНТЕРВАЛ ПЕРЕМЕННОЙ 2 записываем соответственно B2:B9 и C2:C13. Гипотетическая средняя разность должна быть равна 0. Также отмечаем ВЫХОДНОЙ ИНТЕРВАЛ, и в открывшемся поле указываем любую пустую ячейку, в которой будет записан результат счета (Табл.2).

Таблица 2

Двухвыборочный t-тест с различными дисперсиями		
	Переменная 1	Переменная 2
Среднее	84,49	72,29
Дисперсия	1,36	41,53
Наблюдения	8	12
Гипотетическая разность средних	0	
df	12	
t-статистика	6,40	
P(T<=t) одностороннее	1,70E-05	
t критическое одностороннее	1,78	
P(T<=t) двухстороннее	3,40E-05	
t критическое двухстороннее	2,18	

В качестве значимости p используем "$P(T <= t)$ двухстороннее", которое равно $p = 3,40E{-}05 < 0.05$, что свидетельствует о значимом различии среднеиюльских температур в штатах TX и CA.

3. **Проверка однородности трех выборок одновременно.** Для проверки гипотезы $H_0 : \mu_1 = \mu_2 = \mu_3$ используем путь: СЕРВИС - АНАЛИЗ ДАННЫХ - Однофакторный дисперсионный анализ. В окне ВХОДНОЙ ИНТЕРВАЛ указываем все три столбика одновременно A2:C13. ГРУППИРОВАНИЕ по столбцам. ВЫХОДНОЙ ИНТЕРВАЛ должен содержать любую пустую ячейку, с которой начнется запись результатов счета (табл.3-4).

Помимо итоговых характеристик основной интерес представляет так называемое P-значение, по которому следует ориентироваться в принятии гипотезы. В нашем случае $p = 2,49 \cdot 10^{-05} < 0.05$, поэтому гипотеза о равенстве средних отвергается, т.е. между средними июльскими температурами в штатах OH,TX и CA имеется значимое различие. В противном случае не было бы оснований отвергнуть гипотезу о равенстве средних и мы бы тогда сказали, что различие не значимо и может быть объяснено случайностью.

Таблица 3

Однофакторный дисперсионный анализ				
ИТОГИ				
Группы	Счет	Сумма	Среднее	Дисперсия
Столбец 1	5	365,5	73,1	3,695
Столбец 2	8	675,9	84,49	1,36
Столбец 3	12	867,5	72,29	41,53

Таблица 4

Дисперсионный анализ						
Источник вариации	SS	df	MS	F	P-Значение	F критическое
Между группами	780,20	2	390,10	17,84	2,49E-05	3,44
Внутри групп	481,14	22	21,87			
Итого	1261,34	24				

Поскольку выборки не велики, и согласие с нормальным распределением вряд ли окажется удовлетворительным, то для контроля необходимо проверить гипотезу однородности по критерию Краскела-Уоллеса.

Таблица 5

номер	штат	температура	ранги
1	OH	73,2	11
2	OH	71,9	5,5
3	OH	76,4	14
4	OH	72,1	7
5	OH	71,9	5,5
6	TX	83,5	19
7	TX	85,9	25
8	TX	85	22
9	TX	82,3	18
10	TX	84,5	21
11	TX	85,3	23,5
12	TX	85,3	23,5
13	TX	84,1	20
14	CA	74,3	12
15	CA	71	4
16	CA	69,5	3
17	CA	59,1	1
18	CA	73,1	10
19	CA	75,7	13
20	CA	81,9	17
21	CA	62,1	2
22	CA	72,6	8,5
23	CA	72,6	8,5
24	CA	77,9	16
25	CA	77,7	15

Объединяем три выборки в одну и ставим в соответствие каждому индивиду (городу) его ранг (место в вариационном ряду). Это несложно сделать, если правильно воспользоваться процедурой сортировки (табл.5). Не забудьте, что

одинаковым значениям присваиваются усредненные ранги. Формула для вычисления статистики Краскела-Уоллеса имеет вид:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{r} \frac{R_i^2}{n_i} - 3(n+1), \tag{1}$$

где через r обозначено количество групп ($r = 3$), $n_i,\ i = 1, \ldots, r$, количество элементов в каждой группе, $n_1 = 5, n_2 = 8, n_3 = 12$, $n = \sum\limits_{i=1}^{r} n_i = 25$ объем всей выборки, через $R_i,\ i = 1, \ldots, r$, обозначены суммы рангов в каждой группе. $R_1 = 43, R_1^2/5 = 369,8,\ R_2 = 172, R_1^2/8 = 3698,\ R_3 = 110, R_3^2/12 = 1008,33$. О статистике $H = 15,71$ известно, что при однородности групп она имеет распределение хи-квадрат с $r-1$ степенью свободы. Для вычисления p-значения используем функцию ХИ2РАСП$(15,71; 2) = 0,000387 < 0.05$. Отсюда делаем вывод, что различие между штатами по уровню средней июльской температуре значимо.

3. Для вычисления коэффициента корреляции

$$r = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum\limits_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

можно использовать функцию КОРРЕЛ(). Получаем при $n = 25$ значение коэффициента корреляции, равное $r = 0.44$. Положительность корреляции свидетельствует о том, что в более теплом климате города занимают большие площади. Для проверки значимости этой корреляции, то есть гипотезы о равенстве нулю генерального коэффициента корреляции $H_0 : \varrho = 0$ при альтернативной гипотезе $H_1 : \varrho \neq 0$ используем статистику

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2},$$

которая при справедливости H_0 имеет распределение Стьюдента с числом степеней свободы $df = n-2$. Вычисляем $t_* = 2,44$, доверительный уровень вероятности $p = P\{|t| > |t_*|\} =$ СТЬЮДРАСП $(abs(t_*); n-2; 2) = 0,023 < 0,05$. Гипотеза о некоррелированности признаков отвергается, то есть корреляция значима: в более теплом климате города действительно больше. При $p > 0.2$ нет оснований отвергнуть гипотезу о некоррелированности признаков, и в таком случае мы бы сказали, что корреляция не значима. При $0.05 < p < 0.20$ нельзя сказать ничего определенного, поскольку при увеличении объема выборки ситуация может повернуться как в одну, так и в другую сторону.