БИЛЕТЫ К ЭКЗАМЕНУ ПО ПРИКЛАДНОЙ СТАТИСТИКЕ

ВАРИАНТ 200013

1. Как записывается выборочное среднее для не сгруппированных данных?
2. Для чего нужно вычислять доверительный интервал оценки? Приведите содержательный пример, в котором возникает необходимость найти доверительный интервал.
3. Что такое критерии согласия?
4. С помощью какого критерия можно выявить связь между двумя количественными признаками?
5. В каких случаях в качестве меры близости между объектами используется обычное евклидово расстояние, а в каких — нормализованное евклидово?
6. Что характеризует коэффициент детерминации?

ЗАДАЧИ

1. В файле olimp.sf приведены результаты тестирования и олимпиады у большой группы абитуриентов. Можно ли утверждать, что олимпиада "сложнее", чем тестирование?
2. Есть ли корреляция между результатами олимпиады и тестирования?
3. Пусть по результатам олимпиады и тестирования в 1999 году Вам поручили провести зачисление. Требуется зачислить 82 человека. Проведите такое зачисление в SG. После проведения зачисления в приемной комиссии обнаружилось еще три абитуриента со следующими результатами: тест 93 74 49; олимпиада соответственно 30 70 50. Кого из них мы должны зачислить?

ВАРИАНТ 200014

1. Каков содержательный смысл распределения Бернулли?
2. Что такое статистическая гипотеза?
3. Какому условию должны удовлетворять выборки, чтобы можно было воспользоваться однофакторным дисперсионным анализом?
4. В каких случаях выборочную функцию регрессии следует искать в виде линейной функции?
5. Какая задача решается методами дискриминантного анализа?
6. Что можно сказать о регрессионной модели, если коэффициент Дарбина-Уотсона близок к четырем ?

ЗАДАЧИ

1. На заводе разработаны две новые технологии T_1 ,T_2. Чтобы оценить, как изменится дневная производительность при переводе на новые технологии, завод в течение 10 дней работал по каждой, включая существующую T_0. Дневная производительность в условных единицах приводится в таблице. Проверить гипотезу об отсутствии влияния технологии на производительность.

№	T_0	T_1	T_2	№	T_0	T_1	T_2
1	46	74	52	6	44	68	70
2	48	82	63	7	66	76	78
3	73	64	72	8	46	88	68
4	52	72	64	9	60	70	70
5	72	84	48	10	48	60	54

2. В файле **BankSwiss.sf3** приведены данные о размерах для 200 банкнот швейцарского банка по 1000 франков, из которых первые 100 – настоящие, следующие 100- фальшивые. (Описание переменных приведено в комментариях к ним). Пусть Вы не знаете, какие из 200 банкнот настоящие, а какие фальшивые. Проведите разбиение всех банкнот на настоящие и фальшивые. Добейтесь максимального совпадения результатов разбиения с истинными. Запишите все параметры метода, при котором получается наилучшее совпадение.

3. В таблице приведены данные об успеваемости 145 студентов первых трех курсов.

	удовл.	хор.	отл.
1 курс	45	25	15
2 курс	11	11	13
3 курс	9	9	17

Зависит ли успеваемость от курса?

ВАРИАНТ 200015

1. Как записывается несмещенная выборочная дисперсия для не сгруппированных данных
2. Что такое доверительная вероятность?
3. Какая гипотеза проверяется с помощью критерия согласия χ^2? Как следует группировать данные для применения этого критерия?
4. Что характеризует выборочный коэффициент корреляции?
5. Для каких признаков используется Хеммингово расстояние?
6. Чем различаются задачи, решаемые методами дискриминантного и кластерного анализа?

ЗАДАЧИ

1. В таблице приведена урожайность (ц/га) четырех сортов пшеницы (4 уровня фактора А) с использованием пяти типов удобрений (5 уровней фактора В); данные получены на 20 участках одинакового размера и почвенного состава.

Фактор *B* –тип удобрения	Фактор *A*			
	A_1	A_2	A_3	A_4
B_1	19	25	17	21
B_2	22	19	19	18
B_3	26	23	22	25
B_4	18	26	20	23
B_5	21	22	21	24

Зависит ли урожайность от сорта пшеницы? от вида удобрения ?

2. В файле **rost_razmer.sf** переменные **rost** , **obuv**, **ves** приведены данные о росте, размере обуви и весе студентов 4-го курса. Зависит ли размер обуви от роста? от веса? Пусть рост некого студента равен 172 см, а вес — 90 кг. Спрогнозируйте его размер обуви, оцените качество регрессионной модели. Пусть у двух студентов одинаковый рост, а вес одного на девять килограмм больше другого. Как различаются (в среднем) их размеры обуви?

3. В таблице приведены данные об распределении цвета волос на голове и бровей у 46542 человек

Цвет бровей	Цвет волос на голове	
	светлые	темные
Светлые	30472	3238
Темные	3364	9468

Можно ли считать, что данные признаки независимы ?

ВАРИАНТ 200016

1. Каков содержательный смысл распределения равномерного распределения?
2. Что такое простая гипотеза? сложная гипотеза?
3. Что дают критерии Барлетта и Кочрена для однофакторного анализа?
4. Какой метод используется для нахождения коэффициентов линейной функции регрессии?
5. В чем состоит идея метода Варда ?
6. Что можно сказать о регрессионной модели, если коэффициент Дарбина-Уотсона близок к четырем ?

ЗАДАЧИ

1. В файле **BankSwiss.sf3** приведены данные о размерах для 200 банкнот швейцарского банка по 1000 франков, из которых первые 100 – настоящие, следующие 100- фальшивые. (Описание переменных приведено в комментариях к ним). В банк поступили две банкноты – их номера 201 и 202. Есть ли среди этих банкнот фальшивые?

2. Для определения процентного содержания вредных примесей в минерале были взяты образцы одинаковой массы из трех различных месторождений: 3 образца из 1-го, 2 из 2-го и 4 из третьего. Результаты химического анализа (в процентах) даны в таблице

1 мест	2 мест	3 мест
8,35	4,52	8,91
5,40	6,24	7,47
7,16		9,08
		9,94

Можно ли считать, что среднее содержание примесей на всех трех месторождениях одинаковы?

3.В файле **CO_2.sf** приведены данные о содержании углекислого газа в воздухе в зависимости от ряда факторов (описание переменных см. в комментариях). Найдите обобщенные переменные, с помощью которых можно было бы прогнозировать загрязненность воздуха. Дайте их содержательную интерпретацию.

ВАРИАНТ 200017

1. Что такое мода (можно на примере) ?
2. Как записывается доверительный интервал для мат.ожидания?
3. Какие критерии проверки однородности Вы знаете для независимых (непарных) наблюдений?
4. Что характеризует частный коэффициент корреляции?
5. Пусть для построенной регрессионной модели R^2 близок к единице. Что еще следует проверить, прежде чем использовать эту модель для прогноза?
6. Что Вы понимаете под параметрическими методами дискриминантного анализа?

ЗАДАЧИ

1. В файле **olimp.sf** приведены результаты тестирования и олимпиады у большой группы абитуриентов. Можно ли утверждать, что олимпиада "сложнее", чем тестирование?
2. Есть ли корреляция между результатами олимпиады и тестирования?
3. Пусть по результатам олимпиады и тестирования в 1999 году Вам поручили провести зачисление в 2000 году. Предполагается, что уровень сложности не меняется. Известно, что в 1999 году было зачислено 82 человека. Используя этот результат как обучающую выборку, построить дискриминантную функцию и по ней провести зачисление трех абитуриентов со следующими результатами: тест 93 74 49; олимпиада соответственно 30 70 50.

ВАРИАНТ 200018

1. Почему в приложениях чаще других встречается нормальное распределение?
2. Что такое параметрическая гипотеза? Приведите пример.
3. Что вы понимаете под "критериями согласия" ?
4. Что характеризует коэффициент детерминации R^2?
5. Что можно сказать о регрессионной модели, если коэффициент Дарбина-Уотсона близок к двум ?
6. Пусть по выборке из нормальной совокупности построен 95% доверительный интервал для математического ожидания. Можно ли утверждать, что около 95% значений изучаемой случайной величины будут лежать в этом интервале? (Да - нет, почему ?)

ЗАДАЧИ

1. На заводе разработаны две новые технологии T_1 ,T_2. Чтобы оценить, как изменится дневная производительность при переводе на новые технологии, завод в течение 10 дней работал по каждой, включая существующую T_0. Дневная производительность в условных единицах приводится в таблице. Проверить гипотезу об отсутствии влияния технологии на производительность.

№	T_0	T_1	T_2	№	T_0	T_1	T_2
1	46	74	52	6	44	68	70
2	48	82	63	7	66	76	78
3	73	64	72	8	46	88	68
4	52	72	64	9	60	70	70
5	72	84	48	10	48	60	54

2. В файле **CO_2.sf** находятся данные о содержании углекислого газа в водухе в некоторых городах США. Выяснить, от чего зависит содержание углекислого газа. Как можно сделать прогноз по содержанию CO2?
3. В таблице приведены данные об распределении цвета волос на голове и бровей у 46542 человек

Цвет бровей	Цвет волос на голове	
	Светлые	Темные
Светлые	30472	3238
Темные	3364	9468

Можно ли считать, что данные признаки независимы ?

ВАРИАНТ 200019

1. Что такое медиана (можно на примере)?
2. Какое распределение используется для построения доверительного интервала для мат.ожидания?
3. Какие критерии проверки однородности Вы знаете для парных наблюдений?
4. Что Вы понимаете под порядковым признаком?
5. Как записывается расстояние между двумя кластерами по принципу «ближнего соседа»?
6. Какой метод используется для определения коэффициентов линейной выборочной функции регрессии?

ЗАДАЧИ

1. В таблице приведены результаты (количество ошибок на странице), полученные при наборе текста на компьютере в зависимости от стажа работы пользователя (каждый набирал по 4 страницы):

Стаж			
1		*2*	
30	5	2	0
27	8	5	1
23	1	8	3
15	16	11	5

Зависит ли качество данного вида работы от стажа?

2. Проведено обследование с целью определения наличия взаимосвязи между возрастом вступления в первый брак и уровнем доходов молодоженов. Результаты представлены в таблице.

Уровень доходов	Возраст		
	До 18	*18-21*	
Низкий	45	25	15
Средний	35	60	25
Высо-	10	8	24

Каковы Ваши выводы?

3. В файле **BankSwiss.sf** приведены данные о размерах для 200 банкнот швейцарского банка по 1000 франков, из которых первые 100 – настоящие, следующие 100- фальшивые. (Описание переменных приведено в комментариях к ним). В банк поступили две банкноты – их номера 201 и 202. Есть ли среди этих банкнот фальшивые?

ВАРИАНТ 200020

1. Статистическим аналогом какой вероятностной кривой является гистограмма частости? Полигон частости? Кумулята?
2. Что такое ошибка первого рода? второго рода при проверке статистических гипотез?
3. Каким критерием следует воспользоваться, если при однофакторном анализе Вы обнаружили, что нет нормальности?
4. Что такое остаточная дисперсия? Что она характеризует?
5. Как связаны собственные значения и собственные вектора ковариационной матрицы с главными компонентами?
6. Пусть по выборке из нормальной совокупности построен 95% доверительный интервал для математического ожидания. Можно ли утверждать, что около 95% элементов выборки будут лежать в этом интервале? (Да - нет, почему ?)

ЗАДАЧИ

1. В файле **BankSwiss.sf** приведены данные о размерах для 200 банкнот швейцарского банка по 1000 франков, из которых первые 100 – настоящие, следующие 100- фальшивые. (Описание переменных приведено в комментариях к ним). Пусть Вы не знаете, какие из 200 банкнот настоящие, а какие фальшивые. Проведите разбиение всех банкнот на настоящие и фальшивые. Добейтесь максимального совпадения результатов разбиения с истинными. Запишите все параметры метода, при котором получается наилучшее совпадение.

2. Для определения процентного содержания вредных примесей в минерале были взяты образцы одинаковой массы из трех различных месторождений: 3 образца из 1-го, 2 из 2-го и 4 из третьего. Результаты химического анализа (в процентах) даны в таблице

1 мест	2 мест	3 мест
8,35	4,52	8,91
5,40	6,24	7,47
7,16		9,08
		9,94

Можно ли считать, что среднее содержание примесей на всех трех месторождениях одинаковы?

3. В файле **Product. sf** приведены данные по 45 предприятиям легкой промышленности по статистической связи между стоимостью основных фондов *(fonds,* млн руб.) и средней выработкой на 1 работника *(product,* тыс. руб.); z - вспомогательный признак: z = 1 - федеральное подчинение, z = 2 - муниципальное

У некоторого предприятия основные фонды равны 31.5 млн.руб.

1)Спрогнозировать среднюю выработку на данном предприятии без учета Z.

2) Сделать прогноз, если известно, что предприятие федерального подчинения.

ВАРИАНТ 200021

1. Что характеризуют асимметрия и эксцесс?
2. В каком случае требование нормальности распределения изучаемой случайной величины существенно?
3. В чем «идея» критерия знаков?
4. С помощью какого критерия можно выявить связь между двумя порядковыми признаками?
5. Как записывается расстояние между двумя кластерами по принципу «дальнего соседа»?
6. Какие функции используются в качестве дискриминантных в параметрическом методе?

ЗАДАЧИ

1. В файле **olimp.sf** приведены результаты тестирования и олимпиады у большой группы абитуриентов. Можно ли утверждать, что олимпиада "сложнее", чем тестирование?
2. Есть ли корреляция между результатами олимпиады и тестирования?
3. Пусть по результатам олимпиады и тестирования в 1999 году Вам поручили провести зачисление в 2000 году. Предполагается, что уровень сложности не меняется. Известно, что в 1999 году было зачислено 82 человека. Проведите такое зачисление в SG. После проведения зачисления в приемной комиссии обнаружилось еще три абитуриента со следующими результатами: тест 93 74 49; олимпиада соответственно 30 70 50. Кого из них мы должны зачислить ?

ВАРИАНТ 200022

1. Что такое распределение Стьюдента?
2. Что такое наилучшая критическая область при проверке статистических гипотез?
3. Что характеризует частный коэффициент корреляции ?
4. Что можно сказать про остаточную дисперсию, если выборочный коэффициент корреляции близок к 1?
5. Как выбираются векторы главных компонент в k-мерном пространстве?
6. Пусть по выборке из нормальной совокупности построен 95% доверительный интервал для математического ожидания. Можно ли утверждать, что около 95% значений изучаемой случайной величины будут лежать в этом интервале? (Да - нет, почему ?)

ЗАДАЧИ

1. На заводе разработаны две новые технологии T_1 ,T_2. Чтобы оценить, как изменится дневная производительность при переводе на новые технологии, завод в течение 10 дней работал по каждой, включая существующую T_0. Дневная производительность в условных единицах приводится в таблице. Проверить гипотезу об отсутствии влияния технологии на производительность.

№	T_0	T_1	T_2	№	T_0	T_1	T_2
1	46	74	52	6	44	68	70
2	48	82	63	7	66	76	78
3	73	64	72	8	46	88	68
4	52	72	64	9	60	70	70
5	72	84	48	10	48	60	54

2. В файле **CO_2.sf** приведены данные о содержании углекислого газа в воздухе в зависимости от ряда факторов (описание переменных см. в комментариях). Найдите обобщенные переменные, с помощью которых можно было бы прогнозировать загрязненность воздуха. Дайте их содержательную интерпретацию.

3. Проведено обследование с целью определения наличия взаимосвязи между возрастом вступления в первый брак и уровнем доходов молодоженов. Результаты представлены в таблице.

Уровень доходов	Возраст		
	До 18		*18–21*
Низкий	45	25	15
Средний	35	60	25
Высо-	10	8	24

Каковы Ваши выводы?

ВАРИАНТ 200023

1. Для чего используется коэффициент вариации?
2. Какое распределение используется при построении доверительного интервала для дисперсии?
3. В чем «идея» критерия знаковых ранговых сумм ?
4. Для чего используются коэффициенты Спирмена и Кэнделла?
5. Как записывается расстояние между двумя кластерами с использованием расстояния «по центрам тяжести»?
6. При каком предположении в параметрическом методе дискриминантные функции получаются линейными?

ЗАДАЧИ

1. В таблице приведена урожайность (ц/га) четырех сортов пшеницы (4 уровня фактора А) с использованием пяти типов удобрений (5 уровней фактора В); данные получены на 20 участках одинакового размера и почвенного состава.

Фактор *B* –тип удобрения	Фактор *A*			
	A_1	A_2	A_3	A_4
B_1	19	25	17	21
B_2	22	19	19	18
B_3	26	23	22	25
B_4	18	26	20	23
B_5	21	22	21	24

Зависит ли урожайность от сорта пшеницы? от вида удобрения ?

2. В файле **BankSwiss.sf3** приведены данные о размерах для 200 банкнот швейцарского банка по 1000 франков, из которых первые 100 – настоящие, следующие 100- фальшивые. (Описание переменных приведено в комментариях к ним). Пусть Вы не знаете, какие из 200 банкнот настоящие, а какие фальшивые. Проведите разбиение всех банкнот на настоящие и фальшивые. Добейтесь максимального совпадения результатов разбиения с истинными. Запишите все параметры метода, при котором получается наилучшее совпадение.

3. В файле **CO_2.sf** находятся данные о содержании углекислого газа в водухе в некоторых городах США. Выяснить, от чего зависит содержание углекислого газа. Как можно сделать прогноз по содержанию CO2?

ВАРИАНТ 200024

1. Что такое распределение χ^2? Фишера?
2. Что такое наилучшая критическая область (область принятия решения)?
3. В чем состоят основная и альтернативная гипотезы в однофакторном дисперсионном анализе?
4. Для чего используется критерий Дарбина-Уотсона?
5. В чем состоит идея агломерационных методов кластерного анализа?
6. Как можно определить, являются ли остатки в регрессионном анализе коррелированными?

ЗАДАЧИ

1.В файле **BankSwiss.sf3** приведены данные о размерах для 200 банкнот швейцарского банка по 1000 франков, из которых первые 100 – настоящие, следующие 100- фальшивые. (Описание переменных приведено в комментариях к ним). Пусть Вы не знаете, какие из 200 банкнот настоящие, а какие фальшивые. Проведите разбиение всех банкнот на настоящие и фальшивые. Добейтесь максимального совпадения результатов разбиения с истинными. Запишите все параметры метода, при котором получается наилучшее совпадение.

2. В файле **BankSwiss.sf3** приведены данные о размерах для 200 банкнот швейцарского банка по 1000 франков, из которых первые 100 – настоящие, следующие 100- фальшивые. (Описание переменных приведено в комментариях к ним). В банк поступили две банкноты – их номера 201 и 202. Есть ли среди этих банкнот фальшивые?

3. Для определения процентного содержания вредных примесей в минерале были взяты образцы одинаковой массы из трех различных месторождений: 3 образца из 1-го, 2 из 2-го и 4 из третьего. Результаты химического анализа (в процентах) даны в таблице

1 мест	2 мест	3 мест
8,35	4,52	8,91
5,40	6,24	7,47
7,16		9,08
		9,94

Можно ли считать, что среднее содержание примесей на всех трех месторождениях одинаковы?