Программа курса
ПРИКЛАДНАЯ СТАТИСТИКА

Описание курса:

Прикладная статистика – семестровый курс для студентов четвертого года обучения УрГУ. Это - курс для студентов, специализирующихся в области бизнес-информатики, прослушавших стандартный курс теории вероятностей.

В курсе изучаются базовые понятия статистики: описательные статистики, понятие генеральной совокупности и выборки, оценивание параметров, статистическая проверка гипотез и т.п. Особенностью курса является использование на лекциях и в лабораторных занятиях одного из статистических пакетов Statgraphics Centurion, что дает возможность работать не только с модельными, но и с реальными прикладнми задачами.

Цели курса:

Основная цель курса – дать студен там систематические знания в области прикладной статистики. Они должны понимать предмет и освоить основные методы статистического анализа. Студенты должны научиться проводить разведочный анализ данных (находить среднее, медиану, среднеквадратичное отклонение и другие описательные статистики), представлять данные графически. У них должно сложиться понимание различия между генеральной совокупностью и выборкой и, соответственно, между теоретическими и выборочными характеристиками.

Студенты должны научиться формулировать и решать традиционные задачи прикладной статистики: оценивание параметров, статистическая проверка гипотез. При проверке гипотез студент должен научиться определять, в каких случаях следует использовать параметрические, а в каких непараметрические критерии. Изучаются методы выявления связей для различных типов данных. Одна из целей курса знакомство с элементами многомерного статистического анализа: кластерный и дискриминантный анализ, метод главных компонент, основы факторного анализа.

Другая цель курса подготовка студентов к изучению эконометрики на основе изучения простейших моделей парной и множественной регрессии, элементарного анализа временных рядов.

Курс не является математически строгим. Как следствие, доказательства многих теорем и даже точная формулировка результатов обычно опускаются. Важной частью курса является решение прикладных задач. В основе задач – попытка проиллюстрировать различные способы применения теории на практике. В процессе обучения студенты также выполняют компьютерные

задания с реальными данными, вырабатывают практические навыки и интуицию.

По окончании курса студенты должны понимать теорию, лежащую в основе статистической науки, уметь выполнять необходимые вычисления с использованием компьютера и применять стандартные методы на практике.

Методы:

В курсе используются следующие методы и формы работы:
- Лекции (3 часа в неделю)
- Лабораторные занятия (2 часа в неделю)
- Консультации преподавателя
- Еженедельные письменные задания перед выполнением очередной лабораторной работой
- Промежуточная письменная контрольная работа
- Самостоятельная работа с литературой.

На лекциях используется компьютер с проектором, лабораторные занятия производятся в компьютерном классе.

Курс включает 54 часа лекций и 36 часов лабораторных работ.

Принципы оценки работы студентов:

Перед каждой лабораторной работой, студент должен в течение пяти минут ответить на один из вопрос, из списка вопросов по изучаемой теме, приведенных в «Практикуме по прикладной статистике». В середине семестра студенты сдают промежуточный письменный экзамен, а также итоговый письменный экзамен в конце семестра. Экзамен включает набор из шести содержательных вопросов по теории, на которые следует дать короткие ответы, а также четыре небольших прикладных задачи, которые следует решить, используя пакет Statgraphics. . Итоговый экзамен дает 60% конечной оценки за первый семестр, промежуточный экзамен - 20% конечной оценки соответственно. 20% дается за выполнение контрольных вопросов на лабораторных занятиях и работу на занятиях.

Содержание курса:

Разделы курса, темы, их краткое содержание

1. **Обзор общих и специализированных прикладных программных средств, используемых в статистике.**
2. **Предмет, метод и задачи статистики.**

Выборочный метод в статистике. Репрезентативность и однородность выборки.

3. Статистическая сводка и группировка.

Понятие статистической сводки и группировки. Виды группировок. Статистические ряды распределения. Графическое изображение статистических данных. Основные элементы статистических графиков. Классификация статистических графиков. Основные виды графиков, используемые в статистике.

4. **Основные виды распределений, используемых в статистике. Статистические таблицы.**

5 Точечные оценки параметров распределения. Критерии качества оценок: состоятельность, несмещенность и эффективность оценок.

Обобщающие статистические показатели. Оценки среднего и вариации.

6. Методы получения оценок: метод максимального правдоподобия и метод моментов.

7. Интервальные оценки параметров. Доверительные интервалы для математического ожидания и дисперсии.

8. Проверка статистических гипотез.

. Простые и сложные гипотезы. Параметрические и непараметрическиекритерии.Ошибки первого и второго рода. Понятие наилучшей критической области. Типичные задачи_проверки гипотез о математических ожиданиях. Парные иепарные наблюдения. Однофакторный дисперсионный анализ. Критерий Краскелла-Уоллеса. Критерии согласия: критерий χ^2 Пирсона, критерий Колмогорова-Смирнова**.** Проверка однородности выборок. Критерии знаков и знаковых ранговых сумм при проверке однородности для парных наблюдений. Критерий Манна-Уиттни для непарных наблюдений.

9. Выявление связей между признаками. Элементы **корреляционного анализа.**

Выявление связей между качественными признаками. Коэффициенты контингенции и Крамера. Выявление связей для порядковых признаков. Коэффициенты Спирмена и Кэндела. Выявление связей для количественных признаков. Выборочный коэффициент корреляции.

10. Элементы регрессионного анализа.

Теоретическая и выборочная функция регрессии. Метод наименьших квадратов. Линейная выборочная регрессия. Типичные нелинейные регрессионные модели, сводящиеся к линейным. Оценка качества модели. Коэффициент детерминации. Анализ остатков. Значимость коэффициентов.

11. Многомерные статистические методы.

Множественный корреляционный анализ. Парные, частные и множественные коэффициенты корреляции. Модель множественной регессии. Теорема Гаусса-Маркова. Оценка качества модели. Исправленный коэффициент детерминации. Анализ остатков, оценка значимости коэффициентов. Мультиколлинеарность. Причины появления и следствия. Пошаговый отбор

переменных. Метод главных компонент. Метод главных компонент как средство борьбы с мультиколлинеарностью. Постановка задачи факторного анализа. Элементы кластерного анализа. Дискриминантный анализ.

13. Элементы анализа временных рядов.

1. Темы лабораторных занятий.

№ п/п	Темы лабораторных занятий
1.	Первичная обработка и графическое представление статистических данных.
2.	Связь статистики с теорией вероятностей.
3.	Доверительные интервалы.
4.	Проверка статистических гипотез.
5.	Критерии согласия.
6.	Однофакторный дисперсионный анализ.
7.	Выявление связей между признаками.
8.	Простая регрессия.
9.	Множественная регрессия.
10.	Метод главных компонент.
11.	Кластерный анализ.
12.	Дискриминантный анализ.
13.	Временные ряды.

2. Перечень примерных контрольных вопросов и заданий для самостоятельной работы

- Что Вы понимаете под репрезентативностью выборки?
- Что такое гистограмма частостей, статистическим аналогом чего она является?
- Что такое кумулята частостей, статистическим аналогом чего она является?
- Как записывается выборочное среднее для не сгруппированных данных?
- Как записывается выборочное среднее для сгруппированных данных?
- Как записывается несмещенная выборочная дисперсия для не сгруппированных данных ?
- Что такое выборочная мода (можно на примере)? Оценкой какого параметра она является?
- Что такое выборочная медиана (можно на примере)? Оценкой какого параметра она является?
- Что характеризуют асимметрия и эксцесс? Как записываются выборочные асимметрия и эксцесс?
- Для чего используется коэффициент вариации?

- Каков содержательный смысл распределения Бернулли? Приведите пример сл.в., имеющей распределение Бернулли.
- Каков содержательный смысл распределения равномерного распределения? В какой типичной ситуации оно появляется?
- Что такое нормальное распределение? В какой типичной ситуации оно появляется?
- Что происходит с графиком плотности нормального распределения если увеличивать мат.ожидание? Дисперсию?
- Что такое распределение Стьюдента? Что происходит с графиком плотности распределения Стьюдента при увеличении числа степеней свободы?
- Что такое распределение χ^2? Что происходит с графиком плотности распределения χ^2 при увеличении числа степеней свободы?
- Что такое распределение Фишера?
- Что такое доверительный интервал? Для чего он нужен?
- Какое распределение используется при построении доверительного интервала для матожидания? Как записывается доверительный интервал для матожидания?
- Во сколько раз следует увеличить объем выборки, чтобы на порядок уменьшить длину доверительного интервала для мат ожидания?
- В каком случае при построении доверительного интервала требование нормальности существенно?
- Какое распределение используется при построении доверительного интервала для дисперсии?
- Что происходит с длиной доверительного интервала при увеличении доверительной вероятности?
- Что такое статистическая гипотеза?
- Что такое параметрическая гипотеза? Приведите пример.
- Что такое непараметрическая гипотеза? Приведите пример.
- Что такое простая гипотеза? сложная гипотеза?
- Что такое критическая область?
- Что такое наилучшая критическая область (область принятия решения)?
- Что такое ошибка первого рода? второго рода при проверке статистических гипотез?
- Что происходит с вероятностью ошибки второго рода при уменьшении вероятности ошибки первого рода?
- Что такое критерии согласия?
- Какая гипотеза проверяется с помощью критерия согласия χ^2 Пирсона? Как следует группировать данные для применения этого критерия?
- Параметрические или непараметрические гипотезы проверяются с помощью критерия Пирсона? Обоснуйте ответ.
- В чем «идея» критерия знаков?

- В чем «идея» критерия знаковых ранговых сумм?
- В чем разница между парными и независимыми наблюдениями? Приведите примеры.
- Какие критерии проверки однородности Вы знаете для парных наблюдений?
- Какие критерии проверки однородности Вы знаете для независимых (непарных) наблюдений?
- В чем состоят основная и альтернативная гипотезы в однофакторном дисперсионном анализе?
- Каким условиям должны удовлетворять выборки, чтобы можно было воспользоваться однофакторным дисперсионным анализом?
- Что дают критерии Барлетта и Кочрена для однофакторного анализа?
- Таблицы какого распределения используются для принятия решения в одно-(много) факторном дисперсионном анализе?
- Каким критерием следует воспользоваться, если при однофакторном анализе Вы обнаружили, что нет нормальности?
- С помощью какого критерия можно выявить связь между двумя качественными признаками?
- Что характеризует коэффициент Крамера?
- Что Вы понимаете под порядковым признаком?
- С помощью какого критерия можно выявить связь между двумя порядковыми признаками?
- Для чего используются коэффициенты Спирмена и Кэнделла?
- Что характеризует выборочный коэффициент корреляции?
- С помощью какого критерия можно выявить связь между двумя количественными признаками?
- Что такое «ложная корреляция»? Приведите пример.
- Что характеризует частный коэффициент корреляции?
- Что такое функция регрессии? Выборочная функция регрессии?
- В каких случаях выборочную функцию регрессии следует искать в виде линейной функции?
- Какой метод используется для нахождения коэффициентов линейной выборочной функции регрессии?
- Что такое остаточная дисперсия? Что она характеризует?
- Что можно сказать про остаточную дисперсию, если выборочный коэффициент корреляции близок к 1? ; к -1 ?
- Что характеризует коэффициент детерминации R^2?
- Что происходит с коэффициентом детерминации, если в модели увеличивается число независимых переменных?
- Как оценить качество выбранной регрессионной модели?
- Что следует проверить при анализе остатков?
- В чем состоит задача кластерного анализа?

- В каких случаях в качестве меры близости между объектами используется обычное евклидово расстояние, а в каких — нормализованное евклидово?
- Для каких признаков обычно используется Хеммингово расстояние?
- Что можно использовать в качестве расстояния между признаками (не объектами)?
- Как записывается расстояние между двумя кластерами по принципу «ближнего соседа»?
- Как записывается расстояние между двумя кластерами по принципу «дальнего соседа»?
- Как записывается расстояние между двумя кластерами с использованием расстояния «по центрам тяжести»?
- В чем состоит идея агломерационных методов кластерного анализа?
- В чем состоит идея метода Варда?
- Что такое дендрограмма (можно на примере) ?
- Как выбираются векторы главных компонент в к-мерном пространстве?
- Для чего используется метод главных компонент?
- Как связаны собственные значения и собственные вектора ковариационной матрицы с главными компонентами?
- Как выбрать количество оставляемых главных компонент?
- Почему метод главных компонент можно использовать как средство борьбы с мультиколлинеарностью? Каким образом?
- В чем состоит цель факторного анализа?
- Какая задача решается методами дискриминантного анализа?
- Чем различаются задачи, решаемые методами дискриминантного и кластерного анализа?
- Что Вы понимаете под непараметрическими методами дискриминантного анализа?
- Что Вы понимаете под параметрическими методами дискриминантного анализа?
- Какие функции используются в качестве дискриминантных в параметрическом методе?
- При каком предположении в параметрическом методе дискриминантные функции получаются линейными?

На самостоятельное изучение тем содержания дисциплины выделяется 18часов.

1. **Примерный перечень вопросов к экзамену.**

- Понятие выборки и генеральной совокупности. Количественные и качественные признаки. Дискретные и непрерывные случайные величины.
- Сущность выборочного метода исследования. Каким требованиям должны удовлетворять выборки из генеральной совокупности?
- Задачи, решаемые методами математической статистики.
- Понятие вариационного ряда, интервального вариационного ряда. Полигон и гистограмма распределения.
- Эмпирическая функция распределения и ее график.
- Статистические аналоги теоретических законов распределения.
- Понятие точечной оценки. Свойства точечных оценок.
- Точечные оценки генерального математического ожидания. Какие свойства имеют эти оценки.
- Точечные оценки генеральной дисперсии. Свойства этих оценок.
- Распределение “хи-квадрат”, распределение Стьюдента, распределение Фишера.
- Понятие интервальной оценки, доверительного интервала. Уровень значимости, надежность оценки.
- Доверительный интервал для генерального математического ожидания при известной и неизвестной дисперсии.
- Доверительный интервал для генеральной дисперсии при большом и малом объеме выборки.
- Доверительный интервал для вероятности случайного события в схеме Бернулли при большом объеме выборки.
- Понятие статистической гипотезы. Виды статистических гипотез. Процедура проверки статистических гипотез.
- Статистический критерий. Критическая область. Уровень значимости.
- Проверка статистических гипотез о числовом значении математического ожидания генеральной совокупности.
- Проверка статистических гипотез о числовом значении дисперсии генеральной совокупности.
- Проверка статистических гипотез о равенстве математических ожиданий двух генеральных распределений.
- Проверка гипотезы о равенстве дисперсий двух генеральных распределений.
- Проверка гипотезы о законе распределения генеральной совокупности, критерий Пирсона (критерий согласия)
- Критерий согласия Пирсона для проверки гипотезы нормальности распределения.
- Точечная оценка коэффициента корреляции двух нормально распределенных случайных величин.
- Проверка гипотезы об отсутствии корреляционной связи между двумя величинами.

- Эмпирическая линия регрессии. Простая регрессия. Выборочная оценка коэффициентов регрессии.
- Множественная регрессия.
- Метод главных компонент.
- Кластерный анализ.
- Дискриминантный анализ.
- Временные ряды.

III. Распределение часов курса по темам и видам работ

№ п/п	Наименование разделов и тем	ВСЕГО (часов)	Аудиторные занятия (час)		Самостоятельная работа
			в том числе		
			Лекции	Практические (семинары, лабораторные работы)	
1	Выборочный метод в статистике. Первичная обработка данных.		2	2	2
2	Точечные оценки параметров.		4	2	2
3	Доверительные интервалы.		4	2	2
4	Проверка статистических гипотез.		8	4	4
5	Выявление связей между признаками.		6	4	4
6	Простая регессия.		4	4	4
7	Множественная регрессия.		4	4	4
8	.Метод главных		6	4	4

	компонент. Факторный анализ.				
9	Кластерный анализ.		4	2	2
10	Дискриминантный анализ.		6	4	4
11	Временные ряды. Выделение тренда. Анализ остатков..		6	4	4
	ИТОГО:		54	36	36

Рекомендуемая литература

Основная литература:

1. Максимов Ю.Д. Математика Выпуск 8. Математическая статистика. Опорный конспект. СПб., Изд-во СПбГПУ, 2002, 96 с.
2.Математическая статистика. Математика в техн. Университете. Т. XVII/ под ред. Зарубина В.С., Крищенко А.П., М., Изд. МГТУ им Н.Э.Баумана, 2001, 424с.
3. Колемаев В.А, Калинина В.Н., Теория вероятностей и математическая статистика. М., ИНФРА-М, 1997.

Дополнительная литература:

1. С.А. Айвазян, В.С. Мхитарян. Теория вероятностей и прикладная статистика. ЮНИТИ. Москва, 2001.
2. Гмурман В. Е. Теория вероятностей и математическая статистика. М., “Высшая школа”, 1998.
3. Гмурман В. Е. Руководство к решению задач по теории вероятностей и математической статистике. М., “Высшая школа”, 1998.
4. Шведов А.С. Теория вероятности и математическая статистика. ВШЭ, 1995.
5. Сигел Э., Практическая бизнес-статистика. М., Изд. дом «Вильямс», 2004.
6. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. М., «Финансы и статистика», 1998.
7. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. М., ИНФРА-М, 2003, 544с.
8. Дюк В. Обработка данных на ПК в примерах. СПб., 1997, 240с.

II. Ресурсное обеспечение

- Случайные выборки. Двойственность интерпретации. Понятие об оценивании параметров распределения. Выборочное среднее и выборочная дисперсия. Математическое ожидание и дисперсия выборочного среднего. Оценивание пропорций.

- Точечное оценивание. Свойства оценок: несмещенность, эффективность, состоятельность. Оценки среднего и дисперсии.
- Интервальное оценивание. Доверительные интервалы. Оценивание среднего. Нормальная аппроксимация при больших выборках. Случай малых выборок (распределение Стьюдента). Разница двух средних. Пропорции.
- Тестирование гипотез. С использованием доверительных интервалов. С использованием тест-статистик. Двусторонние и односторонние *p*-значения