Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования
Московский государственный университет имени М.В. Ломоносова
Геологический факультет

УТВЕРЖДАЮ
Декан Геологического факультета
академик
_____/Д.Ю.Пущаровский/
«___» _____20 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Банки и базы данных в палеобиологии

Авторы-составители: Барсков И.С., Беляев Р.И.

Уровень высшего образования:

Магистратура

Направление подготовки:

05.04.01 Геология

Направленность (профиль) ОПОП:

Геология и полезные ископаемые

Форма обучения:

Очная

Рабочая программа рассмотрена и одобрена
Учебно-методическим Советом Геологического факультета
(протокол № _____, _____)

Москва

Рабочая программа дисциплины (модуля) разработана в соответствии с самостоятельно установленным МГУ образовательным стандартом (ОС МГУ) для реализуемых основных профессиональных образовательных программ высшего образования по направлению подготовки «Геология» (*программы бакалавриата, магистратуры, реализуемых последовательно по схеме интегрированной подготовки)* в редакции приказа МГУ от 30 декабря 2016 г.

Год (годы) приема на обучение – 2019.

Цель и задачи дисциплины

Целью дисциплины «Банки и базы данных в палеобиологии» является формирование у обучающихся навыков работы с большими массивами данных и их статистического анализа.

Задачи: — формирование представлений о логической структуре баз данных и шкалирование как процедуре перевода исследуемого объекта в формальную (математическую) модель;

— ознакомление с современными палеобиологическими базами данных;

— получение знаний о репрезентативности данных, генеральной и выборочной совокупности, точечных и интервальных оценках;

— освоение методик анализа данных (понимание математической процедуры, ограничения налагаемые на используемые шкалы, формирование навыков применения метода при решения фактических исследовательских задач).

1. Место дисциплины в структуре ОПОП ВО — вариативная часть, профессиональный цикл, обязательные дисциплины, курс – I, семестр – 2.

2. Входные требования для освоения дисциплины, предварительные условия:

Знание ответов на вопросы по вступительным испытаниям в магистратуру.

Дисциплина необходима для научно-исследовательской работы и выполнения выпускных квалификационных работ.

3. Результаты обучения по дисциплине, соотнесенные с требуемыми компетенциями выпускников.

Компетенции выпускников, формируемые (полностью или частично) при реализации дисциплины:

ОПК-5.М Способность использовать современные вычислительные методы и компьютерные технологии для решения задач профессиональной деятельности,

ПК-9.М Способность использовать современные методы обработки и интерпретации комплексной информации для решения производственных задач,

СПК-3.М Способность работать в профильных геологических, биологических и краеведческих музеях и проводить исследования в камеральный и полевой период, как в целом по палеонтологии и стратиграфии, так и по основным их разделам: палеоэкологии, микропалеонтологии, палеоботанике, палеозоологии позвоночных.

Планируемые результаты обучения по дисциплине (модулю):

Знать: принципы построения баз данных; основные типы шкал; понятие репрезентативности, генеральной и выборочной совокупности, точечной и интервальной оценки; основные описательные статистики, включая меры средней тенденции и разброса; понимать процедуру проверки статистических гипотез; основные методики анализа данных;

Уметь: составлять собственные базы данных, осуществляя процедуру шкалирования; осуществлять поиск необходимой информации; определять особенности распределения изучаемой величины; использовать описательные статистики; владеть приемами графического анализа данных; рассчитывать доверительные интервалы; изучать взаимосвязи двух и более переменных; изучать монотонные взаимосвязи между метрическими и ранговыми переменными; применять методы анализа средних; осуществлять процедуру кластерного анализа;

Владеть: навыками работы со специализированным программным обеспечением (MS Excel, R).

4. Формат обучения – семинарские и практические занятия.

5. Объем дисциплины (модуля) составляет **3** з.е. и **108** часов, **58** академических часа, отведенных на контактную работу обучающихся с преподавателем (**26** часов – практические занятия, **26** часов – занятия семинарского типа), **56** академических часов на самостоятельную работу обучающихся из них **6** часов – мероприятия текущего контроля успеваемости и промежуточной аттестации. Форма промежуточной аттестации – экзамен.

6. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий

Краткое содержание дисциплины (аннотация):

Курс «Банки и базы данных в палеобиологии» включает в себя ознакомление с принципами построения баз данных, возможностями их применения в палеобиологии, математическими основами современных методов статистики и процедурами проверки статистических гипотез.

Наименование и краткое содержание разделов и тем дисциплины (модуля), **Форма промежуточной аттестации по дисциплине (модулю)**	**Все го** **(часы)**	В том числе				
		Контактная работа (работа во взаимодействии с преподавателем) **Виды контактной работы, часы**				**Самостоятельная работа обучающегося, часы** *(виды самостоятельной работы – дискуссии, контрольные работы, устные опросы)*
		Занятия лекционного типа	Практические занятия	Занятия семинарского типа	**Всего**	
Раздел 1. Введение.			2	2	4	Подготовка к устному опросу и дискуссии, 2 часа
Раздел 2. Описательные статистики.			2	2	4	Подготовка к контрольной работе, устному опросу и дискуссии, 6 часов
Раздел 3. Визуальное представление массивов данных.			4	3	7	Подготовка к устному опросу и дискуссии, 4 часа
Раздел 4. Репрезентативность.			2	3	5	Подготовка к устному опросу и дискуссии, 4 часа
Раздел 5. Создание баз данных в различном ПО			4	3	7	Подготовка к контрольной работе и дискуссии, 4 часа
Раздел 6. Двумерный анализ данных.			2	3	5	Подготовка к устному опросу и дискуссии, 4 часа
Раздел 7. Меры связи для номинальных и порядковых переменных.			2	2	4	Подготовка к устному опросу и дискуссии, 6 часов

Раздел 8. Методики анализа средних.			2	2	4	Подготовка к устному опросу и дискуссии, 6 часов
Раздел 9. Исследования монотонных взаимосвязей			2	2	4	Подготовка к устному опросу и дискуссии, 6 часов
Раздел 10. Кластерный анализ.			2	2	4	Подготовка к устному опросу и дискуссии, 4 часа
Раздел 11. Регрессионный анализ.			2	2	4	Подготовка к устному опросу и дискуссии, 4 часа
Промежуточная аттестация *экзамен*						6
Итого	**108**	**52**				**56**

Содержание разделов дисциплины:

Раздел 1. Введение.

Понятие о базах данных (БД) и системах управления БД, их функциональном назначение, анализ данных (АД), цели АД. Статистические закономерности и динамические законы. Создание формальной модели изучаемого объекта с помощью процедуры шкалирования. Типы шкал (номинальные, дихотомические, порядковые, интервальные, идеальных отношений). Преобразование шкал.

Раздел 2. Описательные статистики.

Абсолютные и относительные частоты. Максимум, минимум, размах. Меры средней тенденции: среднее, мода, медиана, среднее геометрическое, пятипроцентное усеченное среднее. Квантили: квартили, децили, процентили, межквартильный размах, интердецильный размах. Меры разброса: дисперсия (генеральная и выборочная), стандартное отклонение, коэффициент вариации. Ассиметрия (SKEWNESS), эксцесс (KURTOSIS). Описательные статистики в MS Excel: формулы (математические, статистические), дополнительные загружаемые модули.

Раздел 3. Визуальное представление массивов данных.

Плотностное распределение для метрических и неметрических шкал (гистограммы и одномерные частотные распределения). Коробчатые диаграммы. Массовые вымирания и кривые таксономического разнообразия Дж. Сепкоски и Д. Раупа. Расчет абсолютных и относительных показателей скорости видообразования. Использование инструментария PaleoBiology Database для построения кривых биоразнообразия и палеокарт с местами находок таксонов разного уровня.

Раздел 4. Репрезентативность.

Генеральная и выборочная совокупность. Перенос характеристик выборочной совокупности на генеральную. Случайная и систематическая ошибки. Точечная и интервальная оценки. Центральная предельная теорема. Объем выборки. Расчет доверительных интервалов для вероятностей и среднего.

Раздел 5. Создание баз данных в различном ПО.

Правила создания и кодирования переменных в различном программном обеспечении. Ограничения, налагаемые уровнем используемой шкалы. Перекодирование переменных. Исключение данных из рассмотрения (системные пропущенные значения). Расщепление базы данных по переменной.

Раздел 6. Двумерный анализ данных.

Проверка статистических гипотез об отсутствие взаимосвязи. Уровень значимости. Таблицы сопряженности. Ожидаемые и наблюдаемые частоты. Маргинальные значения. Критерий хи-квадрат, ограничения использования критерия хи-квадрат. Нормировки значений хи-квадрат. Остатки, Z-статистики.

Раздел 7. Меры связи для номинальных и порядковых переменных.

Сила связи. Направленная и ненаправленная связь. Положительная и отрицательная связь. Меры связи.

Раздел 8. Методики анализа средних**.**

Проверка гипотезы о равенстве средних. Параметрические тесты. Одновыборочный Т-тест, двухвыборочный Т-тест, двухвыборочный Т-тест для связанных выборок. Одновыборочный дисперсионный анализ (ANOVA). Апостериорные критерии множественных сравнений.

Раздел 9. Исследования монотонных взаимосвязей**.**

Монотонные взаимосвязи. Функциональная и корреляционная взаимосвязь. Проверка гипотезы об отсутствии монотонной взаимосвязи. Ранговая корреляция Спирмена. Коэффициент корреляции Пирсона.

Раздел 10. Кластерный анализ.

Кластер и кластеризация. Структура совокупности. Иерархический кластерный анализ. Стандартизация переменных, выбор способа объединения в кластеры и меры расстояния между переменными. Быстрый кластерный анализ. Метод k-средних. Кластерный анализ в биологии, ирисы Фишера.

Раздел 11. Регрессионный анализ.

Линейный регрессионный анализ. Зависимая и независимые переменные. Коэффициент детерминации. Ограничения модели линейного регрессионного анализа. Применение регрессионного анализа для изучения изменчивости организмов.

Содержание практических занятий.

1. Шкалирование
2. Описательные статистики
3. Массовые вымирания и кривые таксономического разнообразия
4. Инструментарий PaleoBiology Database и TimeScale Creator
5. Выборка и доверительный интервал
6. Процедура создания баз данных в различном ПО
7. Одномерный анализ данных и перекод переменных
8. Двумерный анализ данных
9. Меры связи для номинальных и порядковых переменных
10. Методики анализа средних
11. Корреляция
12. Кластерный анализ
13. Регрессионный анализ

Рекомендуемые образовательные технологии

При освоении дисциплины «Банки и базы данных в палеобиологии» предусматривается широкое использование активных и интерактивных форм проведения занятий.

Образовательные технологии**.** Аудиторные занятия проводятся в специализированной аудитории кафедры палеонтологии Геологического факультета МГУ, оборудованной компьютерами с доступом в Интернет и с использованием специального программного обеспечения.

7. Фонд оценочных средств (ФОС) для оценивания результатов обучения по дисциплине (модулю)

7.1. Типовые контрольные задания или иные материалы для проведения текущего контроля успеваемости.

Для текущего контроля студентов используются такие формы, как дискуссия, контрольная работа и устный опрос.

Примерный перечень вопросов для проведения устных опросов и контрольных работ:

1. Создание простейших формальных моделей случайно выбранных объектов через шкалирование.
2. Решение практических заданий для понимания работы математического аппарата различных методов статистического анализа данных.
3. Расчеты дисперсии, доверительных интервалов, ожидаемых частот, стандартизированных остатков, мер связи, проверка статистических гипотез об отсутствие взаимосвязи между переменными и т.д.
4. Составление таблиц, графиков, диаграмм, частотных и плотностных распределений и сдача расчетно-графических работ.
5. Применение БД для изучения адаптивной радиации крупных таксонов.
6. Построение кривых биоразнообразия и кривых скорости вымирания у конкретных отрядов морских беспозвоночных по базам Дж. Сепкоски и Д. Раупа.

Примерный перечень тем дискуссий:

1. Границы применимости таких статистических закономерностей как закон Гомперца;
2. Нормальное (Гаусса) распределение в биологии и палеонтологии; разнообразие, норма реакции, отбор;
3. Случайная и систематическая ошибка отбора при формирование палеонтологических выборок;
4. Скорость вымирания и видообразования, фоновый уровень вымирания;
5. Проверка статистических гипотез. Возможно ли подтвердить статистическую гипотезу?
6. Эвристический характер корреляции и взаимная обусловленность различных характеристик объекта

7.2. Типовые контрольные задания или иные материалы для проведения промежуточной аттестации.

Примерный перечень вопросов при промеж ут очной ат т ест ации:

1. Содержание понятий: база данных, система управления БД, анализ данных. Назначение баз данных цели анализа данных. Статистическая закономерность и динамический закон.
2. Шкалирование. Типы шкал (номинальная, порядковая, интервальная, идеальных отношений). Количественные и качественные шкалы. Преобразование шкал.
3. Описательные статистики. Меры средней тенденции. Меры разброса. Квантили и их виды. Асимметрия и пикообразность распределения.
4. Визуальное представление данных. Абсолютные и относительные показатели. Плотностные распределения для количественных и качественных шкал.
5. Использование инструментария PaleoBiology Database для иллюстрации динамики развития таксона и его распространения на момент времени.
6. Генеральная и выборочная совокупность. Перенос характеристик выборочной совокупности на генеральную. Доверительные интервалы.
7. Понятие двумерного анализа данных. Наблюдаемые и ожидаемые частоты. Логика проверки гипотезы об отсутствие взаимосвязи между переменными.
8. Понятие меры связи. Сила связи между переменными. Направленная и ненаправленная связь. Положительная и отрицательная связь.
9. Проверка гипотезы о равенстве средних. Назначение одновыборочного, двухвыборочного и двухвыборочного Т-теста для связанных выборок. Сравнение средних во множестве групп наблюдения (одновыборочный дисперсионный анализ и апостериорные критерии множественных сравнений).

10. Функциональная и корреляционная взаимосвязь. Корреляция между метрическими переменными. Ранговая корреляция.
11. Кластер и кластеризация. Структура совокупности. Необходимость стандартизации переменных.
12. Регрессионный анализ. Зависимая и независимые переменные. Коэффициент детерминации.

Шкала и критерии оценивания результатов обучения по дисциплине.

Результаты обучения	«Неудовлетворительно»	«Удовлетворительно»	«Хорошо»	«Отлично»
Знания: принципы построения баз данных; основные типы шкал; понятие репрезентативности, генеральной и выборочной совокупности, точечной и интервальной оценки; основные описательные статистики; основные методики анализа данных.	Знания отсутствуют	Фрагментарные знания	Общие, но не структурированные знания	Систематические знания
Умения: составлять собственные базы данных, осуществлять поиск необходимой информации; использовать методы статистического анализа данных	Умения отсутствуют	В целом успешные, но не систематические умения, допускаются неточности непринципиального характера	В целом успешное, но содержащее отдельные пробелы умение поиска информации, составления баз данных, использования методов статистического анализа данных.	Успешное умение составления баз данных, поиска информации, использования методов статистического анализа данных
Владения: навыками работы со специализированным программным обеспечением (MS Excel, R)	Навыки работы со специализированным программным обеспечением отсутствуют	Фрагментарное владение навыками работы в указанных программах	В целом сформированные навыки работы в MS Excel, R; затруднение с одним из заданий.	Полное владение специализированными программами (MS Excel, R)

8. Ресурсное обеспечение:

А) Перечень основной и дополнительной литературы.

— основная литература:

1. Берк К., Кейри П. Анализ данных с помощью Microsoft Excel/ Пер. с англ. М.: Издательский дом «Вильямс», 2005. 560 с.
2. Гнеденко Б.В. Курс теории вероятностей. М.: Наука, 1988. 446 с.
3. Лагутин М.Б. Наглядная математическая статистика. М.: БИНОМ. Лаб. знаний, 2007. 472 с.

— дополнительная литература:

1. Венцель Е.С. Теория вероятностей. – М.: Высшая школа, 1999. 576 с.
2. Миркин Б.Г. Введение в анализ данных: учебник и практикум. М.: Юрайт, 2015.
3. Райгородский А.М. Комбинаторика и теория вероятностей. МФТИ, 2012 или Интеллект, 2013

4. Diez, Barr, Cetinkaya-Rundel, Dorazio. Advanced High School Statistics. OpenIntro, Inc., 2015. 458 p.
5. Mirkin B. Core Concepts in Data Analysis: Summarization, Correlation, Visualization, Springer-London, 2012.

Б) Перечень лицензионного программного обеспечения: пакеты программ MS Excel, MS R.

В) Перечень профессиональных баз данных и информационных справочных систем

Г) программное обеспечение и Интернет-ресурсы:

1. Онлайн-курс по введению в статистику (http://onlinestatbook.com).
2. Портал по статистике и анализу данных (http://statistica.ru).
3. Электронные базы данных по палеозоологии на сайтах: paleobiodb.org, fossilworks.org, helsinki.fi/science/now/, gni.globalnames.org, organismnames.com, uio.mbl.edu, lib.vsu.ru.
4. Базы научной литературы: sci-hub.tw, libgen.io, evolbiol.ru, paleo.ru, jurassic.ru.

Д) Материально-технического обеспечение: — мультимедийный проектор, персональные компьютеры, экран, выход в Интернет.

9. Язык преподавания – русский.

10. Преподаватель (преподаватели) – Барсков И.С., Беляев Р.И.

11. Автор (авторы) программы – Барсков И.С., Беляев Р.И.