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

УТВЕРЖДАЮ
и.о. декана Геологического факультета
чл.-корр. РАН _____/Н.Н.Ерёмин/
«___» _____20__ г.

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ**

# Банки и базы данных в палеобиологии

Авторы-составители: Ростовцева Ю.И., Беляев Р.И., Орлова О.А.

**Уровень высшего образования:**

**Магистратура ИМ**

**Направление подготовки:**

**05.04.01 Геология**

**Направленность (профиль) ОПОП:**

**Геология и полезные ископаемые**

**Магистерская программа**

**Палеонтология и стратиграфия**

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 магистратуры, реализуемые последовательно по схеме интегрированной подготовки).*

Год (годы) приема на обучение: 2022

**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:**

**—** ознакомление с современными палеобиологическими базами данных;

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

**—** 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).

**Краткое содержание дисциплины (аннотация):**

Учебный 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.

**1. Место дисциплины (модуля) в структуре ОПОП –** относится к вариативной части ОПОП, является обязательной для освоения.

**2. Входные требования для освоения дисциплины, предварительные условия:**

Знания в части общекультурной и общенаучной подготовки – на уровне требований Образовательного стандарта МГУ направление «Геология», уровень бакалавриат, знания в области геологии в соответствии с требованиями вступительного экзамена в магистратуру.

***3. Планируемые результаты обучения по дисциплине (модулю), соотнесенные с требуемыми компетенциями выпускников.***

| **Компетенции выпускников (коды)** | **Индикаторы (показатели) достижения компетенций** | **Планируемые результаты обучения по дисциплине (модулю), сопряженные с компетенциями** |
|---|---|---|
| **ПК-2М.** Способен самостоятельно проводить научные исследования с помощью современного оборудования, информационных технологий, с использованием новейшего отечественного и зарубежного опыта; | **М.ПК-2. И-1.** Критически анализирует новейший отечественный и зарубежный опыт научно-исследовательских работ по тематике собственного исследования.<br>**М.ПК-2. И-3.** Обрабатывает полученные результаты, формулирует выводы и рекомендации по использованию полученных результатов. | ***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; |

| | | |
|---|---|---|
| **ОПК-6М.** Способен использовать современные вычислительные методы и компьютерные технологии для решения задач профессиональной деятельности. | **М.ОПК-6. И-1.** Выбирает способы обработки данных и программные средства для решения задач профессиональной деятельности с учетом основных требований информационной безопасности. | ***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;<br>***Владеть:*** навыками работы со специализированным программным обеспечением (MS Excel, R). |

**4. Объем дисциплины (модуля)** составляет 3 з.е., в том числе 56 академических часов на контактную работу обучающихся с преподавателем (14 часов - лекции, 14 часов - семинары и 28 часов - практические занятия), 52 академических часа на самостоятельную работу обучающихся. Форма промежуточной аттестации – экзамен.

**5. Формат обучения** не предполагает электронного обучения и использования дистанционных образовательных технологий (за исключением форс-мажорных обстоятельств – пандемии и т.п.)

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

| Наименование и краткое содержание разделов и тем дисциплины (модуля),<br><br>Форма промежуточной аттестации по дисциплине (модулю) | **Всего (часы)** | В том числе | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Контактная работа (работа во взаимодействии с преподавателем)** ***Виды контактной работы, часы*** | | | | **Самостоятельная работа обучающегося** ***Виды самостоятельной работы, часы*** | | |
| | | Лекции | Практические занятия | Семинары | **Всего** | Устный опрос | Рефераты | **Всего** |
| Раздел 1. Введение | 2 | 1 | | 1 | 2 | | | |
| Раздел 2. Описательные статистики | 15 | 2 | 4 | 1 | 7 | 2 | 6 | 8 |
| Раздел 3. Визуальное представление массивов данных | 16 | 3 | 4 | 2 | 9 | 2 | 5 | 7 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Раздел 4. Репрезентативность и создание баз данных в различном ПО | 19 | 2 | 6 | 3 | 11 | 2 | 6 | 8 |
| Раздел 5. Двумерный анализ данных. Меры связи для номинальных и порядковых переменных | 20 | 2 | 6 | 3 | 11 | 2 | 7 | 9 |
| Раздел 6. Методики анализа средних и исследования монотонных взаимосвязей | 17 | 2 | 4 | 2 | 8 | 2 | 7 | 9 |
| Раздел 7. Кластерный и регрессивный анализы | 17 | 2 | 4 | 2 | 8 | 2 | 7 | 9 |
| Промежуточная аттестация *экзамен* | 2 | *Экзамен* | | | | 2 | | |
| **Итого** | **108** | **56** | | | | **52** | | |

**Содержание разделов дисциплины**

**Содержание лекций:**

**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 и создание баз данных в различном ПО.** 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. 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 5. Двумерный анализ данных. Меры связи для номинальных и порядковых переменных.** 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. Меры связи для номинальных и порядковых переменных. Сила связи. Направленная и ненаправленная связь. Положительная и отрицательная связь. Меры связи.

**Раздел 6. Методики анализа средних и исследования монотонных взаимосвязей.** 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. Исследования монотонных взаимосвязей. Монотонные взаимосвязи. Функциональная и корреляционная взаимосвязь. Проверка гипотезы об отсутствии монотонной взаимосвязи. Ранговая корреляция Спирмена. Коэффициент корреляции Пирсона.

**Раздел 7. Кластерный и регрессивный анализы.** 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. 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
5. Инструментарий TimeScale Creator
6. Выборка и доверительный интервал
7. Процедура создания баз данных в различном ПО
8. Одномерный анализ данных и перекод переменных
9. Двумерный анализ данных
10. Меры связи для номинальных и порядковых переменных
11. Методики анализа средних
12. Корреляция
13. Кластерный анализ
14. Регрессионный анализ

**Содержание семинаров:**

1. Создание формальной модели изучаемого объекта с помощью процедуры шкалирования.
2. Описательные статистики в MS Excel: формулы (математические, статистические), дополнительные загружаемые модули.
3. Расчет абсолютных и относительных показателей скорости видообразования.
4. Правила создания и кодирования переменных в различном программном обеспечении
5. Проверка статистических гипотез об отсутствие взаимосвязи
6. Проверка гипотезы о равенстве средних. Параметрические тесты
7. Применение регрессионного анализа для изучения изменчивости организмов.

**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 успеваемости студентов используются такие формы, как устные опросы, написание реферата для оценки степени усвоения материала по разделам курса. По итогам обучения в первом семестре в зачетную сессию проводится устный экзамен.

***Примерный перечень вопросов для проведения устных опросов:***

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,
4. Проверка статистических гипотез об отсутствие взаимосвязи между переменными и т.д.
5. Составление таблиц, графиков, диаграмм, частотных и плотностных распределений и сдача расчетно-графических работ.
6. Перекодирование переменных.
7. Применение БД для изучения адаптивной радиации крупных таксонов.
8. Кластерный анализ в биологии, ирисы Фишера
9. Построение кривых биоразнообразия у отрядов морских беспозвоночных по базам Дж. Сепкоски и Д. Раупа
10. Построение кривых скорости вымирания у конкретных отрядов морских беспозвоночных по базам Дж. Сепкоски и Д. Раупа.

***Примерный перечень тем рефератов***

1. Границы применимости таких статистических закономерностей как закон Гомперца.
2. Методы проверки нормальности распределения (визуальный метод, критерий Комогорова-Смирнова).
3. Нормальное (Гаусса) распределение в биологии и палеонтологии; разнообразие, норма реакции, отбор.
4. Случайная и систематическая ошибка отбора при формирование палеонтологических выборок;
5. Роль и место банков данных в информационных системах.
6. Создание реляционной базы данных
7. Массовые вымирания и кривые таксономического разнообразия Дж. Сепкоски и Д. Раупа
8. Скорость вымирания и видообразования, фоновый уровень вымирания.
9. Проверка статистических гипотез. Возможно ли подтвердить статистическую гипотезу?
10. 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 промежуточной аттестации:***

1. Содержание понятий: база данных, система управления БД, анализ данных. Назначение баз данных цели анализа данных.
2. Статистическая закономерность и динамический закон.
3. Шкалирование. Типы шкал (номинальная, порядковая, интервальная, идеальных отношений).

4. Количественные и качественные шкалы. Преобразование шкал.
5. Описательные статистики. Меры средней тенденции. Меры разброса.
6. Описательные статистики. Квантили и их виды. Асимметрия и пикообразность распределения.
7. Визуальное представление данных. Абсолютные и относительные показатели.
8. Визуальное представление данных. Плотностные распределения для количественных и качественных шкал.
9. Использование инструментария PaleoBiology Database для иллюстрации динамики развития таксона и его распространения на момент времени.
10. Генеральная и выборочная совокупность.
11. Перенос характеристик выборочной совокупности на генеральную. Доверительные интервалы.
12. Понятие двумерного анализа данных. Наблюдаемые и ожидаемые частоты. Логика проверки гипотезы об отсутствие взаимосвязи между переменными.
13. Понятие меры связи. Сила связи между переменными.
14. Направленная и ненаправленная связь. Положительная и отрицательная связь.
15. Проверка гипотезы о равенстве средних. Назначение одновыборочного, двухвыборочного и двухвыборочного Т-теста для связанных выборок.
16. Сравнение средних во множестве групп наблюдения (одновыборочный дисперсионный анализ и апостериорные критерии множественных сравнений).
17. Функциональная и корреляционная взаимосвязь.
18. Корреляция между метрическими переменными. Ранговая корреляция.
19. Кластер и кластеризация. Структура совокупности. Необходимость стандартизации переменных.
20. Регрессионный анализ. Зависимая и независимые переменные. Коэффициент детерминации.

Шкала и критерии оценивания результатов обучения по дисциплине

| Результаты обучения, *соответствующие виды оценочных средств* | «Неудовлетворительно» | «Удовлетворительно» | «Хорошо» | «Отлично» |
|---|---|---|---|---|
| **Знания:** принципов построения баз данных; основных типов шкал; понятий репрезентативности, генеральной и выборочной совокупности, точечной и интервальной оценки; основных описательных статистик; процедур проверки статистических гипотиз и основных методик анализа данных *(устный опрос)* | Знания отсутствуют | Фрагментарные знания | Общие, но не структурированные знания | Систематические знания |
| **Умения** составлять собственные базы данных, осуществлять 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 | 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 |

| | | | | |
|---|---|---|---|---|
| более переменных; изучать монотонные взаимосвязи между метрическими и ранговыми переменными; применять методы анализа средних; осуществлять процедуру кластерного анализа *(устный опрос)* | | | | |
| **Навыки** работы со специализированным программным обеспечением (MS Excel, R) *(устный опрос)* | 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:**

**- лицензионное**

не требуются

**- нелицензионное и свободного доступа**

пакет программ Open Office

**В) Перечень профессиональных баз данных и информационных справочных систем**

-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.

- Базы научной литературы: sci-hub.tw, libgen.io, evolbiol.ru, paleo.ru, jurassic.ru.

**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).

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

**9. Язык преподавания –** русский.

**10. Преподаватель (преподаватели) –** Ответственный за курс — доцент кафедры палеонтологии Ростовцева Ю.И., преподаватели — Ростовцева Ю.И., Крутых А.В., Мамонтов Д.А., Кожанова Д.А.

**11. Разработчики программы: –** доцент Ростовцева Ю.И., младший научный сотрудник ИПЭЭ РАН Беляев Р.И., доцент Орлова О.А.