Статистика и анализ данных

Оценочные средства текущего контроля:

Устный опрос:

1. Как создать переменную в R и присвоить ей значение?

2. Какие существуют правила при выборе имени переменной?

3. Какие логические операции вы знаете?

4. Что такое вектор в R и как его создать?

5. Правила сложения векторов.

6. Типы переменных в R

7. Правила и возможные способы преобразования типов

8. Программа Rstudio, её окна и основные вкладки

9. Что такое скрипт на языке R? Как его создать и запустить?

10. Как установить рабочую директорию в программе Rstudio?

11. Что такое data mining, или многомерный анализ данных?

12. На какие группы методов можно разделить многомерный анализ данных?

13. Какая функция R предназначена для попарной визуализации количественных переменных в данных? Приведите пример её использования.

14. Теоретические основы и назначение метода PCA

15. Что такое график biplot и как его можно интерпретировать?

16. Какие методы кластеризации данных вы знаете?

17. В чём сходства и различия кластеризации и классификации данных?

18. Особенности кластеризации методом k средних

19. Описать процедуру иерархической кластеризации снизу-вверх

20. Основные способы вычисления расстояний в кластеризации

Компьютерная программа:

1. Исследовать влияние витамина С и апельсинового сока в дозировке 0.5 мг на рост зубов морских свинок (встроенный в R набор данных ToothGrowth).

2. Используя встроенный в R набор данных chickwts, построить коробчатый график для веса цыплят и определить, какому типу корма соответствует цыплёнок с минимальным и максимальным весом. Охарактеризовать распределение весов цыплят для типа корма, которому принадлежит цыплёнок с максимальным весом.

3. Используя встроенный в R набор данных CO2, содержащий информацию о темпах поглощения двуокиси углерода растением Echinochloa crusgalli (ежовник обыкновенный), произрастающем в Квебеке и Миссисипи, в зависимости от концентрации CO2 в окружающем воздухе и того факта, было ли растение охлаждено накануне проведения эксперимента или нет, составить таблицу описательных статистик.

4. Построить коробчатый график для зависимости темпов поглощения двуокиси углерода от фактора происхождения растения (Type) и того факта, было ли оно предварительно охлаждено (Treatment). Охарактеризовать зависимость темпов поглощения CO2 от этих двух факторов и определить, какой из них, предположительно, определяет большую изменчивость признака.

5. Построить коробчатый график для зависимости темпов поглощения двуокиси углерода от его концентрации (conc) и охарактеризовать данную зависимость.

6. Найти растения с минимальным и максимальным средним темпом поглощения двуокиси углерода.

7. Используя команду source ("http://www.openintro.org/stat/data/present.R"), сохранить фрейм данных под названием present с сайта OpenIntro. Этот массив данных содержит количества рождений мальчиков и девочек с 1940 по 2002 год в США. Рассчитайте абсолютные различия между количеством мальчиков и девочек, родившихся в каждом году, и определите, в каком году была самая большая абсолютная разница в количествах новорожденных девочек и мальчиков?

8. На одном графике отразите динамику рождений мальчиков и девочек. На основе графика определите, справедливо ли утверждение, что каждый год рождалось больше девочек, чем мальчиков.

9. Постройте график доли мальчиков с течением времени. На основе графика определите, справедливо ли утверждение, что доля мальчиков, родившихся в США, уменьшилась с течением времени.

10. Построить диаграмму рассеяния для зависимости расхода топлива автомобилей из встроенного набора данных mtcars от количества лошадиных сил.

11. Рассчитайте коэффициенты корреляции Пирсона и Спирмена между количеством лошадиных сил и расходом топлива (набор данных mtcars).

12. Визуализируйте все возможные парные диаграммы рассеяния между количественными параметрами набора данных mtcars (на одном графике). Между какими параметрами есть очевидная зависимость?

13. Составьте уравнение линейной регрессии, описывающее зависимость расхода топлива от количества лошадиных сил. Добавьте на диаграмму рассеяния красную линию, соответствующую уравнению регрессии.

14. Используя составленное уравнение регрессии, предскажите расход топлива для новых наблюдений в таблице (100, 150, 129, 300 лошадиных сил).

15. Составьте уравнение линейной регрессии для зависимости расхода топлива от факторной переменной – числа цилиндров.

16. Построить матрицу диаграмм рассеяния между всеми парами измерений из набора встроенных в R данных iris.

17. Используя команду source (http://www.openintro.org/stat/data/cdc.R), загрузите в рабочее пространство базу данных cdc, в которой представлены переменные: genhlth, exerany, hlthplan, smoke100, height, weight, wtdesire, age, и gender. Добавьте новую переменную индекс массы тела (BMI) во фрейме данных cdc. Как зависит BMI от состояния здоровья (genhlth) респондентов? Отобразите результат на графике.

18. Загрузите базу данных nc в рабочее пространство RStudio с http://d396qusza40orc.cloudfront.net/statistics/lab_resources/nc.RData. Связан ли вес ребенка (weight) с весом, набранным матерью во время беременности (gained)? Постройте график зависимости и охарактеризуйте связь между этими переменными.

19. Связан ли вес ребенка (weight) с возрастом матерей? Постройте график зависимости и охарактеризуйте связь между этими переменными.

20. Дана таблица сопряжённости курения и наличия артериальной гипертонии. Исследовать взаимосвязь между этими двумя факторами.

21. Даны наблюдаемые частоты расщепления по фенотипу в F2 дигибридного скрещивания. Проверить, соответствует ли оно ожидаемому.

22. Дана таблица сопряжённости между типом анестезии (галотан/морфин) и операционной летальностью. Исследовать достоверность этой взаимосвязи.

23. Дана таблица сопряжённости между неким фактором 1 с тремя градациями и фактором 2 с двумя. Проверить нулевую гипотезу об отсутствии взаимосвязи этих факторов с помощью критерия хи-квадрат.

24. Рассчитать точную вероятность случайно получить заданную или ещё более несбалансированную таблицу сопряжённости между неким фактором 1 и фактором 2 (оба с двумя градациями).

25. В лаборатории работает 5 юношей и 8 девушек. Можно ли сказать, что такое распределение статистически значимо отличается от равномерного?

26. Изучалась эффективность высокочастотной стимуляции нерва в качестве обезболивающего средства при удалении зуба. Все больные подключались к прибору, но в одних случаях он работал, в других был выключен. Ни стоматолог, ни больной не знали, включён ли прибор. Дана таблица сопряжённости. Позволяют ли эти данные считать высокочастотную стимуляцию нерва действенным анальгезирующим средством?

27. Составить уравнение логистической регрессии для данных о студентах (дана таблица сопряжённости сдачи/несдачи экзамена и наличия и отсутствия подготовки).

28. Мендель выращивал горох трёх цветов. В одном из опытов эмпирическое распределение частот некоторого цвета гороха приняло следующий вид: 18, 55, 27. Предполагаемое теоретическое распределение имеет следующий вид: 1:2:1. Рассчитайте расстояние хи-квадрат для этого примера.

29. Одна из причин инсульта - окклюзия сонной артерии. Чтобы выяснить, какое лечение - медикаментозное или хирургическое - даёт в этом случае лучшие результаты, исследователи сравнили долгосрочный прогноз у

пациентов, на которых применялись два этих метода. Дана таблица сопряжённости. Можно ли говорить о превосходстве одного из видов лечения?