Практическая работа №11
Анализ категорийных данных в R
Проверка статистических гипотез

Цель: сопоставлять эмпирическое распределение признака с одним из теоретических законов распределения, сравнивать два эмпирических распределения между собой с целью выявления сходства или различия в форме распределения, научиться строить таблицы сопряженности, проводить анализ категорийных данных с помощью столбчатых диаграмм, изучать связь между двумя и более переменными, используя критерии Пирсона χ^2, Фишера для проверки гипотез о зависимости переменных, проверять данные на однородность (критерий Бартлетта) и нормальность (критерий Шапиро-Уилка) распределения, проверять различия в средних с использованием t-критерия Стьюдента и критерия Манна-Уитни.

Задания

1. Загрузить в новую переменную **sal** набор данных **Salaries.csv** (данные о зарплатах профессорского состава университета).
2. Посмотреть структуру данных переменной **sal**.
3. Выбрать одну из переменных, являющихся фактором. Построить сводную таблицу для выбранной переменной. Сохранить сводную таблицу в переменную **tab**.
4. Определить размерность переменной **tab.** Сделать выводы (записать их в отчет).
5. Найти сколько женщин и мужчин (**sex**) преподает разные дисциплины (**discipline**), сохранить результат в переменную **tab1**.
6. Указать в названиях где какая переменная.
7. Определить размерность таблицы **tab1**. Сделать выводы (записать их в отчет).
8. Создать таблицу (**tab2**) по трём переменным: пол (**sex**), дисциплина (**discipline)** и ранг (**rank**).
9. В таблице **tab2** назвать переменные: Пол, Дисциплина, Ранг.
10. Определить размерность таблицы **tab2**. Сделать выводы (записать их в отчет).
11. Посмотреть измерения таблицы и их названия.
12. Вывести данные: какие дисциплины ведут женщины разного ранга.
13. С помощью функции **margin.table()** посчитать частоты по всем трем переменным таблицы **tab2**.
14. Вывести подробную информацию из сводных таблиц **tab, tab1** в виде пропорций.
15. Из таблицы **tab2** найти пропорции по переменной ранг.
16. Добавить к таблице **tab2** сумму по столбцам и строкам с помощью функции **addmargins()**.
17. Вычислить суммы таблицы **tab1** и по строкам, и по столбцам

одновременно.

18. Добавить к таблице **tab1** сумму по строкам.
19. Добавить к таблице **tab1** сумму по столбцам.
20. Добавить к таблице **tab1** сумму и по столбцам, и по строкам, в пропорциях.
21. Вычислить сумму таблицы **tab1** в процентах только по строкам.
22. Вычислить сумму таблицы **tab1** в процентах только по столбцам.
23. Создать таблицу сопряженности из набора данных **sal**, используя функцию **CrossTable()** из пакета **gmodels**. Взять для таблицы переменные Пол (**sex**) и Дисциплина(**discipline**).
24. Используя функцию **xtabs** создать таблицу (**tab3**) по трём переменным: пол (**sex**), дисциплина (**discipline)** и ранг (**rank**).
25. Напечатать таблицу **tab3**.
26. Сравнить таблицы **tab2** и **tab3**.
27. Построить столбчатую диаграмму для переменной **tab**. Столбик, показывающий количество женщин закрасить розовым цветом, а мужчин – голубым. Подписать оси графика.
28. Построить столбчатую диаграмму для таблицы **tab1**. Добавить на график легенду в центре наверху. Подписать названия всех значений переменных (Дисциплина A, Дисциплина B), оси графика. Цвета распределить также: женщины – розовый, мужчины – голубой.
29. Проверить зависимость переменных, которые составляют строки и столбцы, двухмерной таблицы **tab1,** применяя тест хи-квадрат (функцию **chisq.test()**). На основе полученных результатов сделать выводы.
30. Проверить нулевую гипотезу о независимости столбцов и строк в таблице сопряженности **tab1,** используя тест Фишера.
31. С помощью критерий t-Стьюдента проверить зависит ли:
    31.1. заработная плата (**salary**) от пола преподавателя (**sex**)
    31.2. заработная плата (**salary**) от ранга преподавателя (**rank**)
    31.3. ранг преподавателя (**rank**) от количества лет, проработанных после получения степени phd (**yrs.service**).
32. Построить простую гистограмму переменной зарплата (**salary**).
33. Построить гистограмму переменной зарплата преподавателей (**salary**) по их рангу (**rank**) с помощью пакета **ggplot2**. Украсить график (определить цвет края, цвет заполнения столбиков и т.д.).
34. Построить диаграмму плотности распределения зарплаты по рангу преподавателя. Указать прозрачность равную 0.4.
35. Выяснить есть ли в данных выбросы с помощью диаграммы размаха.
36. Проверить зарплату на нормальность распределения в целом и по группам в зависимости от ранга.
37. Проверить зарплату на однородность с помощью критерия Бартлетта.
38. Провести сравнение по двум переменным: **зарплате** и **рангу** набора данных **sal1** для независимых выборок, используя t-критерий Стьюдента.

39. Вывести на экран только значение p-уровня значимости.
40. Проверить гипотезу о том, что количество лет, которые были отработаны после получения степени phd (**yrs.since.phd**) и количество лет работы в должности (**yrs.service**) не равны.
41. С помощью критерия χ^2 определить существует ли связь между количеством покупок одежды и семейным положением для женщин и мужчин. Данные для анализа указаны в таблицах 1 - 2.

Таблица 1. – Покупки женщин

Частота покупок	**Семейное положение**	
	Не замужем	**Замужем**
Много	32	25
Мало	21	76
Очень мало	6	8
Итоги	59	109

Таблица 2. – Покупки мужчин

Частота покупок	**Семейное положение**	
	Не женат	**Женат**
Много	13	33
Мало	20	71
Очень мало	6	6
Итоги	39	110

Для этого:

41.1. Объединить 2 таблицы
41.2. Задать в R три переменные:
   1) Покупка (с уровнями: 1 – много, 2 – мало, 3 – Очень мало),
   2) Семейное положение (1 – женат, 2 – неженат),
   3) Частота каждого сочетания.
41.3. Ввести данные согласно таблицам. Например, для женщин:

Частота покупок	**Семейное положение**	**Частота**
Много	Не замужем	32
Мало	Не замужем	21

Очень мало	Не замужем	6
Много	Замужем	25
Мало	Замужем	76
Очень мало	Замужем	8

41.4. С помощью команды **xtabs()** представить эти данные в виде кросс-таблицы.

41.5. Получить кросс-таблицу в стиле SPSS с помощью команды **CrossTable** из пакета **gmodels**.

41.6. Проанализировать полученные данные, о чем говорит критерий Пирсона хи-квадрат.

Контрольные вопросы:

1. Дайте понятие категорийных данных
2. Перечислите способы преобразования категорийных данных в фактор
3. Какие существуют типы сводных таблиц?
4. Какая команда используется для посмотра измерений в таблице?
5. Перечислите отличия сводных таблиц
6. С помощьюкакой функции можно получить из сводных таблиц полную информацию?
7. Что показывает биномиальный тест?
8. Перечислите ограничения применению критерия χ^2 Пирсона?
9. Для чего используется тест χ^2Пирсона?
10. Назовите критерий, предназначенный для проверки нормальности распределения?
11. Перечислите основные показатели Шапиро-Уилка теста?
12. С помощью какого критерия проверяется однородность данных?
13. Какой тест используют, когда возникает необходимость выполнить непараметрический аналог t-критерия Стьюдента?
14. Для чего применяют t-критерий Стьюдента?

Отчет по практической работе №11

1. Титульный лист: название работы, вариант, ФИО, учебная группа.
2. Оглавление.
3. Отчет оформлять на каждый пункт задания, указывая следующее:
   1.1. Задание;
   1.2. Решение с кодом;
   1.3. Результат.
   1.4. Вывод.

Домашнее задание

1. Изучить все возможности функции **CrossTable()**.
2. Определить существует ли зависимость между возрастом туриста и затратами на отдых.

Таблица 3. – Зависимость между возрастом туриста и затратами на отдых

Возраст туриста	Уровень затрат на отдых			Всего
	Низкий	Средний	Высокий	
Молодежь	32	41	26	
Средний возраст	48	64	53	
Пожилые люди	59	24	13	
Всего				

3. Проверить гипотезу: частота обращений в телефонную службу доверия неравномерно распределяется по дням недели (с помощью критерия χ^2 сравнить заданное эмпирическое распределение с равномерным).

Таблица 4. – Распределение обращений в телефонную службу доверия по дням недели

День недели	**Количество обращений**
Понедельник	10
Вторник	7
Среда	8
Четверг	7
Пятница	10
Суббота	16
Воскресенье	19