***Дисперсионный анализ.***

Подготовка:

-создать каталог C:\data и сохранить в него файл anova\_lab1.csv

((**пользователям Mac OS** в терминале перейти в домашнюю папку:

***cd ~alexx*** (для пользователя alexx)

и создать папку data:

***md data.***

Во всех командах вместо C:/data/file.csv использовать /Users/alexx/data/file.csv))

**ВСЕМ**: при необходимости установить в системной консоли модули: pandas, matplotlib, researchpy, scipy, statsmodels:

pip install pandas

Pip install matplotlib

pip install researchpy

pip install scipy

pip install statsmodels

**Задание 1.**

Для анализа качества работы операторов проведено небольшое исследование, в ходе которого зафиксировано время, затраченное на каждый телефонный звонок (в минутах) в течение одного рабочего дня. Перед тем как внести изменения в организацию работы необходимо разобраться в этой ситуации. Продолжительность звонков, сгруппированных по определенным темам, приведена в файле anova1.csv

1) ввести данные в Python:

-В интерпретаторе (консоли) Python:

import pandas as pd

import numpy as np

import researchpy as rp

import matplotlib.pyplot as plt

import matplotlib.image as mpimg

df = pd.read\_csv("C:/data/anova\_lab1.csv")

2) построить в одном масштабе блочные диаграммы для этих четырех видов звонков и описать полученную структуру.

boxplot = df.boxplot(by='call')

plt.show()

3) вычислить описательные статистики для каждого вида телефонных звонков

rp.summary\_cont(df['length'].groupby(df['call']))

import scipy.stats as stats

4) какой из типов звонков имеет наибольшую среднюю продолжительность? какой – наименьшую?

5) выполняется ли для данного набора данных необходимое для однофакторного дисперсионного анализа предположение о нормальном распределении и равной вариации? почему?

- проверить данные на нормальность распределения:

import scipy.stats as stats

stats.shapiro(df['length'])

6) скопируйте данные под новым именем и вычислите для каждого значения натуральный логарифм.

dfLog = df

dfLog[‘length’] = np.log(df[‘length’])

7) Постройте блочные диаграммы для этих логарифмов.

boxplot = dfLog.boxplot(by='call')

plt.show()

Можно ли сказать, что предположение о равной вариации лучше выполняется для логарифмов, чем для исходных данных?

**Задание 2.**

1) Используя значения логарифмов продолжительности телефонных звонков из задания 1, проведите однофакторный дисперсионный анализ:

import statsmodels.api as sm

from statsmodels.formula.api import ols

model = ols('length ~ call', data=dfLog).fit()

aov\_table = sm.stats.anova\_lm(model, typ=2)

aov\_table

2) опишите полученные результаты о различиях между этими четырьмя видами телефонных звонков;

3) Проведите тест наименьшего значимого различия (Tukey’s Honestly Significant Difference Test):

import statsmodels.stats.multicomp as mc

comp = mc.MultiComparison(dfLog['length'], dfLog['call'])

post\_hoc\_res = comp.tukeyhsd()

print(str(post\_hoc\_res.\_results\_table))

Какая пара типов телефонных звонков существенно отличается друг от друга (в терминах среднего значения логарифма продолжительности звонка)?

**Задание 3.**

В таблице файла anova\_lab2.csv приведены оценки рекламы, усредненные в балльную систему. При проведении эксперимента взяли по 4 наблюдения для каждой комбинации значений.

1) Ввести данные в Python:

import pandas as pd

import numpy as np

import researchpy as rp

import matplotlib.pyplot as plt

import matplotlib.image as mpimg

df = pd.read\_csv("C:/data/anova\_lab2.csv")

2) Проведите двухфакторный дисперсионный анализ .

import statsmodels.api as sm

from statsmodels.formula.api import ols

model = ols('effect ~ C(type) + C(parameter) + C(type):C(parameter)', data=df).fit()

sm.stats.anova\_lm(model, typ=2)

3) Среднее значение эффективности у рекламы 1 выше, чем у рекламы 2. Значимо ли оно выше? Как вы это определили?

4) Является ли взаимодействие факторов существенным? О чем это говорит?

5) Является ли воздействие главного эффекта вида рекламы существенным? О чем это говорит?

6) Является ли воздействие главного эффекта параметров рекламы существенным? О чем это говорит?

7) Постройте график средних значений для главных эффектов. Сделайте выводы.

from statsmodels.graphics.factorplots import interaction\_plot

fig=interaction\_plot(df.parameter, df.type, df.effect, colors=['red', 'blue'], markers=['D', '^'])

fig.show()

**Задание 4.**

В таблице файла Лаб\_Дисп3 приведены результаты исследования Kirby and Dardis, изучавшего на протяжении 13 недель цены 20 наименований товаров. В результате исследования выявлено, что цены в универсальных магазинах действительно на 40% выше цен в магазинах производителя.

1. Действительно ли высокие цены (в среднем на 40% выше в универмагах) значимо выше? Почему?
2. Какой вид дисперсионного анализа здесь использован?
3. Определите три фактора в этом анализе. Сколько категорий имеет каждый из факторов?
4. О чем свидетельствует p-значение для главного эффекта фактора В?
5. Значимо ли различаются цены разных недель? Как это узнать?
6. Рассмотрите взаимодействие типа магазина и вида товара. Значимо ли оно? О чем это говорит?
7. Рассмотрите взаимодействие типа магазина и недели? Значимо ли оно? О чем это говорит?
8. Рассмотрите взаимодействие вида товара и недели. Значимо ли оно? О чем это говорит?
9. Обычно мы изучаем p-значения , чтобы посмотреть, достаточно ли они малы, чтобы можно было заявить, что результат является статистически значимым. Однако для трехфакторного взаимодействия p-значение подозрительно велико, что наводит на мысль о том, что здесь значительно *меньше* случайности, чем можно было бы ожидать для этой модели. Какое техническое предположение для проверки гипотезы может быть здесь не выполнено?