Проект по СЕМ – практикум

Веселина Цацева, ИС, 3та група, ф.н. 71943

Анализ на данни за диабет

Данните се състоят от една категорийна променлива - Pregnancies и две числови – Age и BMI. Броя на изследваните пациенти е 523.

1. Едномерен анализ

* Pregnancies

Chart, bar chart

Description automatically generated Тази променлива показва броя бременности, които е имал пациента. Тя се разделя на четири категории – 0, 1, [2,3] и 4+.

Chart, pie chart

Description automatically generatedГорните графики показват честотното и процентното разпределение на данните в тази променлива. От тях се вижда, че най-голям е броят на пациентите с повече от 4 бременности, т.е. модата е 4+.

* Age

Представя възрастта на пациента. Тъй като променливата е числова правим Box plot графика. От нея виждаме, че минималната и максималната възраст на пациентите е съответно 21 и 81г., т.е. обхватът е 60. За локацията на разпределение използваме медиана, тъй като се вижда, че има наличия на outlier-и, чиито стойности са над 58. Тя е 27, което означава, че половината от изследваните пациенти са под 27 годишна възраст, а другата половина на или над тази възраст. Първият квартил е 23, т.е. ¼ от пациентите са под 23 години, а третият квартил е 37, което показва, че ¾ от пациентите са на по-малко от 37 год.

От хистограмата се вижда, че извадката съдържа екстремални стойности и има J - форма на разпределение. IQR е 14, т.е. средните 50% от пациентите са на между 23 и 37 години, или още - възрастта на половината от тях е разпределена между тези 14 стойности от всичките 60. MAD e 7.4 – отклонението на стойностите от медианата.

За теста на разпределението имаме нулева хипотеза, че разпределението е нормално. При Shapiro test получаваме p-value < 2.2e-16, което е по-малко от 0.05, следователно отхвърляме нулевата хипотеза – не е нормално разпределена.

Chart, histogram

Description automatically generatedChart, box and whisker chart

Description automatically generated

* BMI

Chart, box and whisker chart

Description automatically generated Представя индекса на телесната маса на пациент. Подобно на предната променлива и тук данните се представят чрез Box plot.

Минималната и максималната стойности са 18.2 и 67.1, т.е. обхватът е 48.9, приблизително 49. Медианата е 32.7, половината пациенти имат индекс на телесна маса под 32.7, а останалите над това число. Първият и третият квартил са съответно равни на 27.8 и 36.8, т.е. ¼ от пациентите имат BMI под 27.8, а ¾ имат BMI под 36.8. Извадката има outlier-и, чиито стойности са по-големи от 50.

Чрез five number summary виждаме, че медианата и средната стойност са равни. Също и модата е равна на 32. Следователно изглежда, че имаме нормално разпределение. Вижда се и от QQ plot-а.

IQR е 9, т.е. средните 50% от пациентите имат BMI между 27.8 и 36.8, или още - индекса на телесната маса на половината от тях е разпределена между тези 9 стойности от всичките 49. Стандартното отклонение и MAD са близки по стойност, съответно 7 и 6.7, но т.к. има outlier-и MAD е по-подходящо.

Нулевата хипотеза е, че разпределението е нормално. С Shapiro test получаваме p-value = 1.506e-07 < 0.05, следователно отхвърляме нулевата хипотеза.

Chart, line chart

Description automatically generatedChart, histogram

Description automatically generated

1. Многомерен анализ

* Pregnancies VS Age

Графичното представяне е следното:

Chart, box and whisker chart

Description automatically generated

В pregnancies имаме 4 категории, т.е. 4 групи. Проверяваме ги за вида на разпределение чрез функцията aggregate и Shapiro Wilk. Получаваме, че мин. стойност за p-value е < 0.05, т.е. отхвърляме нулевата хипотеза за нормално разпределение. Следователно използваме тест на Крускал и получаваме, че p-value също e по-малко от 0.05, т.е. има статистически значима разлика между групите. Вижда се и на графиката. Медианата на група "4+" значително се различава от тези на другите групи, които са почти еднакви. Също обхватът и е доста по-голям сравнение с останалите, и можем да кажем, че годините на пациентите с повече от 4 бременности варират много повече.

* Pregnancies VS BMI

Проверяваме вида на разпределение аналогично на предната точка и получаваме, че при две от групите p-value > 0.05, а при другите две е по-малко. При групи "0" и "4+" имаме нормално разпределение, а при "1" и "[2, 3]" – не. Пак използваме Kruskal и получаваме p-

value по-малко от 0.05 – има статистически значима разлика между групите.

Chart, box and whisker chart

Description automatically generated От графиката се вижда, че обхватът на група "0" е по-голям от на останалите, както и медианата. Също разпределението на стойностите на BMI на пациентите с 0 или 1 бременност е по-голямо от това на другите групи.

* Age VS BMI

От cor.test() получаваме, че p-value = 0.047 < 0.05, коефициентът на корелация е 0.087 < 0.3, което означава слаба, възходяща корелация - не зависят една от друга. Също така се вижда и от графиката, че променливите нямат линейна връзка.

Следователно индекса на телесната маса на пациентите не зависи от тяхната възраст.

Chart, scatter chart

Description automatically generated