t-критерий Стьюдента — общее название для статистических тестов, в которых статистика критерия имеет распределение Стьюдента. Наиболее часто t-критерии применяются для проверки равенства средних значений в двух выборках. Нулевая гипотеза предполагает, что средние равны (отрицание этого предположения называют гипотезой сдвига). Для применения данного критерия необходимо, чтобы исходные данные имели нормальное распределение.

*Число степеней свободы* – это количество элементов, которые могут варьироваться при расчете некоторого статистического показателя. Подробнее про степени свободы можно почитать здесь: <https://habr.com/ru/company/stepic/blog/311354/>.

Считываем датасет Ирисы Фишера

?iris

df <- iris

str(df)

df1 <- subset(df, Species != "setosa") # информация по 2 видам цветка

str(df1)

table(df1$Species) #таблица сопряженности

Хотим сравнить переменную Sepal.Length по 2-м типам цветка

hist(df1$Sepal.Length) #гистограмма распределения похожа на нормальное распределение, что хорошо для критерия Стьюдента

ggplot(df1, aes(x =Sepal.Length ))+

geom\_histogram(fill = "white", col = "black", binwidth = 0.4)+

facet\_grid(Species ~ .) # с помощью параметра facet\_grid строим 2 гистограммы (для каждого вида цветков)

ggplot(df1, aes(Sepal.Length, fill = Species ))+

geom\_density(alpha = 0.5) # функция плотности, alpha задает здесь прозрачность

ggplot(df1, aes(Species, Sepal.Length))+

geom\_boxplot() # ящик с усами строим

Для применения теста Стьюдента нужно выполнение условий: нормальность распределения переменных и гомогенность дисперсий. Проверим оба условия.

shapiro.test(df1$Sepal.Length) # тест Шапиро-Уилка для проверки на нормальность

Проверяем каждую переменную на нормальность

shapiro.test(df1$Sepal.Length[df1$Species == "versicolor"])

shapiro.test(df1$Sepal.Length[df1$Species == "virginica"])

#The same result in one line

by(df1$Sepal.Length, INDICES = df1$Species, shapiro.test)

Гомогенность дисперсий:

Критерий Бартлетта (англ. Bartlett's test) — статистический критерий, позволяющий проверять равенство дисперсий нескольких (двух и более) выборок. Нулевая гипотеза предполагает, что рассматриваемые выборки получены из генеральных совокупностей, обладающих одинаковыми дисперсиями. Критерий Бартлетта является параметрическим и основан на дополнительном предположении о нормальности выборок данных. Поэтому перед применением критерия Бартлетта рекомендуется выполнить проверку нормальности. Критерий Бартлетта очень чувствителен к нарушению данного предположения.

bartlett.test(Sepal.Length ~ Species, df1) # тест Бартлетт

Проводим тест Стьюдента:

t.test(Sepal.Length ~ Species, df1)

test1 <- t.test(Sepal.Length ~ Species, df1)

str(test1)

test1$p.value

t.test(Sepal.Length ~ Species, df1, var.equal = T) #указываем, то дисперсии гомогенны

t.test(df1$Sepal.Length, mu = 8) # проверяем гипотезу о равенстве 8 среднего значения переменной

Проверим нулевую гипотезу о том, что длина и ширина лепестков равны:

t.test(df1$Petal.Length, df1$Petal.Width, paired = T) # paired означает, что выборки зависимы

ggplot(df1, aes(Species, Sepal.Length))+

stat\_summary(fun.data = mean\_cl\_normal, geom = "errorbar",

width = 0.1)+

stat\_summary(fun.y = mean, geom = "point", size = 4) # отображаем среднее и доверительные интервалы

ggplot(df1, aes(Species, Sepal.Length))+

stat\_summary(fun.data = mean\_cl\_normal, geom = "pointrange",

size = 2)

Критерий Уилкоксона-Манна-Уитни — непараметрический статистический критерий, используемый для оценки различий между двумя выборками по признаку, измеренному в количественной или порядковой шкале. U-критерий является ранговым, поэтому он инвариантен по отношению к любому монотонному преобразованию шкалы измерения.

?wilcox.test

test2 <- wilcox.test(Petal.Length ~ Species, df1)

pv <- test2$p.value

ggplot(df1, aes(Species, Petal.Length))+

geom\_boxplot()

wilcox.test(df1$Petal.Length, df1$Petal.Width, paired = T)

paired\_wtest <- wilcox.test(df1$Petal.Length, df1$Petal.Width, paired = T)

paired\_wtest$p.value

***Памятка***

t-Критерий Стьюдента для независимых выборок

t.test(Var1 ~ Var2, data) # если первая переменная количественная, а вторая фактор

t.test(data$Var1, data$Var2) # если обе переменные количественные

t-Критерий Стьюдента для зависимых выборок

t.test(data$Var1, data$Var2, paired = T)

Проверка на нормальность распределения

shapiro.test(Var1) # проверка на нормальность распределения переменной Var1

# но не удобно когда есть группирующая факторная переменная

Поможет функция by(), которая применяет различные функции на каждом уровне фактора.

by(iris$Sepal.Length, INDICES = iris$Species, shapiro.test) # проверка на нормальность переменной

# Sepal.Length в трех разных группах в соответствии с переменной Species

Проверка на гомогенность дисперсий

bartlett.test(mpg ~ am, mtcars) #Критерий Бартлетта