«КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Лабораторная работа №10 по теме: «Проверка гипотезы о равенстве дисперсий»

Выполнил

Студент 1 курса

группы 09-115(3)

Зиновьев Е. А.

Преподаватель:

Шустова Е.П

Казань 2021

**Подготовка к выполнению лабораторной работы:**

1.         Прочитайте материал параграфа 5.4 (см. стр 147) из книги С.Э. Мастицкий, В.К. Шитиков Статистический анализ и визуализация данных с помощью R.

**Выполните следующее:**

1. Пусть Х – это тот признак, который Вы наблюдаете в двух выборках . Выберите свои данные из репозитория данных для выявления различий в двух выборках, если они сформированы при наблюдениях в разных условиях за одним и тем же признаком Х. Приведите краткое описание этих  данных и признака. Проверьте, что каждая из выборок распределена по нормальному закону. Можете использовать те же выборки что и в ЛР9.
2. Проверьте гипотезу о равенстве дисперсий этих двух выборок.

Ход работы:

1. Был выбран следующий датасет:

<http://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work>

Данный датасет показывает социально-физиологические особенности людей, которые отсутствовали на работе, такие как рост, вес, возраст; рассматриваются категории причин прогулов людей по здоровью, такие как проблемы с дыхательной системой, нервной, кровеносной и т.д.; временные факторы, такие как день, месяц прогула, время до места работы. То есть данный набор данный может рассказать о том, по каким причинам чаще всего отсутствовали люди на работе.

Информация об Атрибутах:

1. Индивидуальная идентификация (ID)

2. Причина отсутствия (МКБ).

Отсутствие, подтвержденное Международным кодексом болезней (МКБ), разделено на 21 категорию (с I по XXI) следующим образом:

I Некоторые инфекционные и паразитарные заболевания

II Новообразования

III Заболевания крови и кроветворных органов и некоторые нарушения, связанные с иммунным механизмом

IV Эндокринные, пищевые и метаболические заболевания

V Психические расстройства и расстройства поведения

VI Заболевания нервной системы

VII Заболевания глаз и придаточного аппарата

VIII Заболевания уха и сосцевидного отростка

IX Болезни системы кровообращения

X Заболевания дыхательной системы

XI Заболевания пищеварительной системы

XII Заболевания кожи и подкожной клетчатки

XIII Заболевания опорно-двигательного аппарата и соединительной ткани

XIV Болезни мочеполовой системы

XV Беременность, роды и послеродовой период

XVI Определенные состояния, возникающие в перинатальный период

XVII Врожденные пороки развития, деформации и хромосомные аномалии

XVIII Симптомы, признаки и аномальные клинические и лабораторные результаты, не классифицированные в других рубриках

XIX Травмы, отравления и некоторые другие последствия внешних причин

XX Внешние причины заболеваемости и смертности

XXI Факторы, влияющие на состояние здоровья и контакты с медицинскими службами.

И 7 категорий без (CID) наблюдения за пациентами (22), медицинской консультации (23), сдачи крови (24), лабораторного обследования (25), необоснованного отсутствия (26), физиотерапии (27), консультации стоматолога (28).

3. Месяц отсутствия

4. День недели (понедельник (2), вторник (3), Среда (4), Четверг (5), Пятница (6))

5. Времена года (лето (1), осень (2), зима (3), весна (4))

6. Транспортные расходы

7. Расстояние от места жительства до работы (км)

8. Время обслуживания

9. Возраст

10. Средняя рабочая нагрузка в день

11. Поразите цель

12. Дисциплинарный проступок (да=1; нет=0)

13. Образование (средняя школа (1), выпускник (2), аспирант (3), магистр и доктор (4))

14. Количество детей

15. Пьющий (да= 1; нет=0)

16. Курильщик (да= 1; нет=0)

17. Домашнее животное (количество домашних животных)

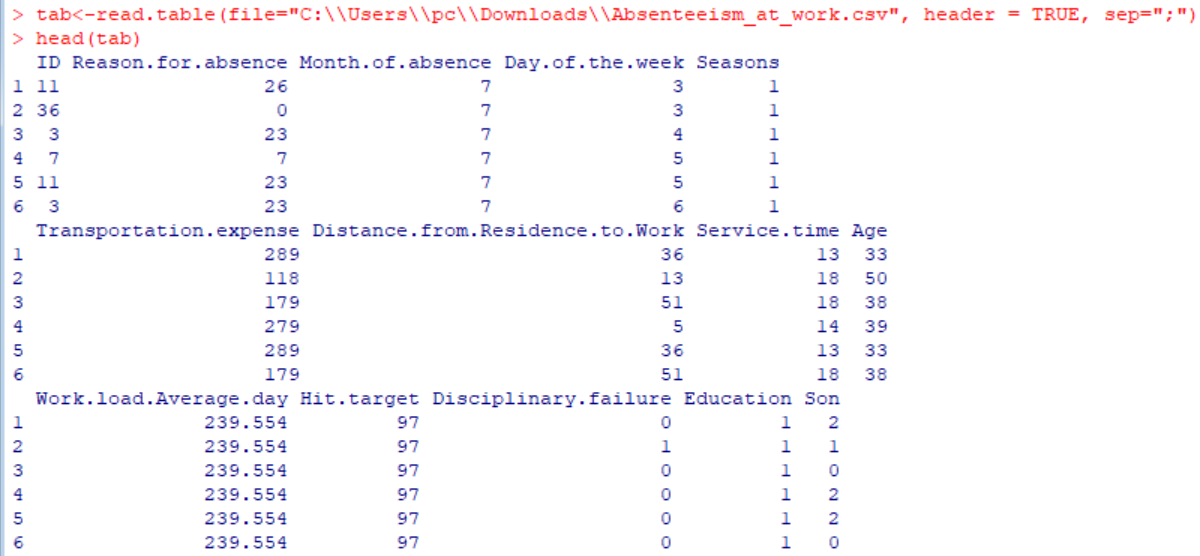
18. Вес

19. Рост

20. Индекс массы тела

21. Время прогула в часах

Выведем этот датасет:

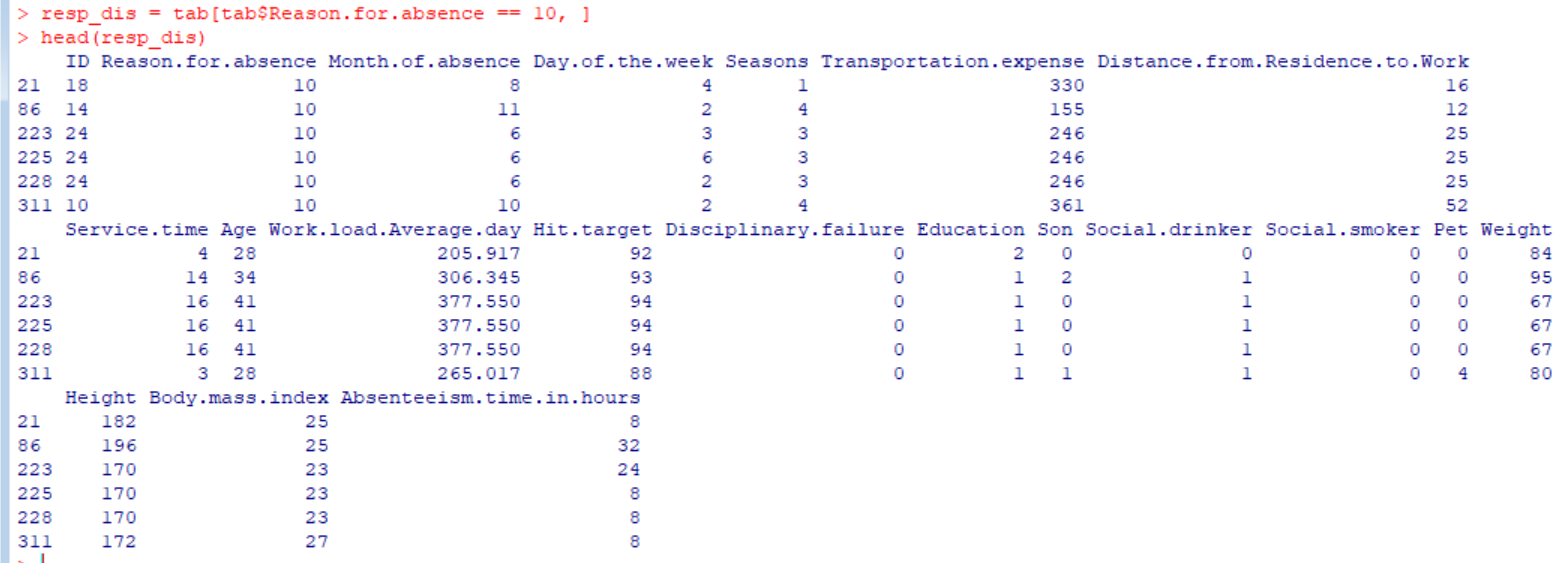


В качестве исследуемого параметра выбран возраст прогульщиков. В качестве условий были выбраны причины прогула, связанные с заболеваниями дыхательной системы (2.Х) и прогулы без обоснованных причин (2.26).

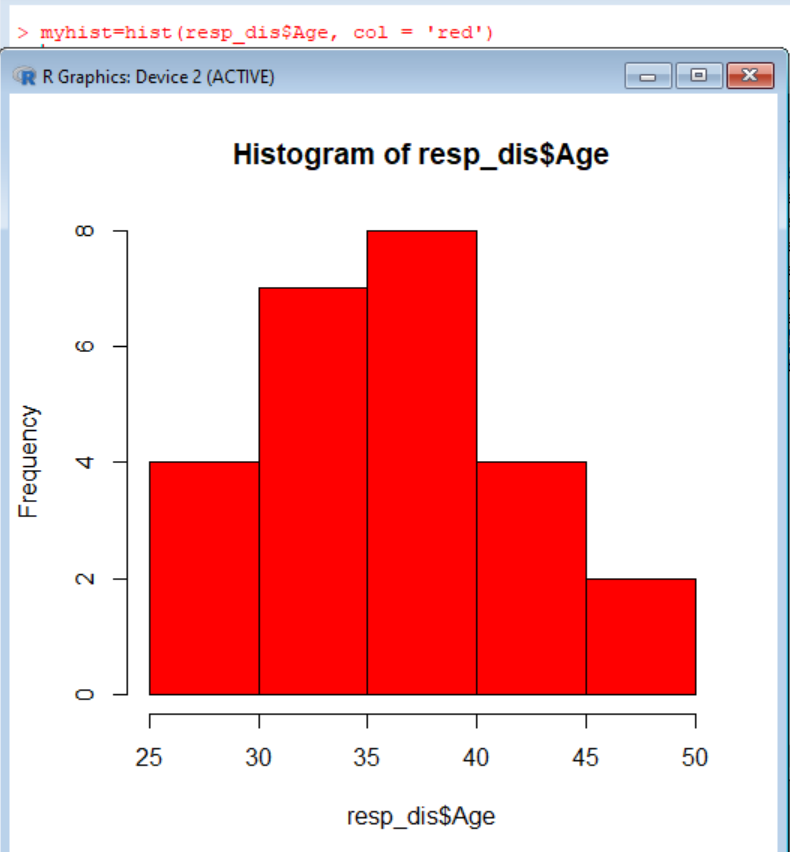
Проверка на нормальность данных исследуемого параметра при указанных условиях:

Первое условие. Проверить распределение данных на нормальность возраста людей, которые прогуливали работу в связи с заболеваниями дыхательной системы.

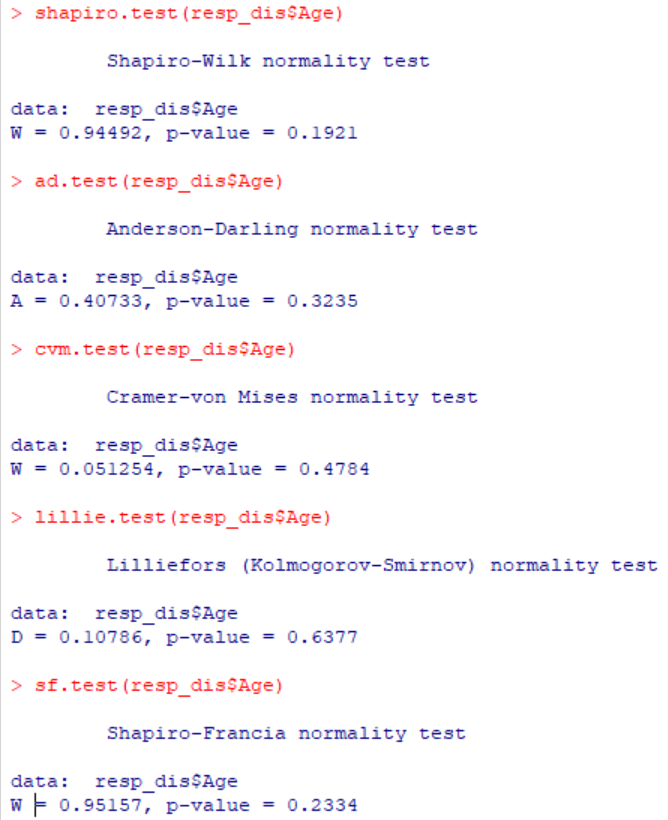
Оставим только строки, в которых причина отсутствия (Reason.for.absence) = 10:



Посмотрим на гистограмму распределения возраста для данной причины:



По гистограмме прослеживается нормальная куполообразная форма. Посмотрим теперь количественно, как проходит данная выборка тесты на нормальность Шапиро-Франсия, Колмогорова-Смирнова, Крамера фон Мизеса, Андерсона-Дарлинга.

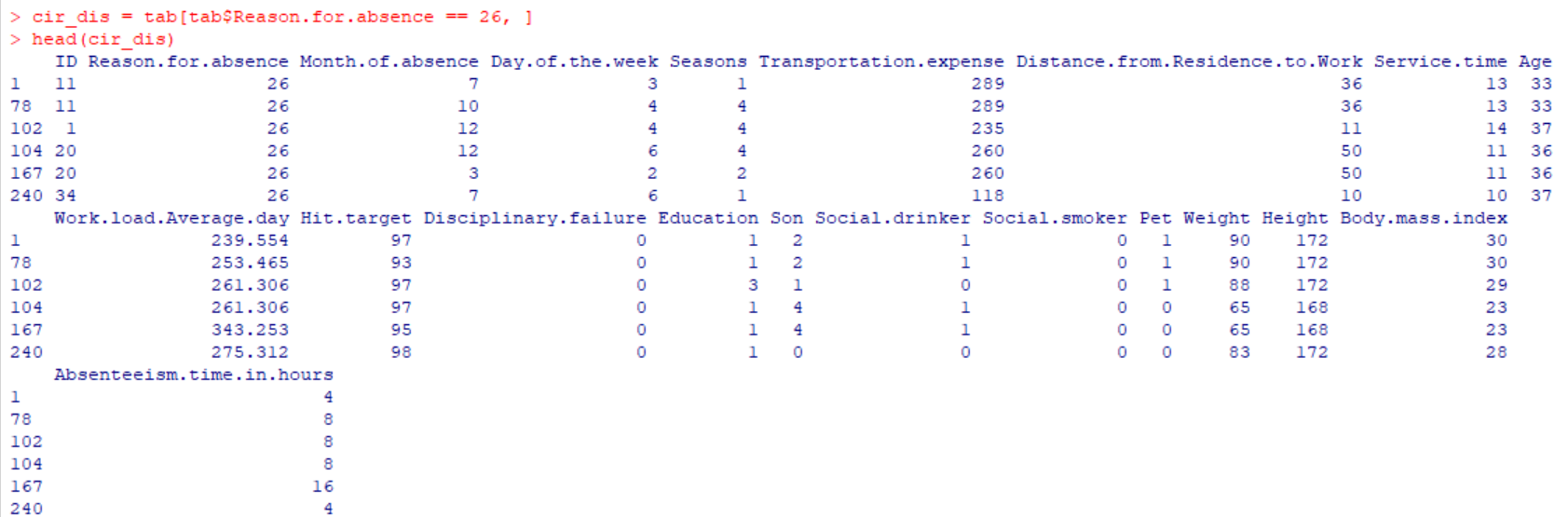


Как видим по p-value > 0.05, ни на одном тесте не была отклонена нулевая гипотеза о нормальности распределения данных.

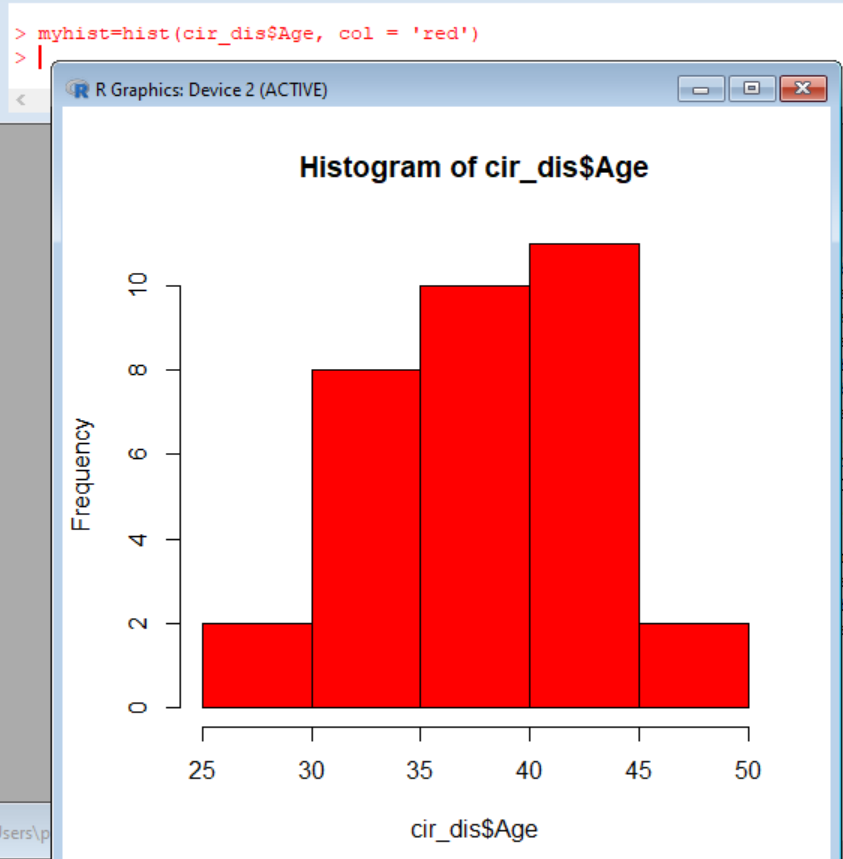
Таким образом, и качественно и количественно выборка прошла тесты на нормальность.

Второе условие. Проверить распределение данных на нормальность возраста людей, которые прогуливали работу в связи с необоснованными причинами.

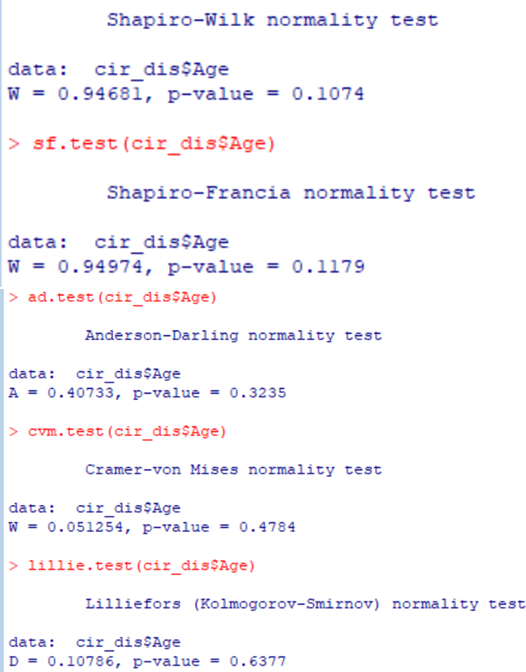
Оставим только строки, в которых причина отсутствия (Reason.for.absence) = 26:



Посмотрим на гистограмму распределения возраста для данной причины:



По гистограмме прослеживается нормальная куполообразная форма, хоть и скошенная набок, говоря о том, что молодые чаще пропускают по необоснованным причинам. Посмотрим теперь количественно, как проходит данная выборка тесты на нормальность Шапиро-Франсия, Колмогорова-Смирнова, Крамера фон Мизеса, Андерсона-Дарлинга.

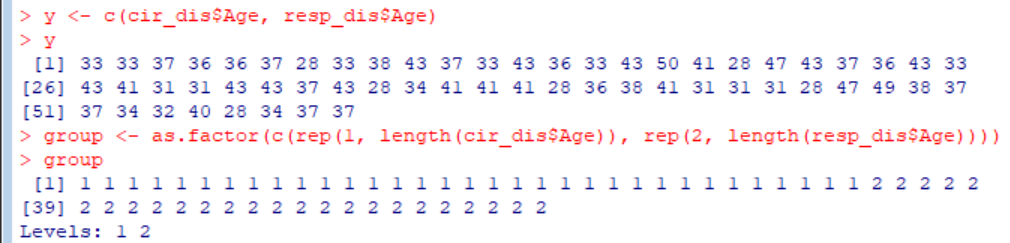


Как видим по p-value > 0.05, ни на одном тесте не была отклонена нулевая гипотеза о нормальности распределения данных.

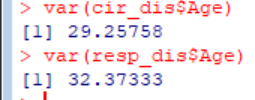
Таким образом, и качественно и количественно выборка прошла тесты на нормальность.

1. Так как 2 выборки распределены нормально, то мы можем проверить на однородность дисперсий этих выборок с помощью использования различных классических и непараметрических тестов: Левене, Бартлетта, Флигнера.

Для этого сначала соединим эти 2 выборки в одну общую и присвоим элементам каждой соответствующую этой группе категорию:

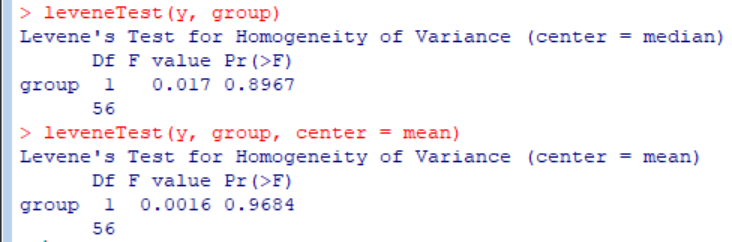


Посмотрим для начала значения дисперсий у этих 2 выборок:



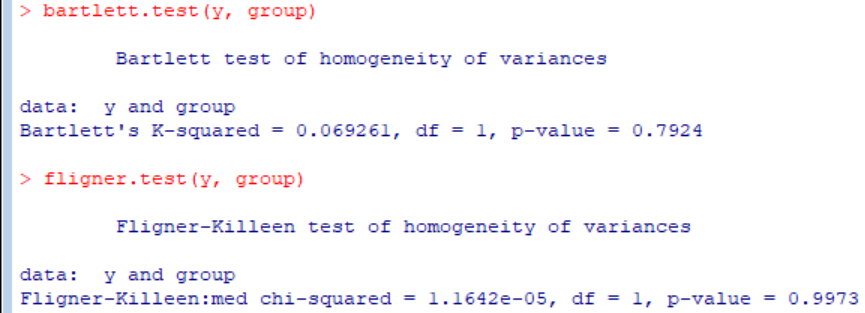
Как видим, они довольно близки.

Проверим по тесту Левене:



По тесту Левене нельзя отклонить гипотезу об однородности дисперсий, как при использовании для анализа среднего, так и медианы, причем довольно уверенно (коэффициент р близок к 1).

Проверим теперь на тестах Бартлетта и Флигнера:

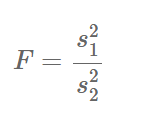


По этим тестам также можем уверенно сказать, что различие дисперсий 2 данных выборок статистически не значимо (p-value близок к 1).

Проверим теперь с помощью критерия Фишера. В общем виде критерий Фишера F, или F-тест, используется для сравнения дисперсий двух генеральных нормально распределенных совокупностей, т.е проверятся следующая нулевая гипотеза:



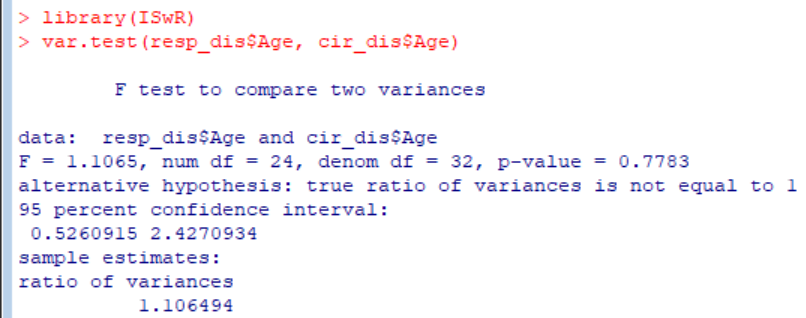
Генеральные дисперсии оцениваются на основе выборок, и сам критерий непосредственно рассчитывается как отношение одной выборочной дисперсии к другой:



На практике в числитель приведенной формулы обычно помещают бóльшую дисперсию, а в знаменатель – меньшую.

Очевидно, что чем ближе рассчитанное значение F к 1, тем больше у нас оснований сделать заключение о справедливости приведенной выше нулевой гипотезы. И наоборот - чем больше это значение, тем больше имеется оснований отклонить нулевую гипотезу о равенстве дисперсий. Критическое значение F, начиная с которого нулевую гипотезу отклоняют, определятся уровнем значимости (например, α = 0.05) и количеством степеней свободы для каждой из сравниваемых дисперсий. Кроме того, нулевую гипотезу можно проверить при помощи Р-значения для F-критерия, т.е. вероятности того, что случайная величина с соответствующим распределением Фишера окажется равной или превысит рассчитанное по выборочным данным значение F.

Для выполнения теста Фишера в R имеется функция var.test():



По выходу данной функции можно увидеть, что р-значение сильно больше 0.05, значит, так же, как и в других выше рассмотренных методах, различие дисперсий 2 данных выборок статистически не значимо. Также на выходе этой функции можно увидеть кол-во степеней свободы (объем) выборок num и denom, значение отношения дисперсий F – как видим, оно очень близко к 1, а также можно сказать, что истинное отношение сравниваемых дисперсий с вероятностью 95% находится в интервале от 0.53 до 2.43 ( 95 percent confidence interval).

Код программы:

tab<-read.table(file="C:\\Users\\pc\\Downloads\\Absenteeism\_at\_work.csv", header = TRUE, sep=";")

head(tab)

resp\_dis = tab[tab$Reason.for.absence == 10, ]

head(resp\_dis)

myhist=hist(resp\_dis$Age, col = 'red')

library(nortest)

shapiro.test(resp\_dis$Age)

ad.test(resp\_dis$Age)

cvm.test(resp\_dis$Age)

lillie.test(resp\_dis$Age)

sf.test(resp\_dis$Age)

cir\_dis = tab[tab$Reason.for.absence == 26, ]

head(cir\_dis)

myhist=hist(cir\_dis$Age, col = 'red')

shapiro.test(cir\_dis$Age)

sf.test(cir\_dis$Age)

ad.test(cir\_dis$Age)

cvm.test(cir\_dis$Age)

lillie.test(cir\_dis$Age)

library(car)

y <- c(cir\_dis$Age, resp\_dis$Age)

group <- as.factor(c(rep(1, length(cir\_dis$Age)), rep(2, length(resp\_dis$Age))))

var(cir\_dis$Age)

var(resp\_dis$Age)

leveneTest(y, group)

leveneTest(y, group, center = mean)

bartlett.test(y, group)

fligner.test(y, group)

library(ISwR)

var.test(resp\_dis$Age, cir\_dis$Age)