«КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Лабораторная работа №9 по теме: «Ошибки I и II рода. Проведение анализа мощности»

Выполнил

Студент 1 курса

группы 09-115(3)

Зиновьев Е. А.

Преподаватель:

Шустова Е.П

Казань 2021

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.

2.    Ответьте на вопрос: по какой формуле рассчитывается величина эффекта? (ответ см. в [2], стр. 341).

3.    Сколько наблюдений должно быть в каждой из групп,  чтобы  обнаружить  эффект  размером:

* + 0.8  с  90%-ной  уверенностью  и  вероятностью  обнаружить  несуществующую закономерность не больше 5%.?
  + 0.9  с  95%-ной  уверенностью  и  вероятностью  обнаружить  несуществующую закономерность не больше 2%.?

4.    Представим, что, сравнивая две группы, вы хотите, чтобы можно было обнаружить различие между выборочными средними в 0.5 стандартного отклонения. Вы хотите снизить шансы *обнаружения различий по ошибке* (то есть обнаружение различий там, где их нет) до 1 из 100. Кроме того, вы можете исследовать только 20 представителей. С какой вероятностью вы сможете выявить различия между выборочными средними при таких условиях:

* если в обоих экспериментальных группах будет одинаковое количество наблюдений?
* если размер экспериментальных групп различается?

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)

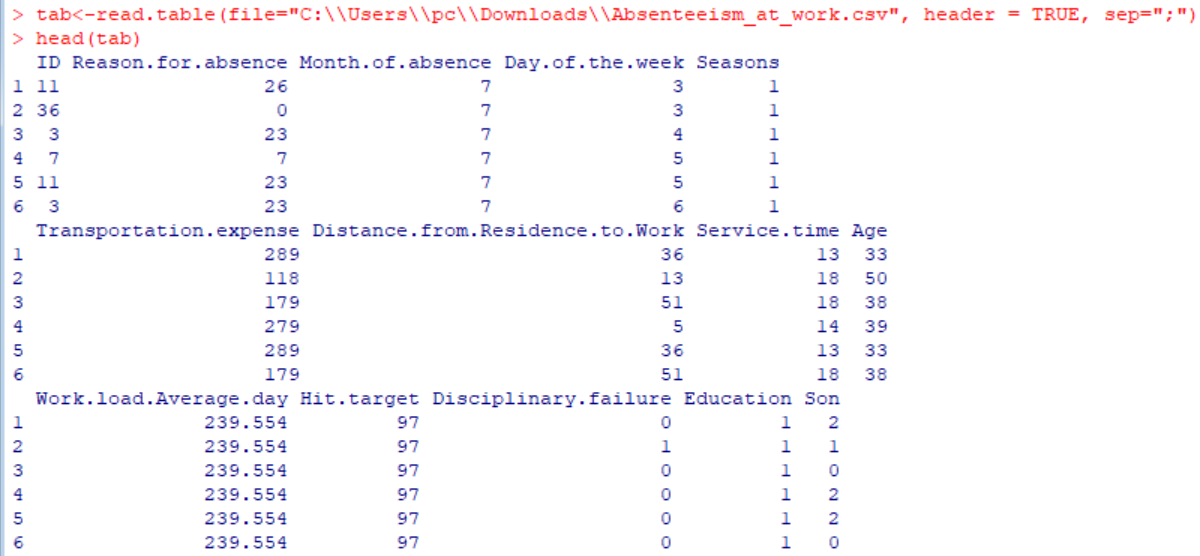
18. Вес

19. Рост

20. Индекс массы тела

21. Время прогула в часах

Выведем этот датасет:

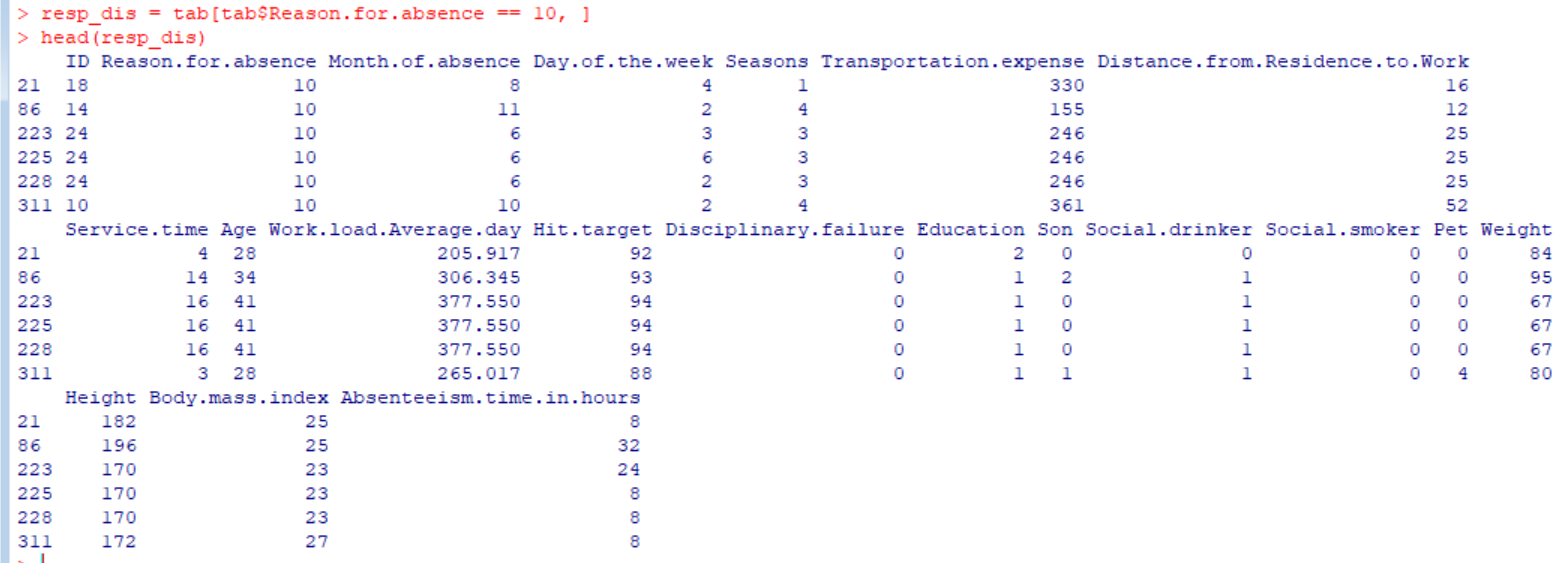


В качестве исследуемого параметра выбран возраст прогульщиков. В качестве условий были выбраны причины прогула, связанные с заболеваниями дыхательной системы (2.Х) и прогулы без обоснованных причин (2.26).

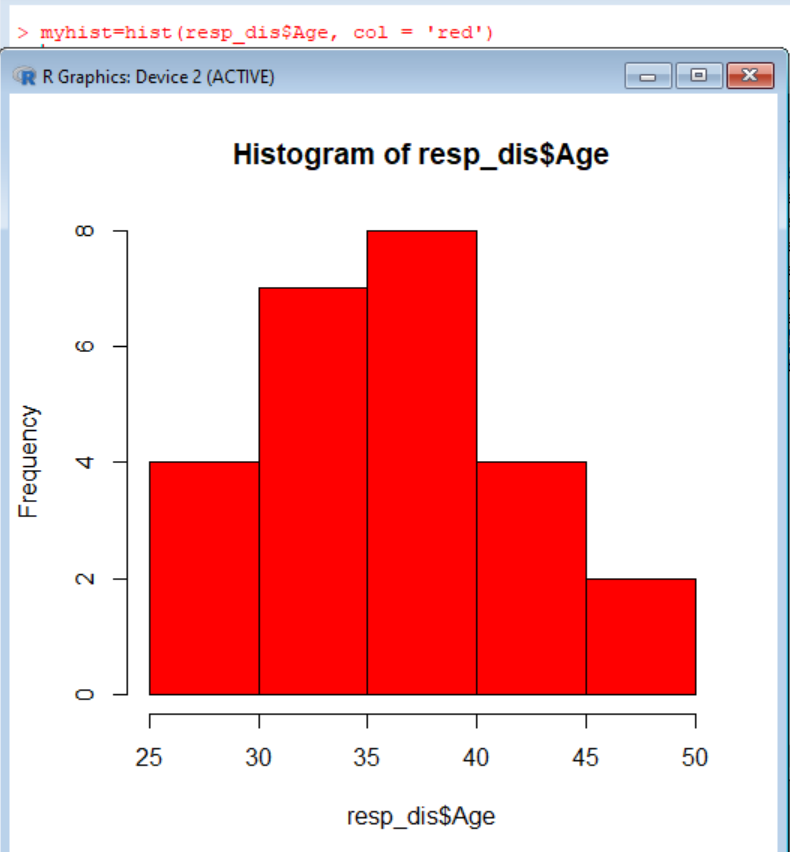
Проверка на нормальность данных исследуемого параметра при указанных условиях:

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

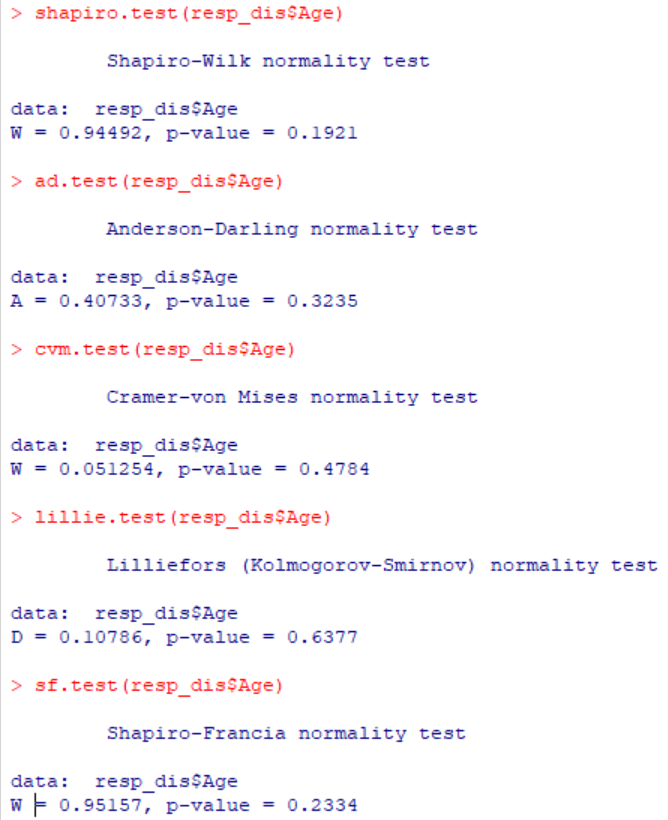
Оставим только строки, в которых причина отсутствия (Reason.for.absence) = 10:



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.

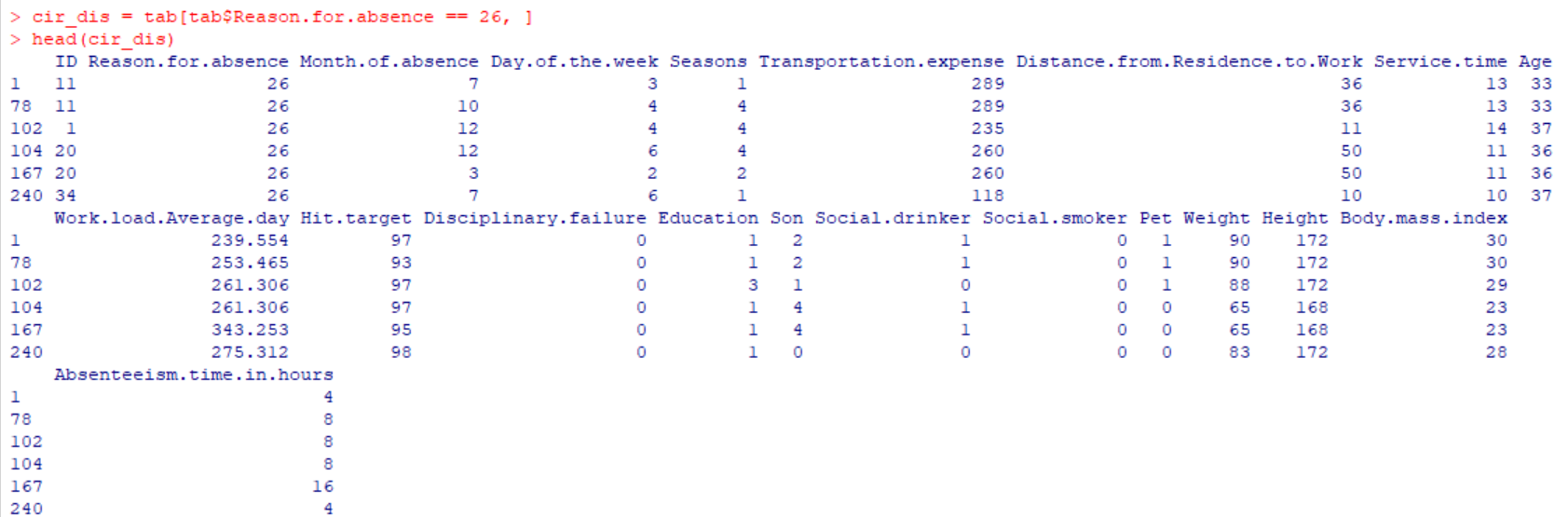


Как видим по p-value > 0.05, ни на одном тесте не была отклонена нулевая гипотеза о нормальности распределения данных.

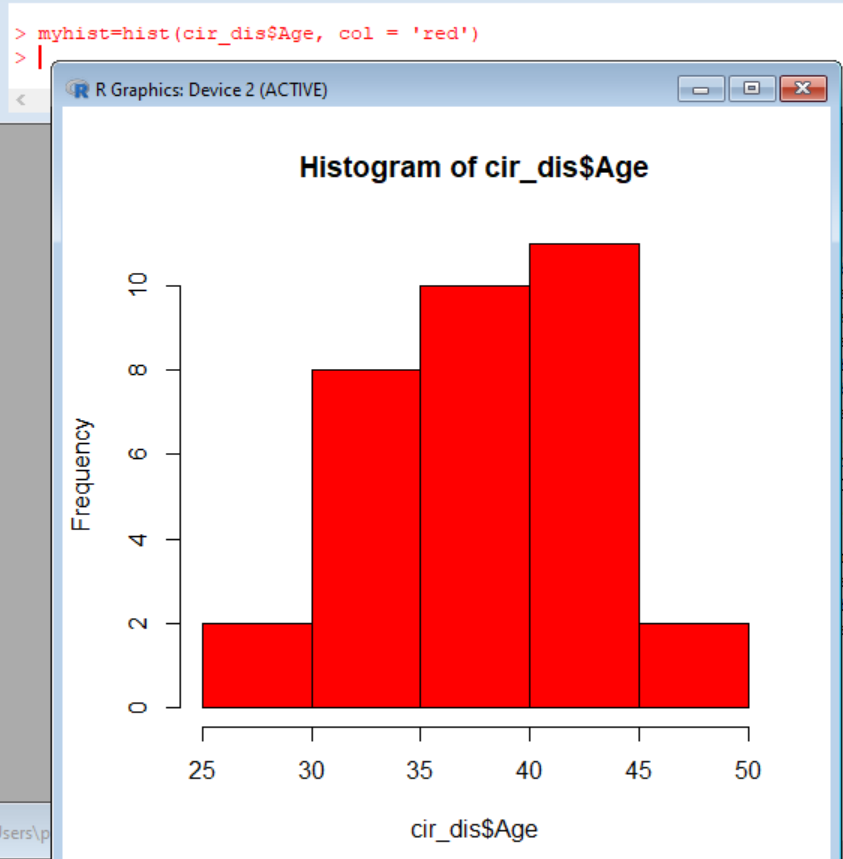
Таким образом, и качественно и количественно выборка прошла тесты на нормальность.

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

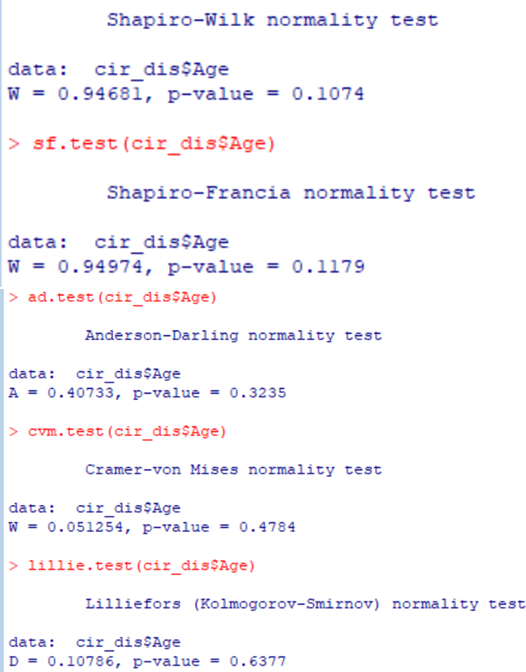
Оставим только строки, в которых причина отсутствия (Reason.for.absence) = 26:



Посмотрим на гистограмму распределения возраста для данной причины:



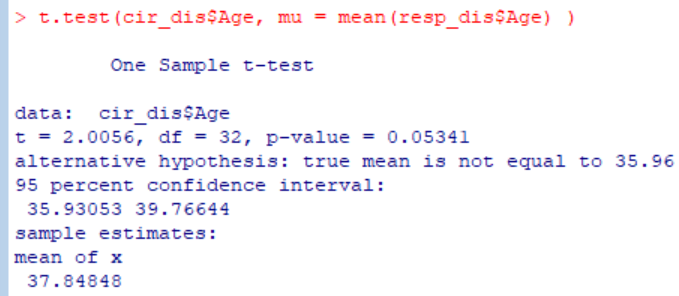
По гистограмме прослеживается нормальная куполообразная форма, хоть и скошенная набок, говоря о том, что молодые чаще пропускают по необоснованным причинам. Посмотрим теперь количественно, как проходит данная выборка тесты на нормальность Шапиро-Франсия, Колмогорова-Смирнова, Крамера фон Мизеса, Андерсона-Дарлинга.



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

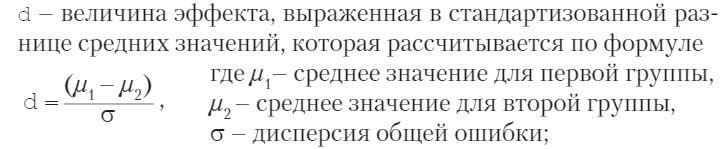
Таким образом, и качественно и количественно выборка прошла тесты на нормальность.

Так как 2 выборки распределены нормально, то мы можем использовать тест Стьюдента для того, чтобы выяснить наличие или отсутствие различий выборочных средних значений между двумя выборками:

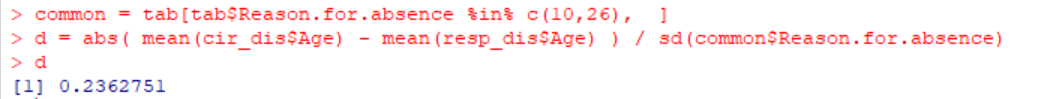


Видим, что для имеющихся выборочных данных t-критерий составляет 2 при 32 степенях свободы (df). Вероятность получить такое (либо большее) значение t при условии, что проверяемая нулевая гипотеза верна, оказалась мала: p-value = 0.053, но она превзошла 5% уровень доверия. Следовательно, мы не можем отклонить проверяемую нулевую гипотезу о равенстве выборочного среднего значения нормативу и принять альтернативную гипотезу (alternative hypothesis: true mean is not equal to 35.96). То есть средний возраст у людей с прогулами работы в связи с заболеваниями дыхательной системы и у людей с прогулами работы без обоснованных причин статистически не различим (у первой выборки – 35.96 лет, у второй – 37.85 лет), хотя и очень слабо.

1. Величина эффекта считается по следующей формуле:



Для наших выборок эффект равен 0.236:

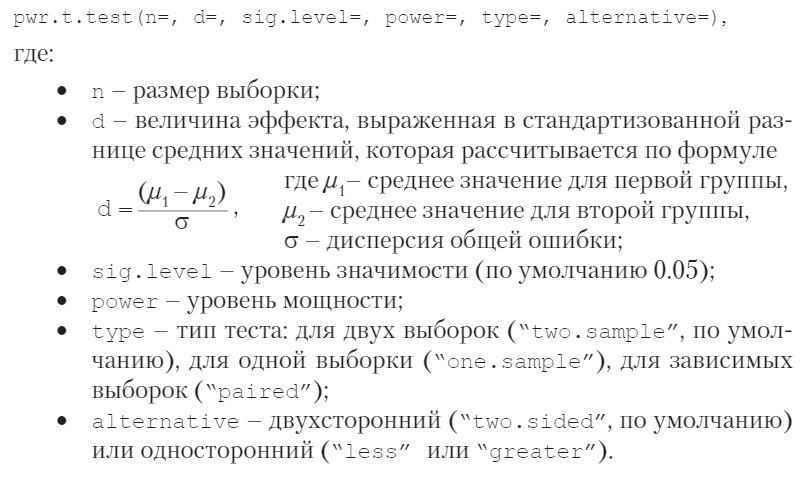


Чем выше мощность статистического теста, тем меньше вероятность совершить ошибку второго рода, что говорит о возможности допуска довольно большой ошибки.

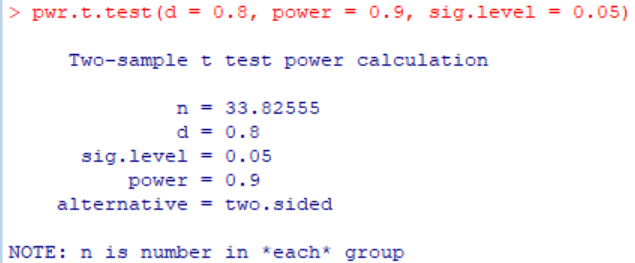
3. Сколько наблюдений должно быть в каждой из групп, чтобы обнаружить эффект размером 0.8 с 90%-ной уверенностью и вероятностью обнаружить несуществующую закономерность не больше 5%.?

Пакет pwr, созданный Стефаном Чемпили (Stéphane Champely), позволяет выполнить анализ мощности.

Когда нужно использовать тест Стьюдента, функция pwr.t.test предоставляет много полезных для анализа мощности возможностей. Формат ее применения таков:



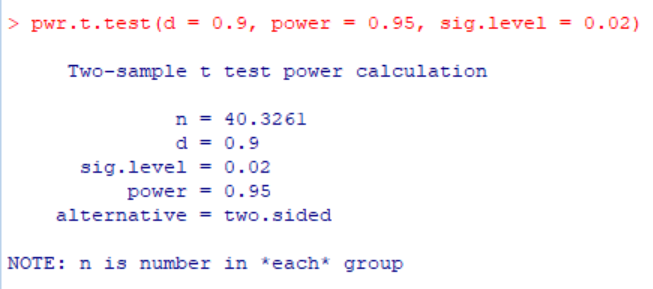
Для наших параметров sig.level, d, power размер выборки будет:



То есть чтобы обнаружить эффект размером 0.8 с 90%-ной уверенностью и вероятностью обнаружить несуществующую закономерность не больше 5% необходимо 68 элементов по 34 элемента в каждой выборке.

Сколько наблюдений должно быть в каждой из групп, чтобы обнаружить эффект размером 0.9 с 95%-ной уверенностью и вероятностью обнаружить несуществующую закономерность не больше 2%.?

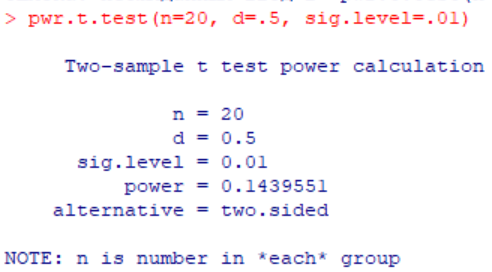
Для наших параметров sig.level, d, power размер выборки будет:



То есть чтобы обнаружить эффект размером 0.9 с 95%-ной уверенностью и вероятностью обнаружить несуществующую закономерность не больше 2% необходимо 82 элементов по 41 элементу в каждой выборке.

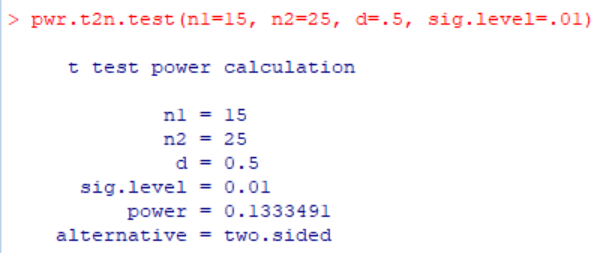
Таким образом, при увеличении размаха, уверенности и уменьшении вероятности обнаружения несуществующих закономерностей размер выборки также должен увеличиться.

4. Представим, что, сравнивая две группы, вы хотите, чтобы можно было обнаружить различие между выборочными средними в 0.5 стандартного отклонения. Вы хотите снизить шансы *обнаружения различий по ошибке* (то есть обнаружение различий там, где их нет) до 1 из 100. Кроме того, вы можете исследовать только 20 представителей. С какой вероятностью вы сможете выявить различия между выборочными средними при таких условиях, если в обоих экспериментальных группах будет одинаковое количество наблюдений?



При 20 участниках в группе, заданном уровне значимости 0.01 и мощности выборки d = 0.5 составляет 14%. Соответственно, вероятность пропустить искомый эффект составляет 86%.

Если размер выборок различается, можно использовать функцию:



При уменьшении одной выборки на 25% и увеличении другой на те же 25% при заданном уровне значимости 0.01 и мощности выборки d = 0.5 составляет 13%. Соответственно, вероятность пропустить искомый эффект составляет 87%, то есть слегка увеличилась по сравнению с двумя равными выборками.

Код:

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)

common = tab[tab$Reason.for.absence %in% c(10,26), ]

d = abs( mean(cir\_dis$Age) - mean(resp\_dis$Age) ) / sd(common$Reason.for.absence)

pwr.t.test(d = 0.8, power = 0.9, sig.level = 0.05)

pwr.t.test(d = 0.9, power = 0.95, sig.level = 0.02)

pwr.t.test(n=20, d=.5, sig.level=.01)

pwr.t2n.test(n1=15, n2=25, d=.5, sig.level=.01)

t.test(cir\_dis$Age, mu = mean(resp\_dis$Age) )