Московский государственный университет имени М. В. Ломоносова
Факультет Вычислительной математики и кибернетики

Никонов Максим Викторович
316 группа

2020

Задание 2: На собственных данных проверить использование методов хи-квадрат, точного теста Фишера, теста МакНемара, Кохрана-Мантеля-Хензеля.

Метод хи-квадрат можно сделать ручками, но есть функция **chisq.test** от “chi square” Параметр – не факторная переменная

```
chisq.test(data$Close)
```

Данные из протокола

```
Chi-squared test for given probabilities

data:  data$Close
X-squared = 12965, df = 1509, p-value < 2.2e-16
```

Для теста Фишера нужна выборка маленьких размеров, тест относится к точным тестам значимости, поскольку не использует приближения большой выборки (асимптотики при размере выборки стремящемся к бесконечности). У меня таких данных практически нет, поэтому сгенинированы столбцы num и num2 для демонстрации. Необходимы переменные с уровнем 2, для построения матрицы 2x2

```
fisher.test(data$num, data$num2)
```

Данные из протокола

```
> fisher.test(data$num, data$num2)

      Fisher's Exact Test for Count Data

data:  data$num and data$num2
p-value = 0.918
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.8234209 1.2458164
sample estimates:
odds ratio
  1.012819
```

Используем дальше столбцы, тк для моих данных либо не выполняются условия, либо данные слишком независимые

```
mcnemar.test(data$num, data$num2, correct = TRUE)
mantelhaen.test(data$Close, data$num, data$num2)
```

Данные из протокола

```
> mcnemar.test(data$num, data$num2, correct = TRUE)

      McNemar's Chi-squared test with continuity correction

data:  data$num and data$num2
McNemar's chi-squared = 1.5352, df = 1, p-value = 0.2153

> mantelhaen.test(data$Close, data$num, data$num2)
exit

      Cochran-Mantel-Haenszel test

data:  data$Close and data$num and data$num2
Cochran-Mantel-Haenszel M^2 = 1423.7, df = 1420, p-value = 0.4676
```

Тест МакНемара рассчитан для двух переменных, используется для анализа таблиц сопряженности размером 2x2 (для дихотомического признака). В отличие от критерия хи-квардрат, критерий МакНемара применяется, когда условие независимости наблюдений не выполняется, но, напротив, учет признака выполняется на одних и тех же субъектах

Тест Кохрана-Мантеля-Хензеля для таблиц размерности 2x2xK

Таблица сопряженности - средство представления совместного распределения двух переменных, предназначенное для исследования связи между ними. Таблица сопряженности является наиболее универсальным средством изучения статистических связей, так как в ней могут быть представлены переменные с любым уровнем измерения.

Итоговый вид программы

```
library(MASS)
library(dplyr)

head <- read.csv("/Users/Nikon/Desktop/CMC MSU/MC/5
sem/Data/AAPL.csv",
                 header = TRUE)
num <- random <- sample(1:2, nrow(head), replace=TRUE)
num2 <- random <- sample(1:2, nrow(head), replace=TRUE)
head <- cbind(head, num)
head <- cbind(head, num2)
data <- head[-c(2,3,4,6)]
data <- mutate_each(data, "factor", Volume, Date)
#data$Date <- as.POSIXct(data$Date, format = "%Y")
chisq.test(data$Close)
fisher.test(data$num, data$num2)
mcnemar.test(data$num, data$num2, correct = TRUE)
     mantelhaen.test(data$Close, data$num, data$num2)
```

Используемые packages. MASS, dplyr
Тестирование. Не требует
Неразрешенные вопросы. Нет
Новые функции. chisq.test, fisher.test, mcnemar.test, mantelhaen.test
Статус компиляции. ОК. Данные из протокола:

```
> ^M
^[[1m^[[7m%^[[27m^[[1m^[[0m
^M ^M^[]7;file://MBP-
Nikon.Dlink/Users/Nikon/Desktop/CMC%20MSU/MC/5%20sem/R/tz9/1^G^M^[[0m^[[27m^[[24m^[[JN
ikon@MBP-Nikon 1 % ^[[K^[[?2004he^Hexit^[[?2004l^M^M

                        Script done on Wed Nov 18 13:49:41 2020
```

Задание 3: Изучить возможности функции power.prop.test() и разобраться с вычислением статистической мощности при сравнении частот

Для расчета статастической значимости значения конверсии используется **power.prop.test()**

Из документации функции

```
power.prop.test(n = NULL, p1 = NULL, p2 = NULL, sig.level = 0.05,
            power = NULL,
            alternative = c("two.sided", "one.sided"),
            strict = FALSE, tol = .Machine$double.eps^0.25)
```

```
power.prop.test(n = 20000, p1 = 0.6, p2 = 0.3, sig.level
= 0.05,
                power = NULL,
                alternative = c("two.sided",
"one.sided"),
                strict = FALSE, tol =
.Machine$double.eps^0.25)

pwr.p.test(h = ES.h(p1 = 0.65, p2 = 0.50),
           sig.level = 0.05,
           power = 0.80)
```

Данные из протокола

```
Two-sample comparison of proportions power calculation

              n = 20000
             p1 = 0.6
             p2 = 0.3
      sig.level = 0.05
          power = 1
    alternative = two.sided
```

```
proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.3046927
              n = 84.54397
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Тест хи-кварта

```
v <- matrix(c(25, 70, 7, 80), ncol = 2, byrow = T)
v
chisq.test(v)
```

Данные из протокола

```
> v
     [,1] [,2]
[1,]   25   70
[2,]    7   80
> chisq.test(v)

      Pearson's Chi-squared test with Yates' continuity correction

data:  v
X-squared = 9.2374, df = 1, p-value = 0.002371
```

Вычисление статистической мощности

```
prop.power <- function(n1, n2, p1, p2) {
  twobytwo=matrix(NA, nrow = 10000, ncol = 4)
  twobytwo[,1] = rbinom(n = 10000, size = n1, prob = p1)
  twobytwo[,2] = n1-twobytwo[,1]
  twobytwo[,3] = rbinom(n = 10000, size = n2, prob = p2)
  twobytwo[,4] = n1 - twobytwo[, 3]
  a = rep(NA, 10000)
  chisq.test.v = function(x)
  as.numeric(chisq.test(matrix(x, ncol = 2), correct =
FALSE)[3])
  a = apply(twobytwo, 1, chisq.test.v)
  power = sum(ifelse(a < 0.05, 1, 0))/10000
  return(power)
}

prop.power(50, 50, 0.50, 0.30)
```

Данные из протокола

```
> prop.power(50, 50, 0.50, 0.30)
[1] 0.5545
```

pwr.p.test() из пакета pwr используется как раз для анализа мощности для одной доли (например, при сравнении ее с каким-либо ожидаемым значением, как сделано выше)

Итоговый вид программы

```
library(reshape2)
library(pwr)
library(dplyr)

head <- read.csv("/Users/Nikon/Desktop/CMC MSU/MC/5
sem/Data/AAPL.csv",
                 header = TRUE)
num <- random <- sample(1:2, nrow(head), replace=TRUE)
num2 <- random <- sample(1:2, nrow(head), replace=TRUE)
head <- cbind(head, num)
head <- cbind(head, num2)
data <- head[-c(2,3,4,6)]
data <- mutate_each(data, "factor", Volume, Date)

power.prop.test(n = 20000, p1 = 0.6, p2 = 0.3, sig.level = 0.05,
                 power = NULL,
                 alternative = c("two.sided", "one.sided"),
                 strict = FALSE, tol = .Machine$double.eps^0.25)

pwr.p.test(h = ES.h(p1 = 0.65, p2 = 0.50),
           sig.level = 0.05,
           power = 0.80)

v <- matrix(c(25, 70, 7, 80), ncol = 2, byrow = T)
v
chisq.test(v)

prop.power <- function(n1, n2, p1, p2) {
  twobytwo=matrix(NA, nrow = 10000, ncol = 4)
  twobytwo[,1] = rbinom(n = 10000, size = n1, prob = p1)
  twobytwo[,2] = n1-twobytwo[,1]
  twobytwo[,3] = rbinom(n = 10000, size = n2, prob = p2)
  twobytwo[,4] = n1 - twobytwo[, 3]
  a = rep(NA, 10000)
  chisq.test.v = function(x)
as.numeric(chisq.test(matrix(x, ncol = 2), correct = FALSE)[3])
  a = apply(twobytwo, 1, chisq.test.v)
  power = sum(ifelse(a < 0.05, 1, 0))/10000
  return(power)
}

prop.power(50, 50, 0.50, 0.30)
```

Используемые packages. reshape2, dplyr, pwr
Тестирование. Не требует
Неразрешенные вопросы. Нет
Новые функции. power.prop.test
Статус компиляции. OK. Данные из протокола:

```
[1] 0.5545^M
> ^M
^[[1m^[[7m%^[[27m^[[1m^[[0m
^M ^M^[]7;file://MBP-
Nikon.Dlink/Users/Nikon/Desktop/CMC%20MSU/MC/5%20sem/R/tz9/2^G^M^[[0m^[[27m^[[24m^[[JNikon@MB
P-Nikon 2 % ^[[K^[[?2004he^Hexit^[[?2004l^M^M

                        Script done on Wed Nov 18 14:40:51 2020
```