Отчет по домашнему заданию

Фахртдинов Т. А.

25 декабря 2019 г.

Шестая задача. Проверка гипотез однородности для независимых выборок.
Вариант 3.

```
TX <- c(27.6, 20.9, 55.6, 69, 23, 19.5, 8.9, 50.4)
CA <- c(39.9, 20.7, 26.6, 13.9, 23.6, 16.2, 29.9, 13.9, 65.2, 31.4, 26, 25)
OH <- c(1.1, 4.6, 0.7, 4, 0.7)
mean.TX <- mean(TX)
mean.CA <- mean(CA)
var.TX <- var(TX)
var.CA <- var(CA)
n1 <- length(TX)
n2 <- length(CA)
```

Критерий Фишера:

```
F <- var.CA / var.TX
```

Значение критерия и p - value:

```
## [1] 0.4394916 0.2844862
```

Найдем значение критерия с помощью встроенной функции:

```
var.test(CA, TX)

##
##  F test to compare two variances
##
## data:  CA and TX
## F = 0.43949, num df = 11, denom df = 7, p-value = 0.215
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.09332083 1.65188989
## sample estimates:
## ratio of variances
##          0.4394916
```

Значение, которое получили мы совпало со значением встроенной функции.
p-value > 0.05 — нет оснований отклонить гипотезу о равенстве дисперсий.

Критерий Стьюдента:

Т.к. мы не отклонили гипотезу о равенстве дисперсий, используется следующий критерий Стьюдента (неизвестные одинаковые дисперсии):

$$T = \frac{(\bar{x} - \bar{y})\sqrt{n_1 + n_2 - 2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}} \sim \mathbf{T}(n_1 + n_2 - 2).$$

$H_0 : \mu_1 = \mu_2$

```
T <- (mean.TX - mean.CA) * sqrt(n1 + n2 - 2) /
     (sqrt(1/n1 + 1/n2) * sqrt((n1-1) * var.TX + (n2 - 1) * var.CA))
```

Значение критерия и p - value:

```
## [1] 0.8519455 0.4054371
```

Найдем значение критерия с помощью встроенной функции:

```
t.test(TX, CA, var.equal = TRUE)

##
##  Two Sample t-test
##
## data:  TX and CA
## t = 0.85195, df = 18, p-value = 0.4054
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.779633 23.121300
## sample estimates:
## mean of x mean of y
##  34.36250  27.69167
```

Значение, которое получили мы совпало со значением встроенной функции.
p-value > 0.05 — нет оснований отклонить нулевую гипотезу о равенстве средних.
Для TX, среднее ±ошибка среднего: 34.363 ±7.480.
Для CA, среднее ±ошибка среднего: 27.692 ±4.049.

Найдем доверительные интервалы для средних, возьмем уровень значимости $\alpha = 0.05$:
Для среднего TX:

```
alpha <- 0.05

q = qt(1 - alpha/2, length(TX) - 1) * sd(TX)/sqrt(length(TX))
c(mean(TX) - q, mean(TX) + q)

## [1] 16.6749 52.0501
```

Для среднего CA:

```
q = qt(1 - alpha/2, length(CA) - 1) * sd(CA)/sqrt(length(CA))
c(mean(CA) - q, mean(CA) + q)

## [1] 18.78011 36.60323
```

Для среднего OH:

```
q = qt(1 - alpha/2, length(OH) - 1) * sd(OH)/sqrt(length(OH))
c(mean(OH) - q, mean(OH) + q)

## [1] -0.1609534  4.6009534
```

Используя статистику Фишера, проверить однородность выборок, относящихся ко всем штатам одновременно.

```
r <- 3
N <- length(c(OH, TX, CA))
len <- c(length(OH), length(TX), length(CA))
m <- mean(c(OH, TX, CA))
x <- c(mean(OH), mean(TX), mean(CA))
Q1 <- sum(len * (x - m)^2)
Q2 <- sum((OH - x[1])^2) + sum((TX - x[2])^2) + sum((CA - x[3])^2)
F <- (Q1/(r - 1)) / (Q2/(N - r))
```

Значение критерия и p-value:

```
## [1] 7.00145557 0.00443543
```

Найдем значение критерия с помощью встроенной функции:

```
CITY <-data.frame(hisp = c(OH,TX,CA),
                  state = rep(c("OH","TX","CA"),
                  len))
summary(aov(hisp~state, data = CITY))

##             Df Sum Sq Mean Sq F value  Pr(>F)
## state        2   3381  1690.5   7.001 0.00444 **
## Residuals   22   5312   241.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Значение, которое получили мы совпало со значением встроенной функции.
p-value < 0.5 — Нулевая гипотеза о равенстве средних отвергается.
Для TX, среднее ±ошибка среднего: 34.363 ±7.480.
Для CA, среднее ±ошибка среднего: 27.692 ±4.049.
Для OH, среднее ±ошибка среднего: 2.22 ±0.858.
Проверяем значимость парных статистик (наведение контрастов):

```
temp = sqrt(Q2 / (N - r))
T1 <- (mean.TX - mean.CA) / (temp * sqrt(1 / n1 + 1 / n2))
T2 <- (mean.TX - mean(OH)) / (temp * sqrt(1 / n1 + 1 / length(OH)))
T3 <- (mean.CA - mean(OH)) / (temp * sqrt(1 / n2 + 1 / length(OH)))
```

Значения критерия и соответствующии p-value:

```
## [1] 0.9405563 3.6284578 3.0795881
## [1] 0.357145817 0.001485936 0.005480267
```

Воспользуемся встроенной функцией:

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  c and y
##
##   1      2
## 2 0.0015 -
## 3 0.0055 0.3571
##
## P value adjustment method: none
```

Значения совпадают с полученными нами.
$0.357145817 > 0.05$, нет оснований отклонить гипотезу о незначительности отклонений внутригрупповых средних у TX и CA.
Для остальных пар p-value < 0.05 и гипотеза о незначительности отклонений внутригрупповых средних отклоняется.

Построим boxplot для наших данных:

```
a <- data.frame(hisp = TX, state = 'TX')
b <- data.frame(hisp = CA, state = 'CA')
c <- data.frame(hisp = OH, state = 'OH')
STATES <- rbind(a, b, c)

means <- aggregate(hisp ~ state, STATES, mean)
means$hisp <- round(means$hisp, 3)
library(ggplot2)
ggplot(data=STATES, aes(x=state, y=hisp)) + geom_boxplot(aes(fill=state)) +
stat_summary(fun.y = mean, colour="red", geom="point",
shape=18, size=2,show.legend  = FALSE) +
geom_text(data = means, aes(label = hisp, y = hisp + 1),size = 2 )
```