Методы статистической обработки информации. Задание 5

Ершов А. С., гр. 22.М04-мм

Вариант 14.

- Вариант для проверки однородности двух независимых выборок: метрическая переменная `SBP.1`, категориальная переменная `tremor.1`.
- Вариант для проверки однородности трех независимых выборок: метрическая переменная `SV.1`, первый фактор (категориальная переменная из варианта) - `insomnia.1`, второй фактор - произвольная категориальная переменная).

Задание

- Задача однородности в случае двух выборок (табл. Варианты для проверки однородности двух независимых выборок). При наличии трех градаций нужно объединить ячейки, чтобы получить две наиболее представленные градации. Проверить на предмет однородности независимых выборок метрическую переменную (столбцы таблицы) в зависимости от категориальной переменной (строки таблицы) по критерию 1) Вилкоксона, 2) Фишера равенства дисперсий, 3) Стьюдента равенства средних. Привести значения средних с ошибками среднего, медиан с интерквартильным размахом, значимости соответствующих критериев.
- Задача однородности в случае более двух выборок (табл. Варианты для проверки однородности трех независимых выборок) Проверить на предмет однородности данные метрической переменной в зависимости от фактора: 1) по критерию Краскела-Уоллиса, 2) при помощи однофакторного дисперсионного анализа. Построить бокс-плот. Применить критерий Стьюдента для множественных сравнений с поправкой Бонферони и критерий Тьюки.
- Выполнить двухфакторный дисперсионный анализ для метрической переменной с двумя факторами (первый из своего варианта, второй произвольный). Сравнить результаты использования моделей с фиксированными и случайными эффектами.

Функции, которые понадобятся для вычислений.

```
suppressWarnings(library('lawstat'))
DescriptiveStat <- function(X, group)
{
  mm <- tapply(X, group, function(x)
    mean(x, na.rm = TRUE))
  mm
  Sd <- tapply(X, group, function(x)
    sd(x, na.rm = TRUE))
  Sd
  nn <- tapply(X, group, function(x)
    length(na.omit(x)))
```

```
  nn
  inqr <- tapply(X, group, function(x)
    IQR(x, na.rm = TRUE))
  inqr
  err <- Sd / sqrt(nn)
  err
  list(mm = mm,
       err = err,
       nn = nn,
       inqr = inqr)
}

Sentence <- function(mm, err, nn, inqr)
{
  A1 <-
    paste(paste(
      format(mm, digits = 3, nsmall = 2),
      format(err, digits = 2, nsmall = 2),
      sep = "±"
    ),
    nn, sep = "/")
  A1. <-
    paste("Средние в группах равны соответственно",
          paste(A1, collapse = ", "))

  A4. <- paste("интерквартильные размахи равны соответственно", inqr[1], "и", inqr[2])
  paste(c(A1., A4.), sep = "\n")
}
```

Прочитаем тестовые данные из файла:

```
data_big <- read.csv("./data_big.csv")
table(data_big$tremor.1)
```

```
##
##  0  1  2
##  1 21 12
```

```
df<-data.frame(group=ifelse(data_big$tremor.1==2,2,1),X= data_big$SBP.1)
```

Задача однородности в случае двух выборок

Посчитаем значения средних с ошибками среднего, медианы с интерквартильным размахом

```
L<-DescriptiveStat(df$X,df$group)

Sentence(L$mm,L$err,L$nn,L$inqr)
```

```
## [1] "Средние в группах равны соответственно 135.18±3.63/22, 153.50±2.72/12"
## [2] "интерквартильные размахи равны соответственно 19.5 и 15"
```

```
p.T <- t.test(X ~ group, df, var.equal = TRUE)$p.value
p.T
```

```
## [1] 0.001655216
```

Значение p-value 0.0016 меньше 0.05, различия между группами в выборке статистически значимы.

Проверим на предмет однородности независимых выборок метрическую переменную в зависимости от категориальной переменной по критериям Вилкоксона, Фишера равенства дисперий, Стьюдента равенства средних.

```
suppressWarnings(p.W <- wilcox.test(X ~ group, df, exact=TRUE)$p.value)
p.W
```

```
## [1] 0.001073001
```

```
p.F <- var.test(X ~ group, df)$p.value
p.F
```

```
## [1] 0.04715123
```

```
p.T <- t.test(X ~ group, df, var.equal = TRUE)$p.value
p.T
```

```
## [1] 0.001655216
```

По результатам вычисления всех трех критериев при уровне значимости 0.05 гипотеза об однородности независимых выборок отвергается.

Задача однородности в случае более двух выборок

Проверим на предмет однородности данные метрической переменной `SV.1` в зависимости от факторов `insomnia.1` и `tremor.1` по критерию Краскела-Уоллиса.

```
kruskal.test(SV.1 ~ insomia.1, data = data_big)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SV.1 by insomia.1
## Kruskal-Wallis chi-squared = 0.50425, df = 2, p-value = 0.7771
```

Значение p-value 0.77 больше 0.05, что означает, что на уровне значимости 0.05 гипотеза о том, что различные группы имеют одинаковые распределения, не может быть отвергнута.

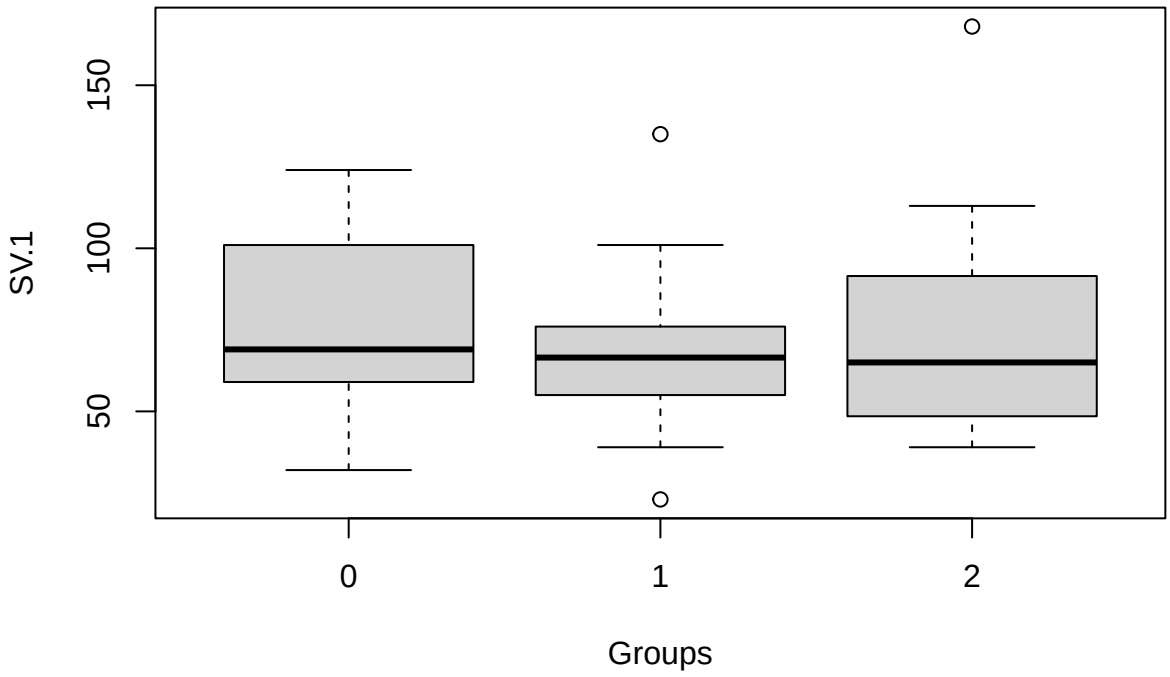
Тоже самое при помощи однофакторного дисперсионного анализа.

```
aov <- aov(SV.1 ~ insomia.1, data = data_big)
summary(aov)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## insomia.1    1      8     7.7   0.007  0.932
## Residuals   31  32179  1038.0
## 1 пропущенное наблюдение удалено
```

По результатам значение p-value 0.93 превышает уровень значимости 0.05, нулевая гипотеза о том, что между группами нет значимых различий, не может быть отвергнута.

```
boxplot(data_big$SV.1 ~ data_big$insomia.1,
        xlab = "Groups", ylab = "SV.1")
```



На boxplot отображается распределение переменной SV.1 для каждой группы insomia.1. Видно, что медианы значений в группах близки, при этом максимальные и минимальные значения, верхний и нижний квартили отличаются. Также в группе 1 и 2 есть выбросы - значения, выходящие за 1.5 IQR (интерквартильного интервала).

Применим критерий Фишера для множественных сравнений с поправкой Бонферони.

```
df <- data.frame(group = as.factor(data_big$insomia.1), X = as.numeric(data_big$SV.1))
table(df$group)
```

```
##
##  0  1  2
##  7 10 17
```

```
suppressMessages(suppressWarnings(library(multcomp)))
suppressMessages(suppressWarnings(library(agricolae)))

aov <- aov(X~group, df)
out <- LSD.test(aov,"group", p.adj="bonferroni",group=FALSE)
out
```

```
## $statistics
##    MSerror Df     Mean       CV
##   1061.515 30 73.72727 44.19111
##
## $parameters
##         test  p.ajusted name.t ntr alpha
##   Fisher-LSD bonferroni  group   3  0.05
##
## $means
##          X      std  r      LCL       UCL Min Max   Q25  Q50    Q75
## 0 77.85714 31.87177  7 52.70773 103.00655  32 124 59.00 69.0 101.00
## 1 69.20000 31.61856 10 48.15850  90.24150  23 135 55.00 66.5  75.75
## 2 74.75000 33.41956 16 58.11523  91.38477  39 168 48.75 65.0  90.25
##
## $comparison
##       difference pvalue signif.       LCL      UCL
## 0 - 1   8.657143      1         -32.05684 49.37112
## 0 - 2   3.107143      1         -34.33175 40.54604
## 1 - 2  -5.550000      1         -38.85388 27.75388
##
## $groups
## NULL
##
## attr(,"class")
## [1] "group"
```

Статистически значимой разницы между группами не обнаружено, p-value > 0.05.

Применим критерий Тьюки.

```
tukey <- TukeyHSD(aov)
print(tukey)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = X ~ group, data = df)
##
## $group
##          diff       lwr      upr     p adj
## 1-0 -8.657143 -48.23965 30.92536 0.8527161
## 2-0 -3.107143 -39.50558 33.29129 0.9759002
## 2-1  5.550000 -26.82834 37.92834 0.9065676
```

Для всех трех пар групп различия не являются статистически значимыми при уровне значимости 0.05.

Двухфакторный дисперсионный анализ

```
df <- data.frame(
  group1 = as.factor(data_big$tremor.1),
  group2 = as.factor(data_big$insomia.1),
  X = as.numeric(data_big$SV.1)
)
ao <- aov(X ~ group1 * group2, df)
SLM <- summary(ao)
SLM
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## group1         2    242   120.8   0.103  0.902
## group2         2    316   158.2   0.135  0.874
## group1:group2  2   1267   633.4   0.542  0.588
## Residuals     26  30362  1167.8
## 1 пропущенное наблюдение удалено
```

Все факторы и их взаимодействие не имеют статистически значимых различий с зависимой переменной, все p-value больше 0.05.

Посчитаем значимость для модели со случайными эффектами.

```
df_ <- SLM[[1]][,1]
y <- SLM[[1]][,3]

F1 <- y[1] / y[3]
p1 <- 1 - pf(F1, df_[1], df_[3])
F2 <- y[2] / y[3]
p2 <- 1 - pf(F2, df_[2], df_[3])

c(p1, p2)
```

```
## [1] 0.8398411 0.8001581
```

У модели со случайными эффектами эффектами значения p-value получились меньше, чем у модели с фиксированными эффектами, но в обоих случаях они больше 0.05, значимость не изменилась.