Методы статистической обработки информации (практика 4)

02.08.2019

Проверка гипотез однородности

Описание данных и задачи

В качестве тестового массива будем использовать кардиологические данные (метрические и категориальные) об ААС (алкогольно-абстинентном синдроме) у финских алкоголиков, измеренные в 1, 2, 3, 9 день поступления в стационар после длительного запоя. Варианты в табл.1. Переменная из $Metr$, группирующая из $Cath$. Последний вариант тестовый.

1. Проверить равенство дисперсий и средних для двух независимых выборок соответственно по критериям Фишера и Стьюдента.
2. Проверить равенство средних по всем трем группам одновременно при помощи однофакторного дисперсионного анализа и применить множественные сравнения: LSD с поправкой Бонферони, Тьюки HSD.
3. Проверить значимость изменения во времени метрической перемнной по разным парам временных точек при помощи критерия Стьюдента и по всем временным точкам при помощи ANOVA Repeated Measures.
4. Показать на примерах использование непараметрических критериев: знаков, рангового Вилкоксона, Фридмана.
5. Добавить категориальную перемнную depress.mod и применить двухфакторный дисперсионный анализ с фиксированными и случайными эффектами.

Чтение данных

```
library("knitr")
data_big <- read.csv("data_big.csv")
```

Таблица 1: Варианты сочетаний метрических переменных и факторов.

num	Cath	Metr	Descr
1	tremor.1	DBP.1	диаст.давл.
2	tremor.1	SBP.1	сист.давл.
3	sweating.1	CI.1	сердечный индекс
4	chest.pain.1	DBP.1	
5	thirst.1	SBP.1	
6	anoreksia.1	DBP.1	
7	craving.to.alcohol.1	HR.1	част.серд.сокр.

Параметрические критерии однородности для независимых выборок

Для группирующей переменной с двумя градациями применяются критерии Стьюдента и Фишера, а с числом градаций более двух — однофакторный дисперсионный анализ с критерием Бартлетта равенства дисперсий с множественными сравнениями по Тьюки.

```
#выбираем вариант
df<-data.frame(group=as.factor(data.c$craving.to.alcohol.1),X=as.numeric(data.m$HR.1))
#определяем число групп и количество наблюдений в каждой группе
table(df$group)
```

```
##
##  0  1  2
## 12 17  5
```

```
name.gr<-"craving.to.alcohol.1"
name.x<-"HR.1"

#критерий равенства дисперсий
bartlett.test(X~group,df)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  X by group
## Bartlett's K-squared = 0.71624, df = 2, p-value = 0.699
```

```
#однофакторный дисперсионный анализ
ao<-aov(X~group,df)
summary(ao)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        2   2589  1294.3    4.85 0.0147 *
## Residuals   31   8273   266.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# графическое представление о различии групп
boxplot(X~group,xlab=name.gr,ylab=name.x,data=df)
```

Видим, что при высокой потребности в алкоголе ЧСС выше. Значимость меньше 0.05, поэтому говорит о том, что это различие нельзя объяснить случайностью. Критерий равенства дисперсий свидетельствует о корректности применения данного метода. Применяем множественные сравнения.

```
library("agricolae") #пакеты для множественных сравнений
library("multcomp")
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```

```
ao<-aov(X~group,df)
#обычные LSD
out <- LSD.test(ao,"group", p.adj="none",group=FALSE)
out
```

```
## $statistics
##    MSerror Df     Mean       CV
##   266.8698 31 81.79412 19.97228
##
## $parameters
##         test p.ajusted name.t ntr alpha
##   Fisher-LSD      none  group   3  0.05
##
## $means
```

```
##          X      std  r      LCL       UCL Min Max  Q25 Q50   Q75
## 0 70.00000 13.85641 12 60.38198  79.61802  49  98 57.5  73  77.5
## 1 87.88235 17.42083 17 79.80160  95.96311  65 117 72.0  85 108.0
## 2 89.40000 18.06378  5 74.49983 104.30017  63 110 85.0  87 102.0
##
## $comparison
##       difference pvalue signif.       LCL       UCL
## 0 - 1 -17.882353 0.0067      ** -30.44439 -5.320311
## 0 - 2 -19.400000 0.0331       * -37.13475 -1.665247
## 1 - 2  -1.517647 0.8563         -18.46798 15.432684
##
## $groups
## NULL
##
## attr(,"class")
## [1] "group"
```

```
# LSD с поправками Бонферони
out <- LSD.test(ao,"group", p.adj="bonferroni",group=FALSE)
out
```

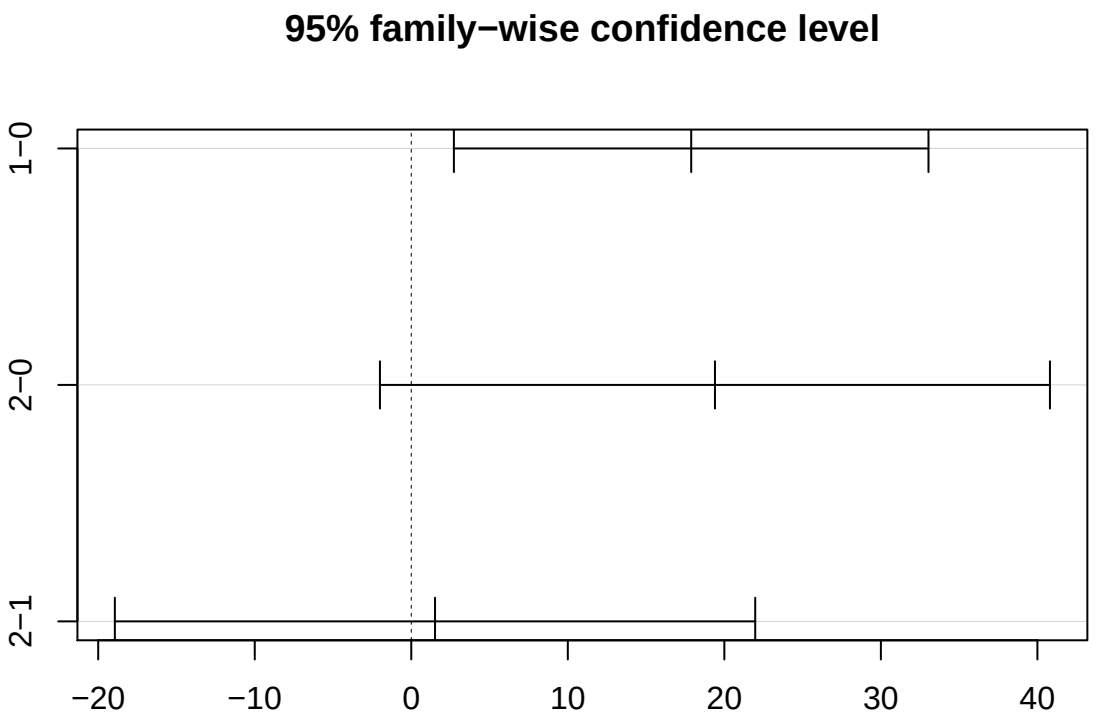
```
## $statistics
##    MSerror Df     Mean       CV
##   266.8698 31 81.79412 19.97228
##
## $parameters
##         test  p.ajusted name.t ntr alpha
##   Fisher-LSD bonferroni  group   3  0.05
##
## $means
##          X      std  r      LCL       UCL Min Max  Q25 Q50   Q75
## 0 70.00000 13.85641 12 60.38198  79.61802  49  98 57.5  73  77.5
## 1 87.88235 17.42083 17 79.80160  95.96311  65 117 72.0  85 108.0
## 2 89.40000 18.06378  5 74.49983 104.30017  63 110 85.0  87 102.0
##
## $comparison
##       difference pvalue signif.       LCL       UCL
## 0 - 1 -17.882353 0.0202       * -33.47117 -2.293539
## 0 - 2 -19.400000 0.0992       . -41.40787  2.607869
## 1 - 2  -1.517647 1.0000         -22.55209 19.516797
##
## $groups
## NULL
##
## attr(,"class")
## [1] "group"
```

```
#Тьюки
TukeyHSD(ao, "group", ordered = TRUE)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##     factor levels have been ordered
##
## Fit: aov(formula = X ~ group, data = df)
##
```

```
## $group
##          diff        lwr      upr     p adj
## 1-0 17.882353   2.723072 33.04163 0.0180070
## 2-0 19.400000  -2.001466 40.80147 0.0816153
## 2-1  1.517647 -18.937215 21.97251 0.9817956
```

```
plot(TukeyHSD(ao, "group"))
```



Наиболее значимое отличие по частоте сердечных сокращений между группой 0 - алкоголики без потребности опохмелиться - и группой 1 (с потребностью). Однако различие между группами 1 и 2 (с повышенной потребностью) незначимо. Поэтому есть смысл проверить значимость отличия группы 0 от 1и 2 вместе. Для этого строим сравнение с коэффициентами (-1, 1/2,1/2).

```
#сравнения
contr <- rbind(  "0 - 12" = c(-1, 1/2, 1/2),
                 "1 - 02" = c(1/2, -1, 1/2),
                 "2 - 01" = c(1/2, 1/2, -1))

GL<-glht(ao, linfct = mcp(group = contr))
summary(GL)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: aov(formula = X ~ group, data = df)
##
## Linear Hypotheses:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## 0 - 12 == 0   18.641      6.285   2.966   0.0147 *
## 1 - 02 == 0   -8.182      5.882  -1.391   0.3508
## 2 - 01 == 0  -10.459      7.928  -1.319   0.3882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Параметрические критерии Фишера и Стьюдента

Посмотрим, получим ли мы значимое различие между группами 0 и 1 по обычному двухвыборочному критерию.

```
#критерий Фишера для проверки равенства дисперсий
df.<-subset(df,df$group!=2)
S2<-(with(df.,tapply(X,group,'sd')))^2;S2
```

```
##        0        1        2
## 192.0000 303.4853       NA
```

```
nn<-with(df.,tapply(X,group,'length'));nn
```

```
##  0  1  2
## 12 17 NA
```

```
F.<-S2[2]/S2[1];p.F<-1-pf(F.,nn[2]-1,nn[1]-1);p.F
```

```
##         1
## 0.2229444
```

```
#Дисперсии можно считать одинаковыми

#критерий Стьдента для проверки равенства средних
t.test(X~group,subset(df,df$group!=2),var.equal=(p.F>0.05))
```

```
##
##  Two Sample t-test
##
## data:  X by group
## t = -2.9524, df = 27, p-value = 0.006455
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -30.310038  -5.454667
## sample estimates:
## mean in group 0 mean in group 1
##        70.00000        87.88235
```

Заметим, что статистики по критерию Стьюдента не совпадают с аналогичными статистиками из множественных сравнений, поскольку в последних используется общая дисперсия, а в обычном критерии дисперсия только по паре выборок.

Непараметрические критерии однородности для независимых выборок

Для проверки гипотезы однородности двух медиан используется критерий Вилкоксона (точный) или его приближенный вариант критерий Манна-Уитни. Для выборок более двух применяется критерий Краскела-Уоллиса или по медианный критерий.

```r
#df. без группы 2

wilcox.test(X~group,subset(df,df$group!=1),exact=TRUE)  # точный Вилкоксона
```

```
## Warning in wilcox.test.default(x = c(98, 56, 79, 64, 49, 58, 70, 80, 77, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  X by group
## W = 10, p-value = 0.03959
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(X~group,df.,exact=FALSE,correct=FALSE)  # Манна-Уитни
```

```
##
##  Wilcoxon rank sum test
##
## data:  X by group
## W = 41, p-value = 0.006855
## alternative hypothesis: true location shift is not equal to 0
```

```r
wilcox.test(X~group,df.,exact=FALSE)  # Манна-Уитни с поправкой
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  X by group
## W = 41, p-value = 0.007326
## alternative hypothesis: true location shift is not equal to 0
```

```r
kruskal.test(X~group,df) #Краскел-Уоллиса
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  X by group
## Kruskal-Wallis chi-squared = 8.5868, df = 2, p-value = 0.01366
```

```r
library(agricolae)

comparison<-with(df,kruskal(X,group, p.adj="bonferroni",group=FALSE, main="HR"))
comparison
```

```
## $statistics
##       Chisq Df     p.chisq
##   8.586762  2 0.01365867
##
## $parameters
##             test  p.ajusted name.t ntr alpha
##   Kruskal-Wallis bonferroni  group   3  0.05
##
## $means
##          X     rank      std  r Min Max  Q25 Q50   Q75
## 0 70.00000 10.75000 13.85641 12  49  98 57.5  73  77.5
## 1 87.88235 20.94118 17.42083 17  65 117 72.0  85 108.0
```

```
## 2 89.40000 22.00000 18.06378  5  63 110 85.0  87 102.0
##
## $comparison
##       Difference pvalue Signif.       LCL        UCL
## 0 - 1 -10.191176 0.0135        * -18.61314 -1.7692143
## 0 - 2 -11.250000 0.0686        . -23.13990  0.6399001
## 1 - 2  -1.058824 1.0000          -12.42282 10.3051771
##
## $groups
## NULL
##
## attr(,"class")
## [1] "group"
```

```
with(df,Median.test(X,group))
```

```
##
## The Median Test for X ~ group
##
## Chi Square = 11.52616   DF = 2   P.Value 0.003141425
## Median = 80
##
##   Median  r Min Max  Q25   Q75
## 0     73 12  49  98 57.5  77.5
## 1     85 17  65 117 72.0 108.0
## 2     87  5  63 110 85.0 102.0
##
## Post Hoc Analysis
##
## Groups according to probability of treatment differences and alpha level.
##
## Treatments with the same letter are not significantly different.
##
##    X groups
## 2 87      a
## 1 85      a
## 0 73      b
```

Двухфакторный дисперсионный анализ

```
df<-data.frame(group1=as.factor(data.c$craving.to.alcohol.1),
               group2=as.factor(data.c$depressed.mood.1 ),
               X=as.numeric(data.m$HR.1))

ao<-aov(X~group1*group2,df)
SLM<-summary(ao)
SLM
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## group1         2   2589  1294.3   4.754 0.0167 *
## group2         1     59    59.2   0.218 0.6445
## group1:group2  2    591   295.5   1.085 0.3516
## Residuals     28   7623   272.2
## ---
```

variant	indep.variable	grouping	Grouping
1	cravin	sex	educat
2	sstati	sex	educat
3	bdi	curwor	educat
4	gaf	ha	educat
5	rabdru	st	educat
6	rubsex	se	educat
7	cravin	curwor	prcod
8	sstati	st	prcod
9	bdi	sex	prcod
10	gaf	se	prcod
11	rabdru	ha	prcod
12	rubsex	st	prcod

Таблица 2: Соответствие заданий номерам вариантов в параметрических критериях для независимых выборок.

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Для моделей с хотя бы одним случайным эффектом значимость по эффекту взаимодействия остается той же, что и в модели с фиксированными эффектами, а значимость по группам считается по отношениям к сумме квадратов, связанной с эффектом взаимодействия.

```
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])
print("Значимости случайных эффектов")
```

```
## [1] "Значимости случайных эффектов"
```

```
c(p1,p2)
```

```
## [1] 0.1858691 0.6981474
```

Параметрические критерии однородности для двух зависимых выборок

В случае двух зависимых выборок применяется критерий Стьюдента, при нескольких дисперсионный анализ для расщепленных блоков (ANOVA Repeated Measures). В качестве данных рассматриваются данные по AAC. Варианты аналогичны заданиям по проверке однородности независимых выборок с отличием в том, что рассматривается несколько временных точек.

По-отдельности для сравнения двух зависимых выборок используем критерий Стьюдента

```
t.test(data_big$HR.1,data_big$HR.2,  paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  data_big$HR.1 and data_big$HR.2
## t = 2.389, df = 33, p-value = 0.02277
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.9863794 12.3077383
## sample estimates:
## mean of the differences
```

```
##                 6.647059
```

```
t.test(data_big$HR.2,data_big$HR.3,  paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  data_big$HR.2 and data_big$HR.3
## t = 0.4143, df = 21, p-value = 0.6829
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.750352  7.113989
## sample estimates:
## mean of the differences
##                1.181818
```

Непараметрические критерии однородности для двух и нескольких зависимых выборок

```
#критерий знаков
tab<-table(sign(data_big$HR.2-data_big$HR.1));tab
```

```
##
## -1  1
## 21 13
```

```
binom.test(min(tab), sum(tab), p = 0.5,
          alternative =  "less",
          conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  min(tab) and sum(tab)
## number of successes = 13, number of trials = 34, p-value = 0.1147
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.5377521
## sample estimates:
## probability of success
##              0.3823529
```

```
#ранговый Вилкоксона
wilcox.test(data_big$HR.2,data_big$HR.1,paired=TRUE,exact = FALSE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  data_big$HR.2 and data_big$HR.1
## V = 174, p-value = 0.03541
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(data_big$HR.2,data_big$HR.3,paired=TRUE,exact = FALSE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
```

```
##
## data:  data_big$HR.2 and data_big$HR.3
## V = 131, p-value = 0.6016
## alternative hypothesis: true location shift is not equal to 0
```

```
#Friedman
dat<-Form.dat(data_big,Name="HR")

friedman.test(as.matrix(dat[,c(1,ncol(dat))]))
```

```
##
##  Friedman rank sum test
##
## data:  as.matrix(dat[, c(1, ncol(dat))])
## Friedman chi-squared = 11.645, df = 1, p-value = 0.0006437
```

```
friedman.test(as.matrix(dat[,-1]))
```

```
##
##  Friedman rank sum test
##
## data:  as.matrix(dat[, -1])
## Friedman chi-squared = 1.8072, df = 2, p-value = 0.4051
```

Параметрические критерии однородности для нескольких зависимых выборок

Для исследования динамики показателя используем метод ANOVA Repeated Measures. Выбираем данные: группирующую переменную craving.to.alcohol.1 и наблюдения показателя ЧСС в три момента времени HR.1,HR.2,HR.3.

```
dat.AR<-na.omit(subset(data_big,select=c(craving.to.alcohol.1,HR.1,HR.2,HR.3)))

#преобразуем данные к виду, необходимому для стандатной обработки

  m<-ncol(dat.AR)-1;m # число временных точек
```

```
## [1] 3
```

```
  Names<-names(table(dat.AR[,1])); Names; K<-length(Names)
```

```
## [1] "0" "1" "2"
```

```
  dat.AR.T<-data.frame( stack(dat.AR[,-1]),
                        sub=as.factor(rep(seq(nrow(dat.AR)),m)),
                        gr=as.factor(rep(dat.AR[,1],m)))
 dat.AR.T$values<-as.numeric(dat.AR.T$values)


 #   применяем ANOVA Repeated Measures

 formula<-values~gr*ind+Error(sub/ind)

 aov.out <- aov(formula, data=dat.AR.T)
 #выводим средние
 model.tables(aov.out,'mean')
```

```
## Tables of means
## Grand mean
##
## 79.92424
##
##  gr
##         0     1     2
##     65.92 84.92 78.13
## rep 12.00 39.00 15.00
##
##  ind
##     HR.1  HR.2  HR.3
##     85.5 77.73 76.55
## rep 22.0 22.00 22.00
##
##  gr:ind
##       ind
## gr     HR.1  HR.2  HR.3
##   0   61.25 67.25 69.25
##   rep  4.00  4.00  4.00
##   1   91.46 83.69 79.62
##   rep 13.00 13.00 13.00
##   2   89.40 70.60 74.40
##   rep  5.00  5.00  5.00
```

```
#таблица дисперсионного анализа
aov.S<-summary(aov.out);
aov.S
```

```
##
## Error: sub
##           Df Sum Sq Mean Sq F value Pr(>F)
## gr         2   3377  1688.6   3.503 0.0507 .
## Residuals 19   9159   482.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Error: sub:ind
##           Df Sum Sq Mean Sq F value Pr(>F)
## ind        2   1041   520.7   4.649 0.0156 *
## gr:ind     4   1027   256.8   2.293 0.0772 .
## Residuals 38   4256   112.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#графики

  interaction.plot(x.factor=dat.AR.T$ind,
                   trace.factor=dat.AR.T$gr,
                   response=dat.AR.T$values,
                   fun = mean,
                   type =  "b", legend = FALSE,
                   trace.label ="group",
                   xlab = "",
                   ylab = 'HR',
```

```
            lty = seq(K), col = seq(K), pch = 20, lwd=2
)
legend('topright',Names,lty = seq(K), col =seq(K), cex=0.7,pch=20)
```