第二次作业

> 此作业源代码丢失，为重写代码/作业，不保证正确率，仅供参考。
>
> 这次作业我的评价是：纯属折磨人。

1 第一题

> 描述全部及分性别研究对象的年龄、教育程度、是否吸烟、是否饮酒、肥胖状况（分三组：正常/超重/肥胖，界值24/28）、收缩压水平、舒张压水平
>
> 要求：绘制统计表，对统计表内容进行必要文字描述，核心步骤简要说明

	女 (N=1271)	男 (N=1135)	总计 (N=2406)
年龄 (岁)			
[15,30)	300 (23.6%)	216 (19.0%)	516 (21.4%)
[30,45)	291 (22.9%)	264 (23.3%)	555 (23.1%)
[45,60)	279 (22.0%)	276 (24.3%)	555 (23.1%)
[60,75)	285 (22.4%)	259 (22.8%)	544 (22.6%)
[75,90]	116 (9.1%)	120 (10.6%)	236 (9.8%)
教育程度			
未上学	239 (18.8%)	243 (21.4%)	482 (20.0%)
小学	172 (13.5%)	181 (15.9%)	353 (14.7%)
中学	422 (33.2%)	364 (32.1%)	786 (32.7%)
大学及以上	424 (33.4%)	336 (29.6%)	760 (31.6%)
未知	14 (1.1%)	11 (1.0%)	25 (1.0%)
吸烟情况			
从不吸烟	684 (53.8%)	682 (60.1%)	1366 (56.8%)
过去吸烟	522 (41.1%)	401 (35.3%)	923 (38.4%)
现在吸烟	65 (5.1%)	52 (4.6%)	117 (4.9%)
饮酒情况			
从不饮酒	792 (62.3%)	721 (63.5%)	1513 (62.9%)
过去饮酒	164 (12.9%)	146 (12.9%)	310 (12.9%)
现在饮酒	315 (24.8%)	268 (23.6%)	583 (24.2%)
BMI指数			
[14,24)	830 (65.3%)	404 (35.6%)	1234 (51.3%)
[24,28)	344 (27.1%)	261 (23.0%)	605 (25.1%)
[28,40]	97 (7.6%)	470 (41.4%)	567 (23.6%)
收缩压 (mmHg)			
[70,90)	116 (9.1%)	71 (6.3%)	187 (7.8%)
[90,110)	484 (38.1%)	435 (38.3%)	919 (38.2%)

[110,130)	298 (23.4%)	278 (24.5%)	576 (23.9%)
[130,150)	302 (23.8%)	252 (22.2%)	554 (23.0%)
[150,170]	71 (5.6%)	99 (8.7%)	170 (7.1%)
舒张压 (mmHg)			
[50,60)	231 (18.2%)	209 (18.4%)	440 (18.3%)
[60,70)	366 (28.8%)	326 (28.7%)	692 (28.8%)
[70,80)	233 (18.3%)	200 (17.6%)	433 (18.0%)
[80,90)	210 (16.5%)	190 (16.7%)	400 (16.6%)
[90,100)	178 (14.0%)	149 (13.1%)	327 (13.6%)
[100,110]	53 (4.2%)	61 (5.4%)	114 (4.7%)

2 第二题、第三题

> 报告全部及分性别研究对象的高血压患病率(粗率及年龄标化率，利用2010年全国第六次人口普查数据作为标准人口)
>
> 报告全部及分性别研究对象的不同年龄高血压患病率 要求:绘制统计图，对统计图内容进行必要文字描述，核心步骤简要说明

	女		总计												
	[15,30)	[30,45)	[45,60)	[60,75)	[75,90]	[15,30)	[30,45)	[45,60)	[60,75)	[75,90]	[15,30)	[30,45)	[45,60)	[60,75)	[75,90]
	(N=300)	(N=291)	(N=279)	(N=285)	(N=116)	(N=216)	(N=264)	(N=276)	(N=259)	(N=120)	(N=516)	(N=555)	(N=555)	(N=544)	(N=236)
is_high_bp															
否	234	207	177	179	72	169	181	175	164	68	403	388	352	343	140
	(78.0%)	(71.1%)	(63.4%)	(62.8%)	(62.1%)	(78.2%)	(68.6%)	(63.4%)	(63.3%)	(56.7%)	(78.1%)	(69.9%)	(63.4%)	(63.1%)	(59.3%)
是	66	84	102	106	44	47	83	101	95	52	113	167	203	201	96
	(22.0%)	(28.9%)	(36.6%)	(37.2%)	(37.9%)	(21.8%)	(31.4%)	(36.6%)	(36.7%)	(43.3%)	(21.9%)	(30.1%)	(36.6%)	(36.9%)	(40.7%)

	女 (N=1271)	男 (N=1135)	总计 (N=2406)
是否患有高血压（粗率）			
否	869 (68.4%)	757 (66.7%)	1626 (67.6%)
是	402 (31.6%)	378 (33.3%)	780 (32.4%)

标准化后:

	男	女	总计
15-29	36283661	35544309	71898546
30-44	54554141	48031986	102281670
45-59	49483629	47686867	97169406
60-74	24570283	24460564	49050267
75-89	8405007	8899033	17433461

标准化率表:

	男	女	总计
15-29	0.22	0.22	0.22
30-44	0.31	0.29	0.3
45-59	0.37	0.37	0.37
60-74	0.37	0.37	0.37
75-89	0.43	0.38	0.41

不分年龄：

男	女	总计
173296721	164622759	337833350

不分年龄标准化率表：

男	女	总计
0.31	0.30	0.30

3 附录

以下附上源代码。第二题和第三题的实现过于复杂，建议自己寻找其他办法。

```r
# -*- coding: utf-8 -*-
# @Author  :Arthals
# @File    :Homework2.r
# @Time    :2023/01/25 18:48:48
# @Software:   Visual Studio Code

rm(list = ls())

# 第一题
rawdata <- read.csv(
    "课件&作业/2作业-分类变量-标准化率/cleandata.csv",
    header = TRUE, stringsAsFactors = FALSE, na.strings = c("", "NA")
) # 读取数据

dim(rawdata) # 查看数据集的维度
names(rawdata)
summary(rawdata)


# 生成年龄分组频数表
rawdata$agegrp <- cut(
    rawdata$age, c(seq(15, 90, 15)),
    include.lowest = TRUE, right = FALSE
```

```r
)
# 这个函数的作用是给原数据新加了一列分类变量agegrp，这一列的值是根据原数据的age列的值来划分的

# 生成教育程度分组频数表
rawdata$edu ← factor(
    rawdata$edu,
    levels = c("未上学", "小学", "中学", "大学及以上", "未知")
)

# 生成吸烟情况分组频数表
rawdata$smk ← factor(
    rawdata$smk,
    levels = c("从不吸烟", "过去吸烟", "现在吸烟")
)

# 生成饮酒情况分组频数表
rawdata$dnk ← factor(
    rawdata$dnk,
    levels = c("从不饮酒", "过去饮酒", "现在饮酒")
)

# 生成BMI指数变量bmi
rawdata$bmi ← rawdata$weight * 10000 / (rawdata$height^2) # 计算BMI指数
# 生成BMI指数频数表
rawdata$bmigrp ← cut(
    rawdata$bmi, c(14, 24, 28, 40),
    include.lowest = TRUE,
    right = FALSE
)

# 生成收缩压分组频数表
rawdata$sbpgrp ← cut(
    rawdata$sbp, c(seq(70, 170, 20)),
    include.lowest = TRUE,
    right = FALSE
)

# 生成舒张压分组频数表
rawdata$dbpgrp ← cut(
    rawdata$dbp, c(seq(50, 110, 10)),
    include.lowest = TRUE,
    right = FALSE
)

# 设置标签、单位
```

```r
label(rawdata$agegrp) ← "年龄"
label(rawdata$edu) ← "教育程度"
label(rawdata$smk) ← "吸烟情况"
label(rawdata$dnk) ← "饮酒情况"
label(rawdata$bmigrp) ← "BMI指数"
label(rawdata$sbpgrp) ← "收缩压"
label(rawdata$dbpgrp) ← "舒张压"
units(rawdata$agegrp) ← "岁"
units(rawdata$sbpgrp) ← "mmHg"
units(rawdata$dbpgrp) ← "mmHg"

# 生成table1
library(table1)
table_one ← table1(~ agegrp + edu + smk + dnk + bmigrp + sbpgrp + dbpgrp |
    sex, data = rawdata, overall = "总计")
table_one

# 第二题
rm(list = ls())
# 报告研究对象的高血压患病率粗率
rawdata ← read.csv(
    "课件&作业/2作业-分类变量-标准化率/cleandata.csv",
    header = TRUE, stringsAsFactors = FALSE, na.strings = c("", "NA")
) # 读取数据
rawdata$agegrp ← cut(
    rawdata$age, c(seq(15, 90, 15)),
    include.lowest = TRUE, right = FALSE
)
rawdata$is_high_bp ← factor(
    ifelse(rawdata$sbp ⩾ 140 | rawdata$dbp ⩾ 90, 1, 0),
    levels = c(0, 1),
    labels = c("否", "是")
)
label(rawdata$is_high_bp) ← "是否患有高血压（粗率）"
is_high_bp_table ← table1(~ is_high_bp | sex, data = rawdata, overall = "总计")
is_high_bp_table

# 报告研究对象的高血压患病率标准化率
standard_population_total ← c(
    328315484, 339918126, 265660198, 132752961, 42857259
)

standard_population_male ← c(
    166750441, 173521604, 135222590, 66986350, 19396171
)
```

```r
standard_population_female <- c(
    161565043, 166396522, 130437608, 65766611, 23461088
)

rawdata$sex <- as.factor(rawdata$sex)
names(rawdata$sex) <- c("男", "女")


# 按照年龄进行标化
male_agegrp <- c()
for (age in levels(rawdata$agegrp)) {
    male_agegrp <- c(
        male_agegrp,
        nrow(
            rawdata[
                rawdata$agegrp == age &
                    rawdata$sex == "男" &
                    rawdata$is_high_bp == "是",
            ]
        ) / nrow(
            rawdata[
                rawdata$agegrp == age &
                    rawdata$sex == "男",
            ]
        )
    )
}
male_agegrp
illed_male <- male_agegrp * standard_population_male
illed_male

female_agegrp <- c()
for (age in levels(rawdata$agegrp)) {
    female_agegrp <- c(
        female_agegrp,
        nrow(
            rawdata[
                rawdata$agegrp == age &
                    rawdata$sex == "女" &
                    rawdata$is_high_bp == "是",
            ]
        ) / nrow(
            rawdata[
                rawdata$agegrp == age &
```

```
                    rawdata$sex == "女",
                ]
            )
        )
}
female_agegrp
illed_female <- female_agegrp * standard_population_female
illed_female

total_agegrp <- c()
for (age in levels(rawdata$agegrp)) {
    total_agegrp <- c(
        total_agegrp,
        nrow(
            rawdata[
                rawdata$agegrp == age &
                    rawdata$is_high_bp == "是",
            ]
        ) / nrow(
            rawdata[
                rawdata$agegrp == age,
            ]
        )
    )
}
total_agegrp
illed_total <- total_agegrp * standard_population_total
illed_total

# round


# 生成标准化率表
sd_table <- rbind(illed_male, illed_female, illed_total)
header <- c("15-29", "30-44", "45-59", "60-74", "75-89")
rownames(sd_table) <- c("男", "女", "总计")
colnames(sd_table) <- header
sd_table

# round the table
sd_table <- round(sd_table)


sd_table <- t(sd_table)
# save as csv
```

```r
write.csv(sd_table, "课件&作业/2作业-分类变量-标准化率/标准化率.csv")

total_table ← cbind(standard_population_male, standard_population_female,
standard_population_total)
colnames(total_table) ← c("男", "女", "总计")
rownames(total_table) ← header

rate_table ← round(sd_table / total_table, 2)
rate_table
write.csv(rate_table, "课件&作业/2作业-分类变量-标准化率/标准化率比.csv")


# sum rows in sd_table
sd_table ← apply(sd_table, 2, sum)
sd_table
total_table ← apply(total_table, 2, sum)
total_table
rate_table ← sd_table / total_table
rate_table
round(rate_table, 2)

# 第三题
options(warn = 0)
table1(~ is_high_bp | sex * agegrp, data = rawdata, overall = "总计")
options(warn = 1)
```