Code ▾

Fisher’s Exact Test of Breast Cancer Recurrence Data

1. Load preprocessed breast cancer dataset

Hide

```
breast_data = read.csv("breast_cancer_new.csv", row.names = 1)
```

Hide

```
# check data structure
str(breast_data)
```

```
'data.frame':   277 obs. of  10 variables:
 $ class         : chr  "no-recurrence-events" "no-recurrence-events" "no-recurrence-events" "no
-recurrence-events" ...
 $ age           : chr  "30-39" "40-49" "40-49" "60-69" ...
 $ menopause     : chr  "premeno" "premeno" "premeno" "ge40" ...
 $ tumor.size    : chr  "30-34" "20-24" "20-24" "15-19" ...
 $ inv.nodes     : chr  "0-2" "0-2" "0-2" "0-2" ...
 $ node.caps     : chr  "no" "no" "no" "no" ...
 $ deg.malignancy: int  3 2 2 2 2 2 2 1 2 2 ...
 $ breast        : chr  "left" "right" "left" "right" ...
 $ breast.quad   : chr  "left_low" "right_up" "left_low" "left_up" ...
 $ irradiation   : chr  "no" "no" "no" "no" ...
```

Hide

```
# factorize and relevel values of some variables
breast_data$class = factor(breast_data$class)
breast_data$menopause = factor(breast_data$menopause)
breast_data$node.caps = factor(breast_data$node.caps)
breast_data$breast = factor(breast_data$breast)
breast_data$breast.quad = factor(breast_data$breast.quad)
breast_data$irradiation = factor(breast_data$irradiation)
breast_data$age = factor(breast_data$age,
                         levels = c("20-29", "30-39" ,"40-49","50-59" ,"60-69",
                                    "70-79"))
breast_data$inv.nodes = factor(breast_data$inv.nodes,
                         levels = c("0-2", "3-5" ,"6-8","9-11" ,"12-14",
                                    "15-17","24-26"))
breast_data$deg.malignancy = factor(breast_data$deg.malignancy,
                                    levels = c("1","2","3"))
breast_data$tumor.size = factor(breast_data$tumor.size,
                                           levels = c("0-4","5-9", "10-14","15-19", "20-24","25-
29" ,"30-34" ,"35-39" ,"40-44" ,"45-49","50-54" ))
```

Hide

```
# check data structure after factoring and releving
str(breast_data)
```

```
'data.frame':	277 obs. of  10 variables:
 $ class         : Factor w/ 2 levels "no-recurrence-events",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ age           : Factor w/ 6 levels "20-29","30-39",..: 2 3 3 5 3 5 4 5 3 3 ...
 $ menopause     : Factor w/ 3 levels "ge40","lt40",..: 3 3 3 1 3 1 3 1 3 3 ...
 $ tumor.size    : Factor w/ 11 levels "0-4","5-9","10-14",..: 7 5 5 4 1 4 6 5 11 5 ...
 $ inv.nodes     : Factor w/ 7 levels "0-2","3-5","6-8",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ node.caps     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ deg.malignancy: Factor w/ 3 levels "1","2","3": 3 2 2 2 2 2 2 1 2 2 ...
 $ breast        : Factor w/ 2 levels "left","right": 1 2 1 2 2 1 1 1 1 2 ...
 $ breast.quad   : Factor w/ 5 levels "central","left_low",..: 2 5 2 3 4 2 2 2 2 3 ...
 $ irradiation   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

2. Fisher’s Exact Test

Hide

```
# get variables
variables = colnames(breast_data)
variables
```

```
 [1] "class"          "age"            "menopause"      "tumor.size"    
 [5] "inv.nodes"      "node.caps"      "deg.malignancy" "breast"        
 [9] "breast.quad"    "irradiation"   
```

Hide

```
# create a counts table of each variable vs the class variable
counts_df = data.frame()
for (i in 2:length(variables))
{
  df1 =  rep(c(variables[i]), times = length(unique(breast_data[[i]]))) %>%
  cbind(as.data.frame(table(breast_data[[i]], breast_data$class))) %>%
  pivot_wider(., names_from = Var2, values_from = Freq)

  counts_df = rbind(counts_df,df1)
}
```

Hide

```
# check column names of counts_df
colnames(counts_df)
```

```
[1] "."                    "Var1"                 "no-recurrence-events"
[4] "recurrence-events"   
```

Hide

```
# rename column names
# rename columns
colnames(counts_df) = c("variable", "category", "no-recurrence-events",
                        "recurrence-events")
```

Hide

```
# view the first ten rows of counts_df
head(counts_df,n=10)
```

variable	category	no-recurrence-events	recurrence-events
<chr>	<fctr>	<int>	<int>
age	20-29	1	0
age	30-39	21	15
age	40-49	62	27
age	50-59	69	22
age	60-69	38	17
age	70-79	5	0
menopause	ge40	90	33
menopause	lt40	5	0
menopause	premeno	101	48
tumor.size	0-4	7	1

1-10 of 10 rows

Hide

```
# perform fisher's exact test for each variable vs class variable in a loop
# save the p values for each variable in vectors
set.seed(12)

p_value = numeric(length(variables)-1)

for (i in 2:length(variables))
{
  fisher_result = suppressWarnings(fisher.test(table(breast_data[[i]],
                                                  breast_data$class),
                                              simulate.p.value = TRUE, B = 10000))
  p_value[i-1] = fisher_result$p.value[[1]]
}
```

Hide

```
# create a dataframe to visualize the p values for each variable
fisher_data = data.frame(variables[-1], p_value)
colnames(fisher_data) =  c("variable","p_value")
fisher_data
```

variable <chr>	**p_value** <dbl>
age	2.869713e-01
menopause	2.404760e-01
tumor.size	1.649835e-02
inv.nodes	9.999000e-05
node.caps	5.004316e-06
deg.malignancy	9.999000e-05
breast	5.111037e-01
breast.quad	5.316468e-01
irradiation	4.045491e-04

9 rows

Hide

```
# print statistically significant associated variables
fisher_data[fisher_data$p_value<0.05,]
```

	variable <chr>	**p_value** <dbl>
3	tumor.size	1.649835e-02
4	inv.nodes	9.999000e-05
5	node.caps	5.004316e-06
6	deg.malignancy	9.999000e-05
9	irradiation	4.045491e-04

5 rows

Hide

```
# merge counts and fisher test results
countsWithFisher = merge(counts_df, fisher_data, by = "variable")
```

Hide

```
head(countsWithFisher, n = 10)
```

	variable <chr>	category <fctr>	no-recurrence-events <int>	recurrence-events <int>	p_value <dbl>
1	age	20-29	1	0	0.2869713
2	age	30-39	21	15	0.2869713
3	age	40-49	62	27	0.2869713
4	age	50-59	69	22	0.2869713
5	age	60-69	38	17	0.2869713
6	age	70-79	5	0	0.2869713
7	breast	left	100	45	0.5111037
8	breast	right	96	36	0.5111037
9	breast.quad	central	17	4	0.5316468
10	breast.quad	left_low	73	33	0.5316468

1-10 of 10 rows