随机模拟方法与应用导论作业三

陈稼霖 *45875852*

2019-10-01

3.5 (Relating age and wage in the twins dataset)

The variables `AGE` and `HRWAGEL` contain the age (in years) and hourly wage (in dollars) of twin 1.

a. Using two applications of the `cut` function, create a categorized version of `AGE` using the breakpoints 30, 40, and 50, and a categorized version of `HRWAGEL` using the same breakpoints as in Section 3.3.

b. Using the categorized versions of AGE and HRWAGEL, construct a contingency table of the two variables using the function `table`.

c. Use the `prop.table` function to find the proportions of twins in each age class that have the different wage groups.

d. Construct a suitable graph to show how the wage distribution depends on the age of the twin.

e. Use the conditional proportions in part (c) and the graph in part (d) to explain the relationship between age and wage of the twins.

a. 首先读取数据文件`twins.dat.txt`，用`cut`函数根据断点30，40和50分割变量`AGE`并展示结果

```
twn = read.table('twins.dat.txt',header = TRUE,sep = ',',na.strings = '.')
c.age = cut(twn$AGE,breaks = c(0,30,40,50,80))
c.age
```

```
##   [1] (30,40] (50,80] (40,50] (30,40] (30,40] (50,80] (30,40] (50,80]
##   [9] (0,30]  (40,50] (50,80] (30,40] (30,40] (40,50] (30,40] (30,40]
##  [17] (0,30]  (0,30]  (30,40] (30,40] (0,30]  (0,30]  (30,40] (0,30]
##  [25] (0,30]  (30,40] (30,40] (0,30]  (40,50] (40,50] (50,80] (0,30]
##  [33] (50,80] (30,40] (50,80] (40,50] (30,40] (0,30]  (30,40] (50,80]
##  [41] (0,30]  (0,30]  (30,40] (40,50] (0,30]  (40,50] (50,80] (30,40]
##  [49] (50,80] (30,40] (50,80] (0,30]  (30,40] (50,80] (50,80] (40,50]
##  [57] (0,30]  (0,30]  (0,30]  (0,30]  (50,80] (50,80] (40,50] (30,40]
##  [65] (0,30]  (30,40] (50,80] (0,30]  (30,40] (30,40] (30,40] (30,40]
##  [73] (30,40] (30,40] (30,40] (0,30]  (30,40] (30,40] (40,50] (30,40]
##  [81] (40,50] (30,40] (0,30]  (40,50] (50,80] (0,30]  (50,80] (40,50]
##  [89] (0,30]  (50,80] (50,80] (50,80] (40,50] (30,40] (50,80] (0,30]
##  [97] (30,40] (50,80] (40,50] (0,30]  (50,80] (0,30]  (50,80] (40,50]
```

```
## [105] (30,40] (40,50] (30,40] (0,30]  (30,40] (30,40] (30,40] (30,40]
## [113] (50,80] (30,40] (0,30]  (40,50] (30,40] (0,30]  (0,30]  (40,50]
## [121] (50,80] (50,80] (30,40] (30,40] (30,40] (30,40] (30,40] (40,50]
## [129] (40,50] (30,40] (30,40] (0,30]  (0,30]  (0,30]  (40,50] (30,40]
## [137] (0,30]  (30,40] (0,30]  (40,50] (40,50] (50,80] (40,50] (0,30]
## [145] (30,40] (30,40] (0,30]  (50,80] (50,80] (50,80] (50,80] (0,30]
## [153] (50,80] (0,30]  (30,40] (0,30]  (0,30]  (50,80] (40,50] (30,40]
## [161] (40,50] (30,40] (50,80] (0,30]  (30,40] (50,80] (30,40] (0,30]
## [169] (0,30]  (30,40] (0,30]  (30,40] (40,50] (0,30]  (30,40] (40,50]
## [177] (0,30]  (30,40] (40,50] (0,30]  (40,50] (0,30]  (0,30]
## Levels: (0,30] (30,40] (40,50] (50,80]
```

然后用cut函数根据3.3节中的断点———0,7,13,20,150———分割变量HRWAGEL并展示结果

```
c.wage = cut(twn$HRWAGEL,breaks = c(0,7,13,20,150))
c.wage
```

```
##   [1] (7,13]   (7,13]   (7,13]   (13,20]  (13,20]  <NA>     (7,13]
##   [8] <NA>     (13,20]  (7,13]   (13,20]  (7,13]   (20,150] (20,150]
##  [15] (7,13]   <NA>     (0,7]    <NA>     (7,13]   (7,13]   <NA>
##  [22] (13,20]  (13,20]  (7,13]   <NA>     <NA>     (13,20]  (7,13]
##  [29] (0,7]    (20,150] (0,7]    (0,7]    <NA>     (7,13]   (7,13]
##  [36] (20,150] (20,150] (0,7]    (7,13]   (13,20]  (13,20]  (0,7]
##  [43] (13,20]  (20,150] (0,7]    (20,150] (20,150] (7,13]   (7,13]
##  [50] (0,7]    (20,150] (7,13]   (7,13]   (7,13]   (13,20]  (20,150]
##  [57] (7,13]   (0,7]    (0,7]    (7,13]   (13,20]  (7,13]   (13,20]
##  [64] (13,20]  (0,7]    (7,13]   <NA>     (0,7]    (20,150] (0,7]
##  [71] (7,13]   (0,7]    (13,20]  (20,150] (0,7]    <NA>     (13,20]
##  [78] (0,7]    <NA>     (0,7]    (0,7]    (13,20]  (7,13]   (7,13]
##  [85] (7,13]   (7,13]   <NA>     (13,20]  (7,13]   (13,20]  <NA>
##  [92] <NA>     (7,13]   (7,13]   (13,20]  (0,7]    (7,13]   (7,13]
##  [99] (20,150] (0,7]    (7,13]   (7,13]   (7,13]   (13,20]  (7,13]
## [106] (20,150] (7,13]   <NA>     (7,13]   (0,7]    (0,7]    (0,7]
## [113] <NA>     (13,20]  (0,7]    (20,150] (13,20]  (0,7]    <NA>
## [120] (0,7]    <NA>     (20,150] (0,7]    (7,13]   (7,13]   (7,13]
## [127] (7,13]   (13,20]  (7,13]   (7,13]   (20,150] (0,7]    (7,13]
## [134] (7,13]   (20,150] (13,20]  (0,7]    (7,13]   (0,7]    (7,13]
## [141] (7,13]   (0,7]    (0,7]    (7,13]   (0,7]    (0,7]    (7,13]
## [148] (0,7]    (13,20]  <NA>     (0,7]    (7,13]   (0,7]    (13,20]
## [155] (0,7]    (7,13]   (0,7]    (0,7]    (13,20]  <NA>     (0,7]
## [162] (7,13]   (0,7]    (13,20]  (13,20]  <NA>     (13,20]  (0,7]
```

```
## [169] (7,13]   (7,13]   (13,20]  (13,20]  (0,7]    (0,7]    (7,13]
## [176] (13,20]  (0,7]    (20,150] (13,20]  (13,20]  (0,7]    (13,20]
## [183] (13,20]
## Levels: (0,7] (7,13] (13,20] (20,150]
```

b. 用table函数构建年龄和小时工资的相依表并展示

```
T = table(c.age,c.wage)
T
```

```
##          c.wage
## c.age     (0,7] (7,13] (13,20] (20,150]
##   (0,30]     20     16       9        0
##   (30,40]    13     26      15        6
##   (40,50]     7      7       7       10
##   (50,80]     7      9       7        3
```

c. 用prop.table函数计算各年龄层次的工资分布情况

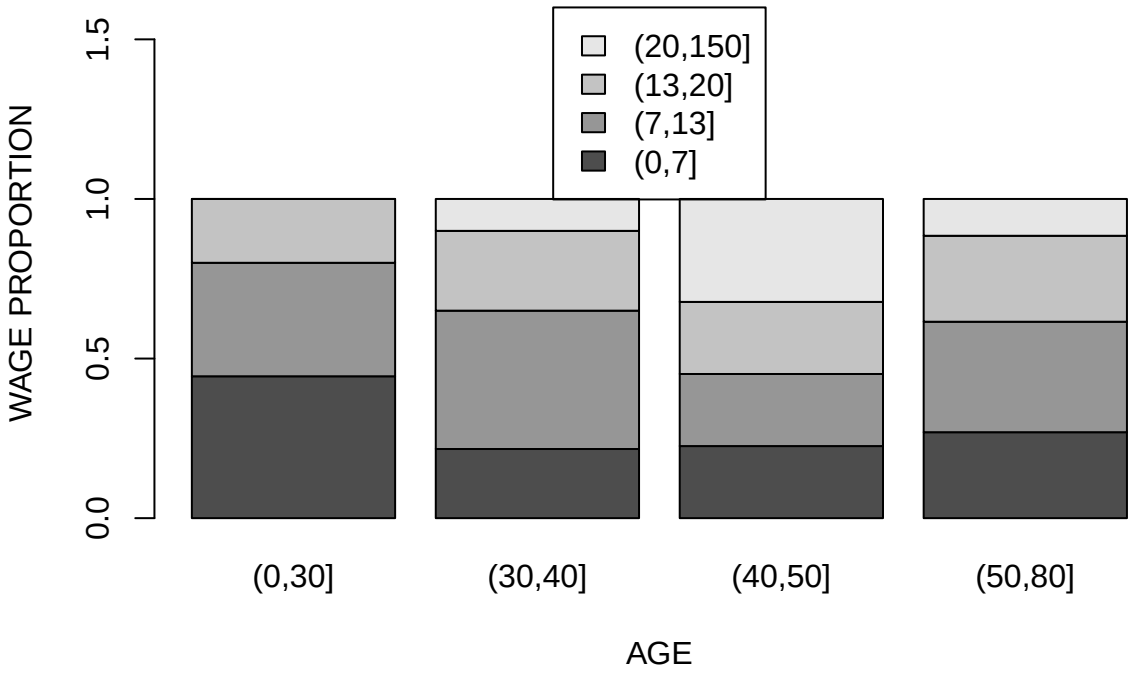
```
P = prop.table(T,margin = 1)
P
```

```
##          c.wage
## c.age         (0,7]    (7,13]   (13,20]  (20,150]
##   (0,30]  0.4444444 0.3555556 0.2000000 0.0000000
##   (30,40] 0.2166667 0.4333333 0.2500000 0.1000000
##   (40,50] 0.2258065 0.2258065 0.2258065 0.3225806
##   (50,80] 0.2692308 0.3461538 0.2692308 0.1153846
```

由此可见，年龄处于0至30岁的年轻人小时工资分布集中在$(0,7]$美元的区间，随着年龄的增长，工资会先有一定程度的提高，如30岁至40岁年龄段的工资分布最多的区间为$(7,13]$美元，40岁至50岁为$(13,20]$美元，但当年龄进一步增长时，工资开始下降，年龄处于50至80岁的老人工资分布最多的区间为$(7,13]$美元。

d. 用barplot函数绘制工资分布随年龄变化的分段条形图

```
barplot(t(P),ylim = c(0,1.6),ylab = 'WAGE PROPORTION',xlab = 'AGE'
        ,legend.text = dimnames(P)$c.wage,args.legend = list(x = 'top'))
```



e. 解释：由c中的条件比例和d中的分段条形图可见，双胞胎的工资随着年龄的增长先有一定程度的提高，这可能是由于随着学历的提高和工龄的增长，双胞胎的工作能力提高，随着年龄的进步增长，工资后下降，这可能是双胞胎由于年老，工作能力下降。