GALLAGHER-BIOS04285\_HW1

Ryan Gallagher

2023-01-29

**NOTE TO PROF**:  
This data set only has the data for the Participants column, and only some the variables of interest. And at that, there is still inconsistency with the numbers I find and the ones reported. I think the simplest and most logical conclusion is that this isnt the FINAL dataset used for this study. With that being said, I’ve created some of the rows found in Table 1 in attempt to meet the expectations of this assignment. If it’s any consolation too, the analysis isn’t exceptionally different per the variable, since they’re categorized to a range and then we just calculate frequency statistics. -Ryan  
———————————————–  
Import the data, set x to be the number of variables, and see the variables available.

dat = read.csv("/Users/ryangallagher/Desktop/HW1Dataset.csv", header=TRUE)  
ls(dat)

## [1] "age" "bcbs" "death" "erpr" "grade" "hospvol"   
## [7] "nci" "node" "size" "stage2000" "surgvol" "t13"   
## [13] "t15" "t9" "time"

x=nrow(dat)

**MAKING THE “AGE GROUP, YEARS” FIGURE FROM TABLE 1**

dat$range=1  
for (i in 1:x) {  
   
 if (dat[i,"age"]==65 | dat[i,"age"]==66 | dat[i,"age"]==67 | dat[i,"age"]==68 | dat[i,"age"]==69) {  
 dat$range[i]="65-69"  
 }  
 else if (dat[i,"age"]==70 | dat[i,"age"]==71 | dat[i,"age"]==72 | dat[i,"age"]==73 | dat[i,"age"]==74) {  
 dat$range[i]="70-74"  
 }  
 else (dat$range[i]="75-89")  
   
}  
table(dat$range)

##   
## 65-69 70-74 75-89   
## 896 970 1217

Above, This shows a different number than is reported in the report. Report has:

"65-69"=972  
"70-74"=984  
"75-89"=1127

In an attempt to see if this data might just not be cleaned yet:

for (i in 1:x) {  
 if (dat[i,"age"] > 89) {  
 print(dat[i,"age"])  
 }  
 }

## [1] 90  
## [1] 90  
## [1] 90

Above yields 3 entries that are greater than 89, not enogh to explain the difference. Odd thing is – they have the same total of observations, I’m not sure why there different Other than that I think this dataset is not the same as the one used in the final results of the paper  
  
Here is the table as found in the academic paper:

range = table(dat$range)  
range = t(range)  
  
ages = as.data.frame(range)  
keep = c("Var2", "Freq")  
ages = ages[keep]  
y = sum(ages$Freq)  
ages$percent = 1  
for (i in 1:3) {ages[i,"percent"]=100\*as.numeric(ages[i,"Freq"])/y}  
for (i in 1:3) {ages[i,"percent"]=round(ages[i,"percent"], digits=1)}  
colnames(ages) = c("Age group, years", "No.", "%")  
ages

## Age group, years No. %  
## 1 65-69 896 29.1  
## 2 70-74 970 31.5  
## 3 75-89 1217 39.5

**MAKING THE “ANNUAL MEDICAL HOSPITAL VOLUME” FIGURE FROM TABLE 1**

dat$hosp=1  
for (i in 1:x) {  
 if (dat[i,"hospvol"]<=11) {  
 dat$hosp[i] = "0-11"  
 }  
 else if (12 <= dat[i,"hospvol"] & dat[i,"hospvol"] <= 23) {  
 dat$hosp[i] = "12-23"  
 }  
 else {dat$hosp[i] = "=>24"}  
 }  
  
hospTab = table(dat$hosp)  
hospTab = t(hospTab)  
  
hospVol = as.data.frame(hospTab)  
keep = c("Var2", "Freq")  
hospVol = hospVol[keep]  
y = sum(hospVol$Freq)  
hospVol$percent = 1  
for (i in 1:3) {hospVol[i,"percent"]=100\*as.numeric(hospVol[i,"Freq"])/y}  
for (i in 1:3) {hospVol[i,"percent"]=round(hospVol[i,"percent"], digits=1)}  
colnames(hospVol) = c("Annual Medicare Hospital Volume, no. of cases", "No.", "%")  
hospVol

## Annual Medicare Hospital Volume, no. of cases No. %  
## 1 =>24 1540 50.0  
## 2 0-11 712 23.1  
## 3 12-23 831 27.0

Once agian, there is a difference in my data as compared to the reported data in the study. With the inconsistency of the variables I have and the one in the report, I can’t help But this I simply have the incomplete dataset.  
**MAKING THE “ANNUAL MEDICAL SURGEON VOLUME” FIGURE FROM TABLE 1**

dat$surg=1  
for (i in 1:x) {  
 if (dat[i,"surgvol"]<=5) {  
 dat$surg[i] = "0-5"  
 }  
 else if (6 <= dat[i,"surgvol"] & dat[i,"surgvol"] <= 11) {  
 dat$surg[i] = "6-11"  
 }  
 else {dat$surg[i] = "=>12"}  
 }  
  
surgTab = table(dat$surg)  
surgTab = t(surgTab)  
  
surgVol = as.data.frame(surgTab)  
keep = c("Var2", "Freq")  
surgVol = surgVol[keep]  
y = sum(surgVol$Freq)  
surgVol$percent = 1  
for (i in 1:3) {surgVol[i,"percent"]=100\*as.numeric(surgVol[i,"Freq"])/y}  
for (i in 1:3) {surgVol[i,"percent"]=round(surgVol[i,"percent"], digits=1)}  
colnames(surgVol) = c("Annual Medicare Surgeon Volume, no. of cases", "No.", "%")  
surgVol

## Annual Medicare Surgeon Volume, no. of cases No. %  
## 1 =>12 1281 41.6  
## 2 0-5 930 30.2  
## 3 6-11 872 28.3

This one seems to be the closest to the report, but still not quite exactly. It looks like this has more inclusion, too.