Data discription and transformation

Yiyang LIU

2019/12/4

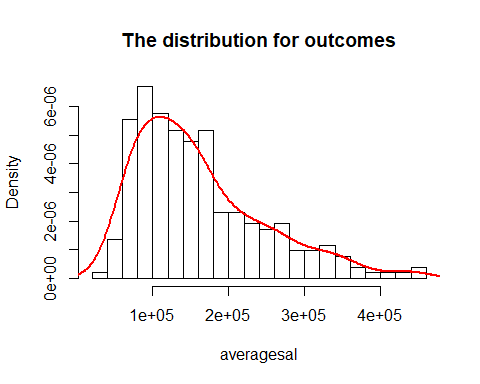
lawsuit =   
 read\_csv("./Lawsuit.csv") %>%   
 janitor::clean\_names()

## Warning in FUN(X[[i]], ...): strings not representable in native encoding  
## will be translated to UTF-8

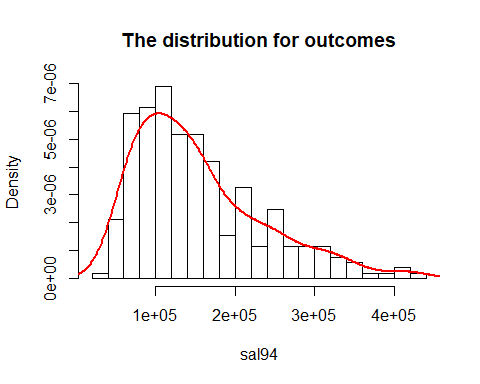
my\_labels <- list(dept = "Department,n(%)", gender = "Gender,n(%)", clin =   
 "Clin,n(%)",cert = "Cert,n(%)", rank = "Rank,n(%)")  
  
# Clean the output  
my\_controls <- tableby.control(  
 total = T,  
 test = T, # No test p-values yet  
 numeric.stats = c("meansd", "medianq1q3"),  
 cat.stats = c("countpct"),  
 digits = 2,  
 stats.labels = list(  
 meansd = "Mean (SD)",  
 medianq1q3 = "Median (Q1, Q3)",  
 countpct = "N (%)"))  
  
# Make some factors to show N (%)  
pb\_1a <- lawsuit %>%   
 mutate(dept = factor(dept, labels = c("Biochemistry/Molecular Biology", "Physiology","Genetics","Pediatrics","Medicine","Surgery"))) %>%  
 mutate(gender = factor(gender, labels = c("Female","Male"))) %>%  
 mutate(clin = factor(clin, labels = c("Primarily research emphasis", "Primarily clinical emphasis"))) %>%  
 mutate(cert = factor(cert, labels = c("Nor certified", "Board certified"))) %>%   
 mutate(rank = factor(rank, labels = c("Assistant", "Associate", "Full professor")))   
   
tab1 = tableby(gender ~ dept + clin + cert + prate + exper + rank + sal94 + sal95, data = pb\_1a, control = my\_controls)  
tab1 %>%   
summary(title = "EDA", labelTranslations = my\_labels, text = T) %>%   
 knitr::kable()

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Female (N=106) | Male (N=155) | Total (N=261) | p value |
| Department,n(%) |  |  |  | < 0.001 |
| - Biochemistry/Molecular Biology | 20 (18.9%) | 30 (19.4%) | 50 (19.2%) |  |
| - Physiology | 20 (18.9%) | 20 (12.9%) | 40 (15.3%) |  |
| - Genetics | 11 (10.4%) | 10 (6.5%) | 21 (8.0%) |  |
| - Pediatrics | 20 (18.9%) | 10 (6.5%) | 30 (11.5%) |  |
| - Medicine | 30 (28.3%) | 50 (32.3%) | 80 (30.7%) |  |
| - Surgery | 5 (4.7%) | 35 (22.6%) | 40 (15.3%) |  |
| Clin,n(%) |  |  |  | 0.197 |
| - Primarily research emphasis | 46 (43.4%) | 55 (35.5%) | 101 (38.7%) |  |
| - Primarily clinical emphasis | 60 (56.6%) | 100 (64.5%) | 160 (61.3%) |  |
| Cert,n(%) |  |  |  | 0.074 |
| - Nor certified | 36 (34.0%) | 37 (23.9%) | 73 (28.0%) |  |
| - Board certified | 70 (66.0%) | 118 (76.1%) | 188 (72.0%) |  |
| prate |  |  |  | 0.004 |
| - Mean (SD) | 5.35 (1.89) | 4.65 (1.94) | 4.93 (1.94) |  |
| - Median (Q1, Q3) | 5.25 (3.73, 7.27) | 4.00 (3.10, 6.70) | 4.40 (3.20, 6.90) |  |
| exper |  |  |  | < 0.001 |
| - Mean (SD) | 7.49 (4.17) | 12.10 (6.70) | 10.23 (6.23) |  |
| - Median (Q1, Q3) | 7.00 (5.00, 10.00) | 10.00 (7.00, 15.00) | 9.00 (6.00, 14.00) |  |
| Rank,n(%) |  |  |  | < 0.001 |
| - Assistant | 69 (65.1%) | 43 (27.7%) | 112 (42.9%) |  |
| - Associate | 21 (19.8%) | 43 (27.7%) | 64 (24.5%) |  |
| - Full professor | 16 (15.1%) | 69 (44.5%) | 85 (32.6%) |  |
| sal94 |  |  |  | < 0.001 |
| - Mean (SD) | 118871.27 (56168.01) | 177338.76 (85930.54) | 153593.34 (80469.67) |  |
| - Median (Q1, Q3) | 108457.00 (75774.50, 143096.00) | 155006.00 (109687.00, 231501.50) | 133284.00 (90771.00, 200543.00) |  |
| sal95 |  |  |  | < 0.001 |
| - Mean (SD) | 130876.92 (62034.51) | 194914.09 (94902.73) | 168906.66 (88778.43) |  |
| - Median (Q1, Q3) | 119135.00 (82345.25, 154170.50) | 170967.00 (119952.50, 257163.00) | 148117.00 (99972.00, 218955.00) |  |

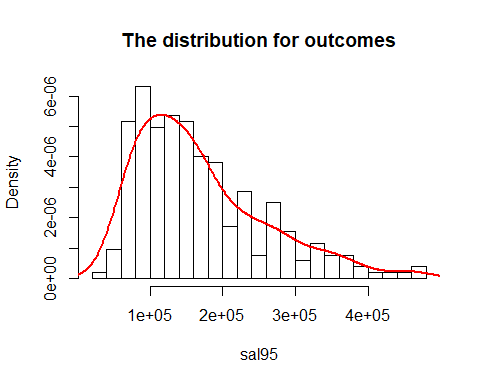
lawsuit =   
 lawsuit %>%   
 mutate(  
 aveg\_sal = (sal94 + sal95)/2  
 )  
  
hist(lawsuit$aveg\_sal,breaks = 20,xlab = "averagesal", freq=F, main="The distribution for outcomes")  
lines(density(lawsuit$aveg\_sal,na.rm=T),col="red",lwd=2)



hist(lawsuit$sal94,breaks = 20,xlab = "sal94", freq=F, main="The distribution for outcomes")  
lines(density(lawsuit$sal94,na.rm=T),col="red",lwd=2)



hist(lawsuit$sal95,breaks = 20,xlab = "sal95", freq=F, main="The distribution for outcomes")  
lines(density(lawsuit$sal95,na.rm=T),col="red",lwd=2)



lawsuit =   
 lawsuit %>%   
 mutate(  
 lg\_aveg\_sal = log(aveg\_sal)  
 )  
  
hist(lawsuit$lg\_aveg\_sal,breaks = 15,xlab = "lg\_averagesal", freq=F, main="The distribution for lg\_outcomes")  
lines(density(lawsuit$lg\_aveg\_sal,na.rm=T),col="red",lwd=2)

