September 25, 2017

STAT 545: Solutions to Homework 2

Dennis D. Cox
Department of Statistics
Rice University

September 25, 2017

1 Exercise 2.15

The first thing to do is to get the data into R. I checked the website for the book and the data are not posted there, so I guess I have to do it by hand. I found it easiest to put it in a spreadsheet and save is as a .csv file and then read it into R. Also, for convenience, I wrote a function to compute the odds ratio from a 2x2 table given as a 4 dimensional vector $(n_{11}, n_{21}, n_{12}, n_{22})$, where the middle two components may be switched (doesn't matter for the computation). Here is the function:

```
OddsRatio = function(nvec){
# function to compute estimated odds ratio from a 4-vector (n11,n21,n12,n22)
theta = nvec[1]*nvec[4]/(nvec[2]*nvec[3])
return(theta)
}
```

Here is the R session:

```
> raw = read.csv("data.csv", header=F)
> raw
   V1  V2  V3  V4
1 512 313  89  19
2 353 207  17   8
3 120 205 202 391
4 138 279 131 244
```

```
5  53 138  94 299
6  22 351  24 317
> apply(raw,2,sum)
  V1   V2   V3   V4
1198 1493  557 1278
> # checks with column totals in the book
> dimnames(raw) = list(rownames=c("A","B","C","D","E","F"),colnames=c("MY","MN","F
> source("OddsRatio.R")
> apply(raw,1,OddsRatio)
        A         B         C         D         E         F
0.3492120 0.8025007 1.1330596 0.9212838 1.2216312 0.8278727
> 512*19/(313*89)
[1] 0.349212
> # just checking
> OddsRatio(apply(raw,2,sum))
     MY
1.84108
```

So, just looking at the (estimated) marginal odds ratio $\hat{\theta}_{AG} = 1.84108$, we might conclude that males have almost twice the odds of acceptance than female applicants. However, the conditional odds ratios, conditioning on a department D, ranges between 0.349 and 1.22, with 4 out of 6 departments having a higher odds for accepting female applicants, and none having an odds ratio as high as the marginal odds ratio. This illustrates some of the concepts described at length in the book, pp. 50-52. It is of interest to look at the proportion of males within departments, the proportion of successful applicants, and the relative size of the departments:

```
> rowsums = apply(raw,1,sum)
> rowprops = rowsums/sum(rowsums)
> rowprops = round(rowprops,digits=3)
> maleprops = (raw[,1]+raw[,2])/rowsums
> maleprops = round(maleprops,digits=3)
> acceptprops = (raw[,1]+raw[,3])/rowsums
> acceptprops = round(acceptprops,digits=3)
> condoddsratio = apply(raw,1,OddsRatio)
> condoddsratio = round(condoddsratio,digits=3)
> cbind(rowprops,maleprops,acceptprops,condoddsratio)
  rowprops maleprops acceptprops condoddsratio
```

```
A    0.206    0.884    0.644    0.349
B    0.129    0.957    0.632    0.803
C    0.203    0.354    0.351    1.133
D    0.175    0.527    0.340    0.921
E    0.129    0.327    0.252    1.222
F    0.158    0.522    0.064    0.828
```

Departments A and B have relatively few female applicants, and they have the lowest two conditional odds ratio, so their small proportions of female applicants means they contribute relatively little to the marginal odds of acceptance of female applicants. Departments A and B also have the highest acceptance rates. Departments C and E have lower proportions of male applicants, and the two highest values in the odds ratio. They also have lower overall acceptance rates than A and B. The ranges in the proportions over all applicants for each department vary from .129 to .206, which isn't that much, so no single department is dominating the marginal result. Thus, it seems the higher acceptance rates in the departments dominated by males (A and B) and the lower acceptance rates in departments with clear majority of female applicants (C and E) is driving the marginal result.

2 Exercise 2.24

I put the data in a spreadsheet again. I had a little bit of a problem because reading it in with the read.csv function creates a data frame, and the cumsum function wasn't operating correctly. Anyway, here's the R code to plot the cdf's coding the levels of job satisfaction as 1,2,3,4 in the order they appear.

```
> jobsat = read.csv("data2.24.csv",header=F)
> jobsat = as.matrix(jobsat)
> cdf = cumsum(jobsat[1,])/sum(jobsat[1,])
> cdf = rbind(cdf,cumsum(jobsat[2,])/sum(jobsat[2,]))
> matplot(t(cdf),xlab="Job Satisfaction",ylab="CDF",type="l")
> title(main="Cumulative Proportions of Job Satisfaction",
+ sub="Black line for Black workers; Red line for white workers")
```

The plot of the two c.d.f.'s is given below, on the last page. One can see that the c.d.f. for the white workers is below the one for the black workers, so the job satisfaction is stochastically ordered with white having more than black.

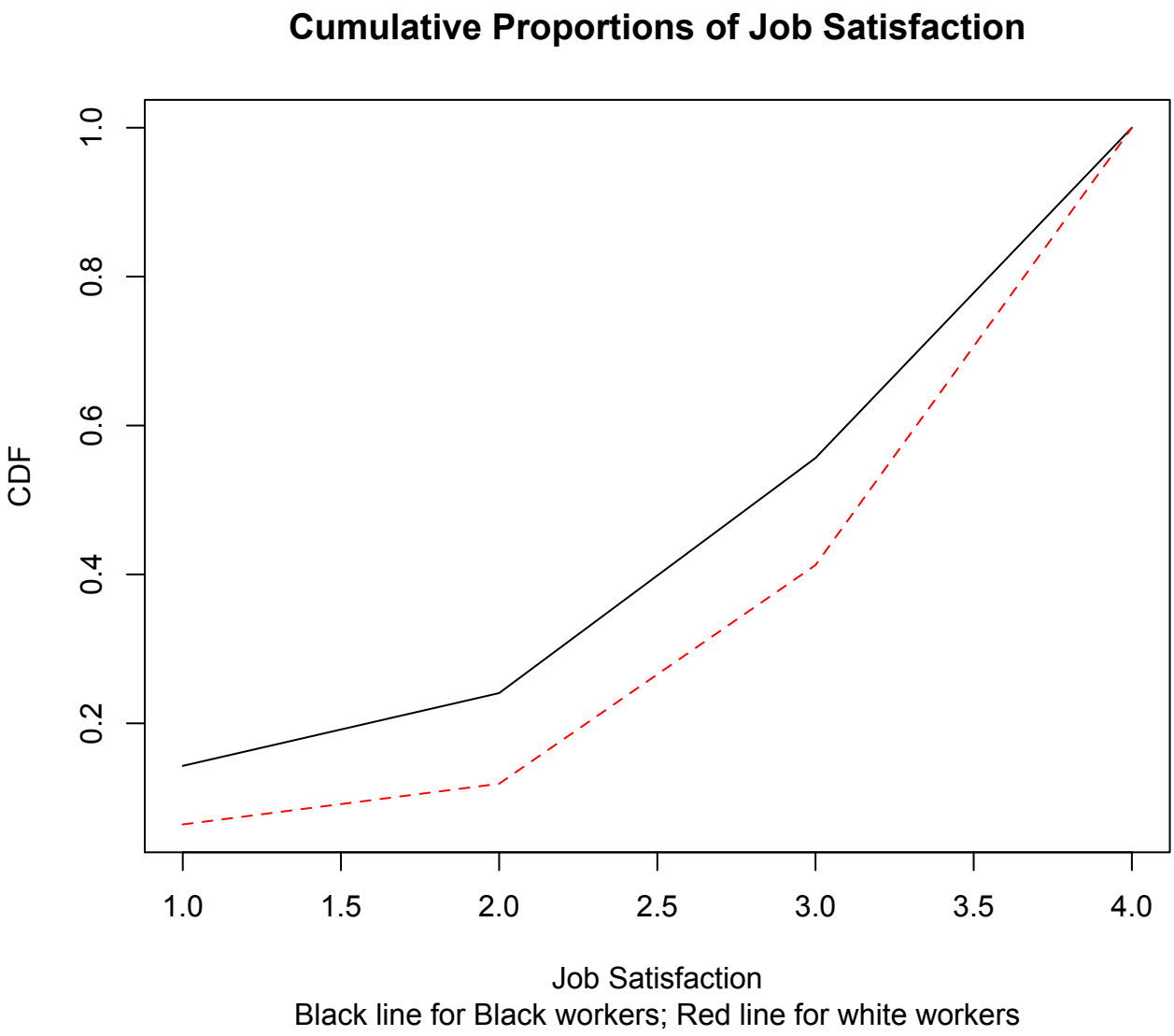

Figure 1: Plot of empirical c.d.f.'s of job satisfaction for black and white workers. The red dashed line (white workers) is below the solid black line (black workers), thus showing a stochastic ordering.

The estimated probability (proportion) that job satisfaction is higher for black workers than white workers is computed next:

```
> pWHB = 19*sum(jobsat[2,2:4])+13*(215+430)+42*430
> pWHB = pWHB/(sum(jobsat[1,])*sum(jobsat[2,]))
> pWHB
[1] 0.4053166
> pBHW = 47*sum(jobsat[1,2:4])+40*(42+59)+215*59
> pBHW = pBHW/(sum(jobsat[1,])*sum(jobsat[2,]))
> pBHW
[1] 0.2268273
> pBHW - pWHB
[1] -0.1784893
```

The proportion of pairs of a black worker and a white worker where the white has higher satisfaction than the black is 0.4053, approximately. The requested difference of probabilities is -.178, approximately. We knew it would be negative because of the stochastic ordering. I don't know how informative the difference is.

3 Exercise 2.25

(a) This has been stated in class - it's just Bayes rule. We give the derivation in detail. Let $Dis \in \{T, F\}$ and $Test \in \{+, -\}$ denote the disease (true or false) and test result (positive or negative).

$$\begin{aligned} & P[Dis = T|Test = +] \\ = & \frac{P[Dis = T \& Test = +]}{P[Test = +]} \\ = & \frac{P[Test = +|Dis = T]P[Dis = T]}{P[Test = +|Dis = T]P[Dis = T] + P[Test = +|Dis = F]P[Dis = F]} \\ = & \frac{\pi_1 \rho}{\pi_1 \rho + \pi_2(1-\rho)}. \end{aligned}$$

(b) Plugging in the numbers and recalling that π_2 is one minus specificity,

$$\begin{aligned} P[Dis = T|Test = +] &= \frac{0.95 * 0.005}{0.95 * 0.05 + 0.05 * 0.995} \\ &= 0.08715596. \end{aligned}$$

Thus, even though the test is very accurate (with high sensitivity and specificity), a positive test result doesn't mean a high probability of actually having disease.

(c) I don't really like this part of the exercise - the result above is not surprising since the prevalence ρ is small. In terms of odds, we have the simple formula

$$\text{(posterior odds)} \;=\; \text{(likelihood ratio)} \times \text{(prior odds)}.$$

In this setting, that means

$$Odds[Dis = T|Test = +] \;=\; \frac{P[Test = +|Dis = T]}{P[Test = +|Dis = F]} \times \frac{\rho}{1-\rho}. \;=\; \frac{\pi_1}{\pi_2} \times \frac{\rho}{1-\rho}.$$

The prior odds of disease, $0.005/0.995 = 1/199$ are quite small, and the likelihood ratio is only $.95/(1-.95) = 19$. In order to get the posterior odds near 1 (which is 0.5 probability) we would need the likelihood ratio near 199.

I guess the grader can just give credit if you did something reasonable with this. I don't know that the tree or joint distribution provides more insight than what I have presented here.

(d) I think I answered it in part (d). Plugging in the new prevalence, we get

$$Odds[Dis = T|Test = +] \;=\; 19 * (0.10/0.90) \;=\; 19/9 \;=\; 2.1111.$$

So the positive predictive value is

$$P[Dis = T|Test = +] \;=\; 19/(19+9) \;=\; 0.6785714.$$

4 Exercise 2.30

I guess what we need to show is

$$\left|\frac{\pi_1}{\pi_2} - 1\right| \;\leq\; \left|\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2} - 1\right|.$$

Of course, the right hand side of the equation may be written as

$$\left|\frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} - 1\right|.$$

Now if $\pi_1 \geq \pi_2$ then $(1-\pi_1) \leq (1-\pi_2)$ and

$$1 \;\leq\; \frac{\pi_1}{\pi_2} \;\leq\; \frac{\pi_1}{\pi_2} * \frac{1-\pi_2}{1-\pi_1},$$

and then the desired inequality holds. If $\pi_2 \geq \pi_1$, then the corresponding inequalities hold with the two π's switched, and when we take reciprocals it reverses all inequalities, so in this case we get

$$1 \geq \frac{\pi_1}{\pi_2} \geq \frac{\pi_1}{\pi_2} * \frac{1-\pi_2}{1-\pi_1},$$

and the desired inequality holds in this case as well.