The Mark of a Criminal Record Revisited

The dataset is called exam3.csv. You may not need to use all of these variables for this activity. We’ve kept these unnecessary variables in the dataset because it is common to receive a dataset with much more information than you need.

|  |  |
| --- | --- |
| Name | Description |
| jobid | Job ID number |
| callback | 1 if tester received a callback, 0 if the tester did not receive a callback. |
| black | 1 if the tester is black, 0 if the tester is white. |
| crimrec | 1 if the tester has a criminal record, 0 if the tester does not. |
| interact | 1 if tester interacted with employer during the job application, 0 if tester doesn’t interact with employer. |
| city | 1 is job is located in the city center, 0 if job is located in the suburbs. |
| distance | Job’s average distance to downtown. |
| custserv | 1 if job is in the costumer service sector, 0 if it is not. |
| manualskill | 1 if job requires manual skills, 0 if it does not. |

The problem will give you practice with:

* constructing confidence intervals
* difference-of-means tests
* p-values
* type I and type II errors

## Question 1

Begin by loading the data into R and explore the data. How many cases are there in the data? Run summary() to get a sense of things. In how many cases is the tester black? In how many cases is he white?

## Answer

#Read the data   
df <- read.csv("exam3-1.csv")  
summary(df)

## jobid callback black crimrec   
## Min. : 1.00 Min. :0.0000 Min. :0.000 Min. :0.0000   
## 1st Qu.: 87.75 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000   
## Median :1024.50 Median :0.0000 Median :1.000 Median :0.0000   
## Mean : 658.57 Mean :0.1638 Mean :0.569 Mean :0.4986   
## 3rd Qu.:1112.25 3rd Qu.:0.0000 3rd Qu.:1.000 3rd Qu.:1.0000   
## Max. :1200.00 Max. :1.0000 Max. :1.000 Max. :1.0000   
##   
## interact city distance custserv   
## Min. :0.0000 Min. :0.0000 Min. : 0.00 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.: 8.00 1st Qu.:0.0000   
## Median :0.0000 Median :0.0000 Median :12.00 Median :1.0000   
## Mean :0.2428 Mean :0.3919 Mean :11.96 Mean :0.6282   
## 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:16.00 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :25.00 Max. :1.0000   
## NA's :2 NA's :2 NA's :2   
## manualskill   
## Min. :0.0000   
## 1st Qu.:0.0000   
## Median :0.0000   
## Mean :0.4813   
## 3rd Qu.:1.0000   
## Max. :1.0000   
## NA's :2

n\_black <- sum(df$black)  
n\_white <- sum(df$white)

There are 696 cases in data. There are in 396 cases the tester is black, and zero cases of white testers.

## Question 2

Now we examine the central question of the study. Calculate the proportion of callbacks for white applicants with a criminal record, white applicants without a criminal record, black applicants with a criminal record, and black applicants without a criminal record.

white.crime <- sum(df[(df$black == 0) & (df$crimrec == 1), "callback"])/nrow(df)  
white.nocrime <- sum(df[(df$black == 0) & (df$crimrec == 0), "callback"])/nrow(df)  
black.crime <- sum(df[(df$black == 1) & (df$crimrec == 1), "callback"])/nrow(df)  
black.nocrime <- sum(df[(df$black == 1) & (df$crimrec == 0), "callback"])/nrow(df)

The rate of callbacks with regard to race and criminal record is: for whites with criminal record - 3.6%, for whites with no criminal record - 7.3%, while for blacks with criminal record the rate is 1.4% and for blacks without a criminal record - 4%. These rates demonstrate that a white person without a criminal record is almost twice (0.073/0.040=1.8) likely to a get a callback in comparison to a black person without criminal record. White people with criminal background are (0.0359/0.0144=2.49) 2.5 times more likely to receive a callback in comparison to black person with criminal record.

## Question 3

Now consider the callback rate for white applicants with a criminal record. Construct a 95% confidence interval around this estimate. Also, construct a 99% confidence interval around this estimate.

#defining two alphas and computing the critical z-scores  
alpha\_05 <- 0.05  
alpha\_01 <- 0.01  
z\_05 <- qnorm(1-alpha\_05/2)  
z\_01 <- qnorm(1-alpha\_01/2)  
#computing CI for 95%  
ci\_95 <- c(NA, NA)  
ci\_95[1] <- white.crime - z\_05\*sqrt(white.crime\*(1 - white.crime)/nrow(df))  
ci\_95[2] <- white.crime + z\_05\*sqrt(white.crime\*(1 - white.crime)/nrow(df))  
  
#computing CI 99%  
ci\_99 <- c(NA, NA)  
ci\_99[1] <- white.crime - z\_01\*sqrt(white.crime\*(1 - white.crime)/nrow(df))  
ci\_99[2] <- white.crime + z\_01\*sqrt(white.crime\*(1 - white.crime)/nrow(df))

With estimated proportion 0.0396 we can state that tthe true value lies between 0.0221 and 0.0497 with 95% confidence. Similarly, the true value of callback for whites with criminal record lies between 0.0178 and 0.0541 with confidence of 99%.

## Question 4

Calculate the estimated effect of a criminal record for white applicants by comparing the callback rate in the treatment condition and the callback rate in the control condition. Create a 95% confidence interval around this estimate. Next, describe the estimate and confidence interval in a way that could be understood by a general audience.

#creating a sample of only whites  
white <- df[df$black == 0,]  
#calculating number of whites with and without a crim record  
wcrime.count <- nrow(white[white$crimrec == 1,])  
wncrime.count <- nrow(white[white$crimrec == 0,])   
#computing the callback rate from white df   
wcrime.rate <- sum(white[white$crimrec == 1, "callback"])/wcrime.count  
wncrime.rate <- sum(white[white$crimrec == 0, "callback"])/wncrime.count  
  
#finding ATE and std error  
ATE <- wcrime.rate - wncrime.rate  
ATE\_se <- sqrt((wcrime.rate\*(1 - wcrime.rate))/wcrime.count + (wcrime.rate\* (1 - wcrime.rate))/ wcrime.count)  
  
#building CI 95%  
ATE\_ci <- c(NA, NA)  
ATE\_ci[1] <- ATE - qnorm (1 - 0.05/2)\* ATE\_se  
ATE\_ci[2] <- ATE + qnorm (1 - 0.05/2)\* ATE\_se  
print(ATE\_ci)

## [1] -0.25767675 -0.08898991

ATE = -0.1733 and 95% CI is (-0.258. -0.089) For white population there is a 17.3% decrease of getting a callback if applicant has a criminal record. For instance, the if we compare two white applicants the one with criminal record has 17.3% lower chance of receiving a callback. However, this callback rate is calculated based on 300 white population and with 95% of confidence interval the true effect lies between -0.258 and -0.089.

## Question 5

Assuming a null hypothesis that there is no difference in callback rates between white people with a criminal record and white people without a criminal record, what is the probability that we would observe a difference as large or larger than the one that we observed in a sample of this size?

white\_test <- t.test(data = white, callback ~ crimrec)  
white\_test$p.value

## [1] 0.0005207343

The probability of observing a difference as large or larger than the one that we observed in a sample is basically a probability of getting a z-score as large or even larger than observed z- score. It is the p-value from the two- sample t-test. We use our white only sample and conduct t-test for a callback rates using criminal records as the factor variable. Hence, the probability is just 0.5% which is less than 1%.

## Question 6

Imagine that we set up a hypothesis test where the null hypothesis is that there is no difference in callback rates between whites with and without a criminal record. In the context of this problem, what would it mean to commit a type I error? In the context of this problem, what would it mean to commit a type II error? If we set for a two-tailed test are we specifying the probability of type I error or type II error?

## Answer 6

The type I error is rejecting the null hypothesis when it is true. Since our null hypothesis is that having criminal record does not influence white applicant’s chances of receiving a callback, assuming it is true the type I error would be to assume that criminal record influences chances of receiving a callback (while it is not true, assuming H0 holds) The type II error happens when we fail to reject the H0 which is false. In our example it means that in reality criminal background changes the chances of white applicant to receive a callback (since we assume that here H0 is false) but we fail to reject it and wrongfully conclude that there is no difference in callback rates between white applicants with and without criminal record. Alpha is probability of concluding that we have a positive result when we actually reject H0. Hence, alpha is probability of committing a type I error. If we set alpha = 0.05 it means that we leave a 5% chance of wrongfully keeping the H0 which is false.