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 CUDJOE SMITH

UIN: 231001378

Description: Exams 3

`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

```
mark<- read.csv("exam3-1.csv")
nrow(mark)
```

```
## [1] 696
```

```
nrow(subset(mark,select = "black",subset  = black==1))
```

```
## [1] 396
```

```
nrow(subset(mark,select = "black",subset  = black==0))
```

```
## [1] 300
```

```
summary(mark)
```

```
##      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
```

```
# There are 696 cases in total. 396 of the testers are black and 300 are
white.
```

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.

```
#Proportion of white applicants with criminal records
ncase<-nrow(mark)
white_crime<-nrow(subset(mark,select = c("callback"),subset = c(black==0 &
callback==1 & crimrec==1)))
```

```
prop_white_crime<-white_crime/ncase
prop_white_crime
```

```
## [1] 0.03591954
```

```
#Proportion of white applicants without criminal records
ncase<-nrow(mark)
white_no_crime<-nrow(subset(mark,select = c("callback"),subset = c(black==0 &
callback==1 & crimrec==0)))
prop_white_no_crime<-white_no_crime/ncase
prop_white_no_crime
```

```
## [1] 0.07327586
```

```
# Proportion of callbacks for black applicants with a criminal record
ncase<-nrow(mark)
black_crime<-nrow(subset(mark,select = c("callback"),subset = c(black==1 &
callback==1 & crimrec==1)))
prop_black_crime<-black_crime/ncase
prop_black_crime
```

```
## [1] 0.01436782
```

```
# Proportion of callbacks for black applicants without a criminal record
ncase<-nrow(mark)
black_no_crime<-nrow(subset(mark,select = c("callback"),subset = c(black==1 &
callback==1 & crimrec==0)))
prop_black_no_crime<-black_no_crime/ncase
prop_black_no_crime
```

```
## [1] 0.04022989
```

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.

```
#Confidence Interval of callback rate of white applicant with criminal record
# 95% confidence interval

n<-white_crime
CI_95<-c(prop_white_crime-sqrt(prop_white_crime*(1-
prop_white_crime)/n)*1.96,prop_white_crime+sqrt(prop_white_crime*(1-
prop_white_crime)/n)*1.96)
CI_95
```

```
## [1] -0.03702757  0.10886665
```

```
#Confidence interval of callback rate
# 99% confidence interval
CI_99<-c(prop_white_crime-sqrt(prop_white_crime*(1-
```

```
prop_white_no_crime)/n)*2.58,prop_white_crime+sqrt(prop_white_crime*(1-
prop_white_crime)/n)*2.58)
CI_99
```

```
## [1] -0.05822396  0.13194176
```

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.

```
# Estimate of effect
est_effect<- prop_white_crime-prop_white_no_crime
est_effect
```

```
## [1] -0.03735632
```

```
n1<-white_crime
n2<-white_no_crime
st_error_prop_white_crime<-sqrt(prop_white_crime*(1-prop_white_no_crime)/n1)
st_error_prop_white_no_crime<-sqrt(prop_white_crime*(1-
prop_white_no_crime)/n2)
st_error_est_effect<-sqrt(st_error_prop_white_crime^2 +
st_error_prop_white_no_crime)
CI_est_95<-c(est_effect- st_error_est_effect*1.96,est_effect +
st_error_est_effect*1.96)
CI_est_95
```

```
## [1] -0.3586972  0.2839846
```

```
# The estimate of the difference between the treatment and the condition is -
0.03736. This means it is estimated that just about 4% more of the
application without criminal records than those with criminal records are
likely to have callback. Also the confidence interval indicates that there is
no statistical difference between the rate at which white with criminal
records and those without criminal records receive a callback.
```

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?

```
test_stat<-est_effect/st_error_est_effect
pnorm(test_stat)
```

```
## [1] 0.4098804
```

```
#A p value of 0.409 confirms that there is no statistical difference between
the treatment and the control. There is about 41% chance of observing a
difference at least larger than the observed sample.
```

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 $\alpha = 0.05$ for a two-tailed test are we specifying the probability of type I error or type II error?

```
#A type I error is committed when we say there is a difference, when there is
actually no difference in callback rate between whites with and without
criminal record.

# A type II error is committed when there is actually a difference, but we
say that there is no difference between whites with and without criminal
record.

#When $\alpha = 0.05$ we are actually specifying a the probability of
commiting a type I error.
```