MA677_Final_Report

Chaoqun Yin

5/7/2019

1 Statistics and the Law

Question description: ACORN made a statstical arguement that the difference between the rates of mortgage application refusals of wihte applicants and minority applicatns constitued evidence of discrimination. Use ACORN's data and create the arguements that (1) the data are sufficient evidence of discrimination to warrant corrective action and (2) the data are not sufficient.

This question is equivalent to a test and a power analysis: * H0: The refusal rate of white applicants equals the refusal rate of minority applicants. * H1: The refusal rate of white applicants is higher than the refusal rate of minority applicants.

```
#prepare the data
MIN<-c(20.90,23.23,23.10,30.40,42.70,62.20,39.5,38.40,26.20,55.90,
       49.70,44.60,36.40,32.00,10.60,34.30,42.30,26.50,51.50,47.20) #minority
applicants
WHITE<-c(3.7,5.5,6.7,9.0,13.9,20.6,13.4,13.2,9.3,21.0,20.1,19.1,16.0,
         16.0,5.6,18.4,23.3,15.6,32.4,29.7) #white applicants
df_dis <- data.frame(group = rep(c("MIN", "WHITE"), each = length(MIN)),
                    percent = c(MIN,  WHITE))
#t-test
res_dis<-t.test(percent ~ group, data = df_dis,
         var.equal = TRUE, alternative = "greater")
res_dis$p.value #1.279668e-07

## [1] 1.279668e-07
```

As we can see that the p-value of the t-test is pretty small (smaller than 0.05), so we should reject hypoyhesis 0. The conclusion is that the refusal rate of white applicants (WHITE) is higher than the refusal rate of minority applicants (MIN).

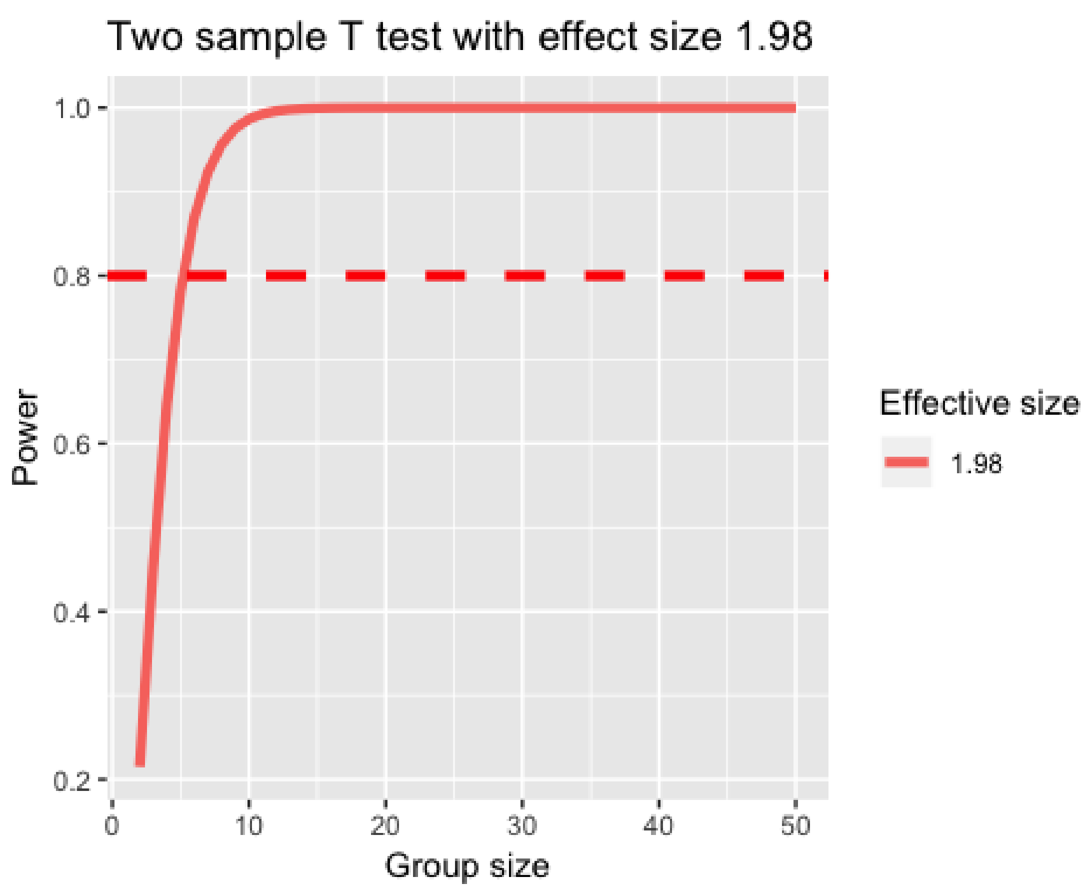
Then the following is a power analysis for confirming data's sufficiency.

```
#Calculate the effect size
sd_p <- sqrt((sd(WHITE) ^ 2 + sd(MIN) ^ 2) / 2)
eff_size <- abs((mean(WHITE) - mean(MIN)) / sd_p)

ptab1 <- cbind(NULL)
n <- seq(2, 50, by = 1)
```

```
for (i in seq(2, 50, by = 1)) {
  pwrt1 <- pwr.t2n.test(
    n1 = i, n2 = i,
    sig.level = 0.05, power = NULL,
    d = eff_size, alternative = "two.sided"
  )
  ptab1 <- rbind(ptab1, pwrt1$power)
}

ptab1<-as.data.frame(ptab1)
colnames(ptab1)[1]<-"num"
ggplot(ptab1) +
  geom_line(aes(x = n, y = num, colour = "orange"), size = 1.5) +
  scale_color_discrete(name = "Effective size",
                       labels =c(round(eff_size, 2))) +
  geom_hline(yintercept = 0.8, linetype = "dashed",
             color = "red", size = 1.5) +
  ylab("Power") +
  scale_y_continuous(breaks = seq(0, 1, by = 0.2)) +
  ggtitle("Two sample T test with effect size 1.98") +
  xlab("Group size")
```



Because of effect size of 1.98, an acceptable level of 0.8 requires more than 5 samples in each group, and we have 20 samples in our data in each group. Therefore, the data is sufficient for this case.

2 Comparing Suppliers

Question description: Acme Student Products sources ornithopters from high schools where students make orithopters as projects in a kinetics sculptor class. Not all of the ornithopers fly. Not all of them look good enough. Revenue aside, which of the three schools produces the higher quality ornithopters, or are do they all produce about the same quality?

This question is equivalent to this test below: * H0: The three schools produce the same quality. * H1: The three shcools produce different level of quality.

```
data_qua <- matrix(c(12,23,89,8,12,62,21,30,119),ncol=3,nrow = 3,byrow=TRUE)
colnames(data_qua) <- c("dead","art","fly")
rownames(data_qua) <- c("Area51","BDV","Giffen")
fly <- as.table(data_qua)
chisq.test(data_qua,correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data_qua
## X-squared = 1.3006, df = 4, p-value = 0.8613
```

The chi-squared result shows that the p-value is extremely large (much larger than 0.05), which leads us to reject H0. The conclusion is that three schools produce the same quality.

3 How deadly are sharks?

Question description: Now that you have the data, please help me sort out how U.S. sharks compare with Australian sharks. Explain your analysis in terms that are simple but technically correct, make sure to include an analysis of statistical power.

For this case, it is equivalent to this test: * H0: Sharks in Australia were, on average, are the same as the sharks in the United States. * H1: Sharks in Australia were, on average, are more vicious than the sharks in the United States.

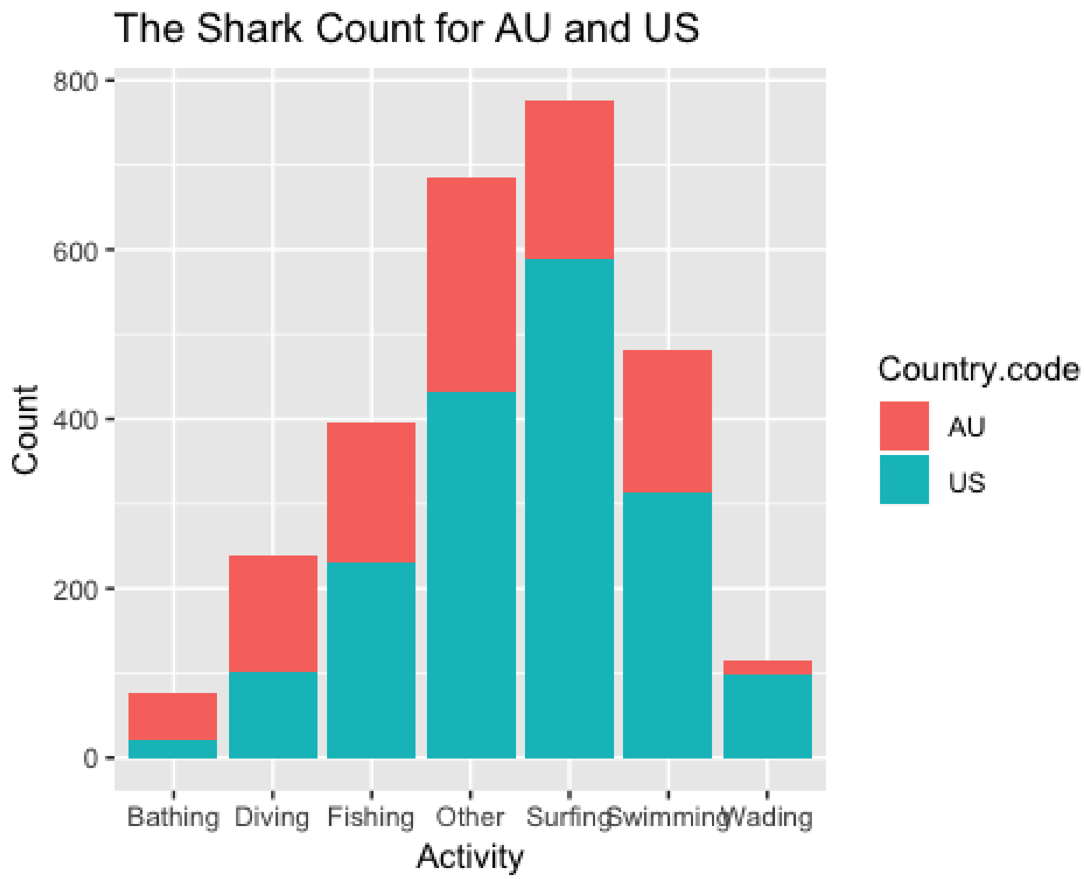
```
df_shark <- read.csv("material/sharkattack.csv")
#In this case we only need sharks in the US and the AU.
df_shark %>% filter(Country.code=="US" | Country.code=="AU") %>%
  filter(Type=="Provoked" | Type=="Unprovoked") -> shark_m
```

```r
shark_m %>%
  group_by(Country.code, Activity) %>%
  summarise(count=n()) %>%
  ungroup() %>%
  group_by(Country.code) %>%
  mutate(percent=count/sum(count)) -> shark_f
kable(shark_f)
```

Country.code	Activity	count	percent
AU	Bathing	55	0.0560652
AU	Diving	137	0.1396534
AU	Fishing	164	0.1671764
AU	Other	254	0.2589195
AU	Surfing	187	0.1906218
AU	Swimming	168	0.1712538
AU	Wading	16	0.0163099
US	Bathing		

22
0.0123043
US
Diving
102
0.0570470
US
Fishing
231
0.1291946
US
Other
431
0.2410515
US
Surfing
590
0.3299776
US
Swimming
313
0.1750559
US
Wading
99
0.0553691

```
#plot the stats
ggplot(shark_f) +
  geom_col(aes(x = Activity,y = count,fill = Country.code)) +
  ylab("Count") +
  ggtitle("The Shark Count for AU and US")
```



From this plot we can see that US sharks' attack incidents are more frequent than AU's.

```
shark_ff <- matrix(c(23,120,275,547,615,347,107,54,129,95,210,186,165,12),
                   nrow=2,dimnames = list(c("AU","US"),
c("Bathing","Diving","Fishing","Other","Surfing","Swimming","Wading")))
chisq.test(shark_ff)

##
##  Pearson's Chi-squared test
##
## data:  shark_ff
## X-squared = 378.78, df = 6, p-value < 2.2e-16
```

The result from chi-square test shows that p-value is smaller than 0.05, so H0 is rejected and the conclusion is that sharks in US are different from those in Australia. From empirical plot, we can see that sharks in US make attack more frequent.