RedSox2018

Kevin Cummiskey

November 14, 2019

Chapter 6.1 Comparing Proportions

Question: Are the Red Sox better at Fenway Park?

The data consists of results from 182 Red Sox games in the 2018 season. Data for this activity is available at https://www.baseball-reference.com/teams/BOS/2018-schedule-scores.shtml

```
head(redsox %>% select(`Gm#`,Tm,Opp,Result, Field))
```

```
## # A tibble: 6 x 5
##   `Gm#` Tm    Opp   Result Field
##   <dbl> <chr> <chr> <fct>  <chr>
## 1     1 BOS   TBR   L      Away
## 2     2 BOS   TBR   W      Away
## 3     3 BOS   TBR   W      Away
## 4     4 BOS   TBR   W      Away
## 5     5 BOS   MIA   W      Away
## 6     6 BOS   MIA   W      Away
```

```
summary = redsox %>%
  group_by(Field)%>%
  count(Result) %>%
  spread(key = Field, value = n)
kable(summary, caption = "Results of the Red Sox 2018 Season")
```

Table 1: Results of the Red Sox 2018 Season

Result	Away	Home
W	51	57
L	30	24

Measures of Association

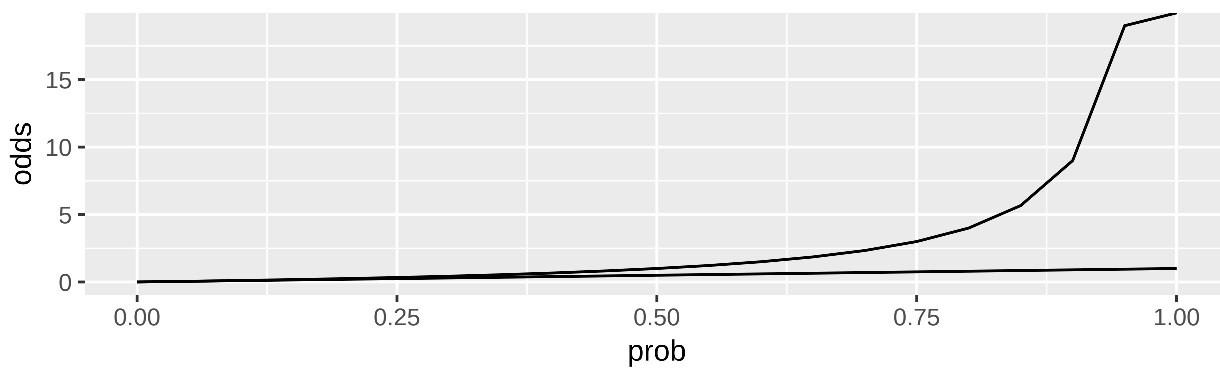
Calculate the *conditional proportions* of wins for home games and away games. (You will also hear conditional proportions referred to as *chances, likelihood, risk*).

Calculate the difference in conditional proportions (also called *risk difference*) comparing home games to away games.

Calculate the *relative risk* for a win comparing home and away games. How does the risk difference and relative risk tell us something different?

Calculate the *odds* of winning at home and away. What are the smallest and largest values the odds can take? (see plot below)

```r
measures = data.frame(prob = seq(0,1, by = 0.05))
measures = measures %>% mutate(odds = prob/(1-prob))
measures %>% ggplot(aes(x = prob, y = odds)) +
  geom_line() +
  geom_line(aes(y = prob))
```



Calculate the *odds ratio* for wins comparing home and away games. What are the smallest and largest values the odds ratio can take? Let's say we take to log of the odds ratio - what are the smallest and largest values the *log odds ratio* can take?

Inference on Difference in Proportions

What are the null and alternative hypotheses for this test?

What is the statistic of interest for this test?

Theory-based test (two sample z-test)

```
# two-sample z-test
phat_home = 57/81
phat_away = 51/81
phat = 108/162
#standardized statistic (pg 420)
z = (phat_home - phat_away)/sqrt(phat*(1-phat)*(1/81 + 1/81))
#p-value
2*(1-pnorm(z,0,1))
```

```
## [1] 0.3173105
```

Theory-based test (χ^2 test)

Fill in the expected values in the table below if home/away has no effect and the Red Sox won 108 games.

Result	Away	Home	Total
W			108
L			54
Total	81	81	182

The χ^2 test compares the observed counts in each cell to the expected counts.

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Calculate the χ^2 statistic.

```
#calculate overall win/loss percentage
summary_overall = redsox %>% group_by(Result) %>%
  count() %>% group_by() %>% mutate(perc = n/sum(n)) %>%
  select(-n)
#calculate expected wins
summary_homeaway = redsox %>%
  group_by(Field,Result) %>%
  count() %>%
  left_join(summary_overall, by = "Result") %>%
  group_by(Field) %>%
  mutate(expected = perc*sum(n))
summary_homeaway
```

```
## # A tibble: 4 x 5
## # Groups:   Field [2]
##   Field Result     n  perc expected
```

```
##   <chr> <fct>  <int> <dbl>    <dbl>
## 1 Away  W         51 0.667       54
## 2 Away  L         30 0.333       27
## 3 Home  W         57 0.667       54
## 4 Home  L         24 0.333       27
```

```r
#calculate chi-square statistic
chisq = summary_homeaway %>% group_by() %>%
  summarise(chisq = sum((n - expected)^2/expected))

1- pchisq(chisq$chisq, 1)
```

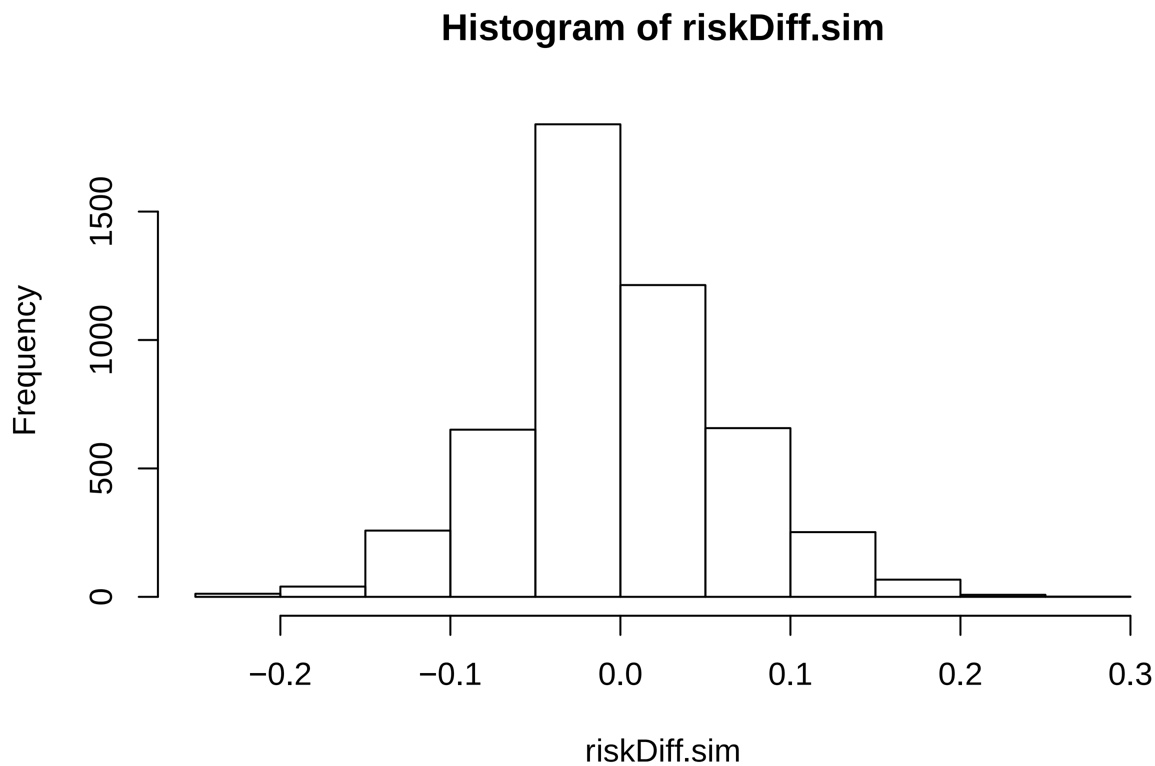
```
## [1] 0.3173105
```

Simulation-based test

```r
redsox.sim = redsox %>% select(Result, Field)
riskDiff.sim = c()
n.sims = 5000

for(i in 1:n.sims){
  summary.sim = redsox.sim %>%
    mutate(Result.sim = sample(Result)) %>% #shuffle wins
    group_by(Field) %>%
    count(Result.sim) %>% mutate(p = n/sum(n)) #calculate win percentages
  riskDiff.sim[i] = summary.sim$p[3]-summary.sim$p[1]
}

hist(riskDiff.sim)
```




```r
sum( abs(riskDiff.sim) > (phat_home - phat_away))/n.sims
```

```
## [1] 0.2372
```

What would we conclude from these tests?

Is confounding an issue in this analysis? What variables might we want to control for in order to reduce confounding?

Intro to Logistic Regression

Let Y_i be whether or not the Red Sox win game i such that $Y_i \sim \text{Bernoulli}(\pi_i)$ be the probability the Red Sox win game i.

Here is our model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{Field}_i$$

where Field_i is whether game i was played on the home or away field.

How do we interpret β_0, β_1? Why is there no ϵ_i in this model?

Let's fit the model.

```r
#reverse factor levels for result
#so win is 1 and loss is 0
redsox$Result = factor(redsox$Result,
                       levels = c("L","W"))
model_homeaway = glm(Result ~ Field,
                     data = redsox,
                     family = "binomial")
summary(model_homeaway)
```

```
##
## Call:
## glm(formula = Result ~ Field, family = "binomial", data = redsox)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5597  -1.4094   0.8383   0.9619   0.9619
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.5306     0.2301   2.306   0.0211 *
## FieldHome     0.3344     0.3349   0.998   0.3181
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 206.23  on 161  degrees of freedom
## Residual deviance: 205.23  on 160  degrees of freedom
## AIC: 209.23
##
## Number of Fisher Scoring iterations: 4
```

Have we seen these estimates before?