RedSox2018

Kevin Cummiskey

November 14, 2019

Chapter 6.1 Comparing Proportions

Question: Are the Red Sox better at Fenway Park?

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, `Home/Away`))
```

```
## # A tibble: 6 x 5
##   `Gm#` Tm    Opp   Result `Home/Away`
##   <int> <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(`Home/Away`)%>%
  count(Result) %>%
  spread(key = `Home/Away`, 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

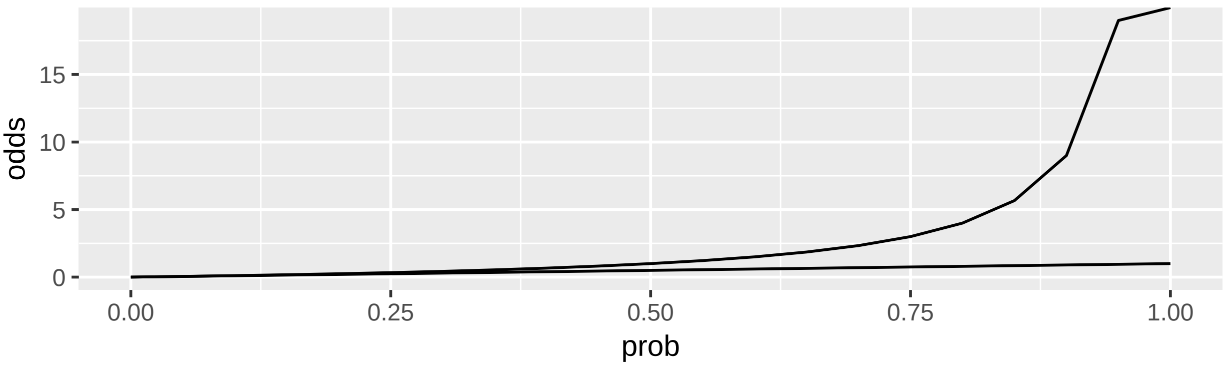
Calculate the *conditional proportions* of wins at home and away. (other names: *chances, likelihood, risk*).

Calculate the difference in conditional proportions (also called *risk difference*).

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 a win at home and away. What are the smallest and largest values the odds of an event 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-square test)

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

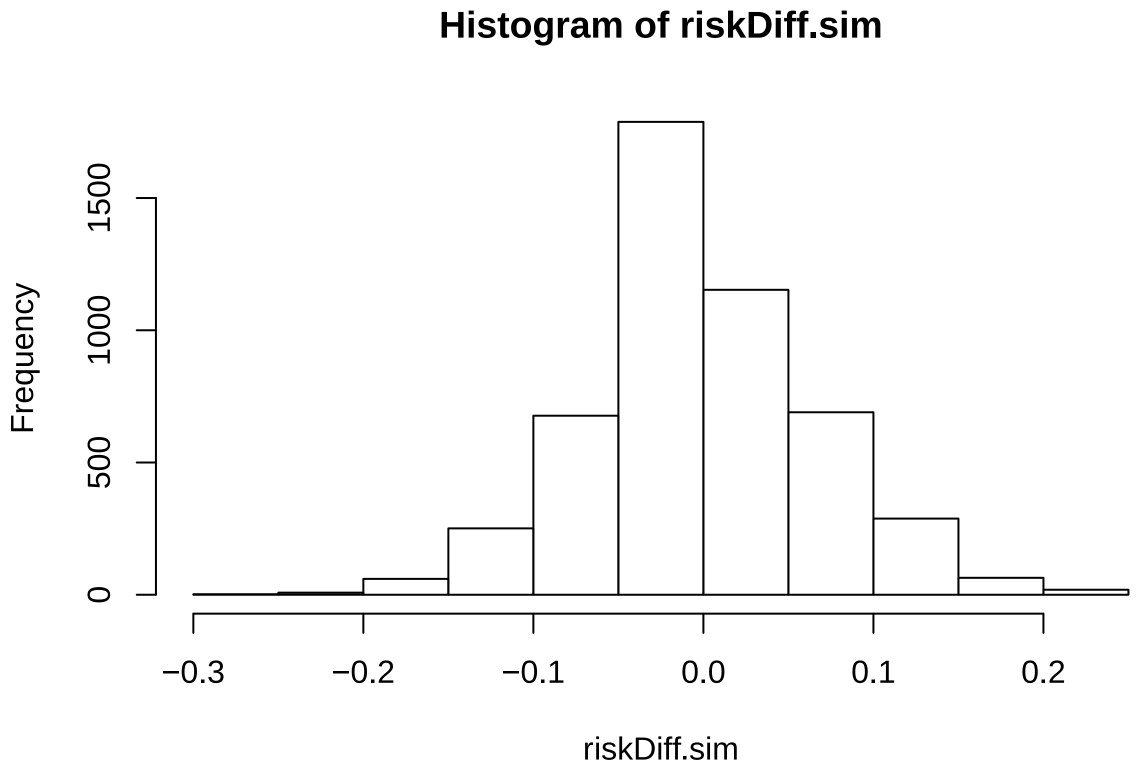
Result	Away	Home	Total
W			108
L			54
Total	81	81	182

Simulation-based test

```
redsox.sim = redsox %>% select(Result, `Home/Away`)
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(`Home/Away`) %>%
    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] 0.2504
```

What would we conclude from these tests?