Inference Examples

Alan Arnholt

Oct 25, 2017

Contents

1 Ideas

- Null hypothesis (H_0): The claim that is not interesting - no real effect. This is the status quo in the absence of the data providing convincing evidence to the contrary.
- Alternative hypothesis (H_A): The claim corresponding to the research hypothesis - there is a real effect.
- Test statistic: a numerical function of the data whose value determines the result of the test. The function itself is generally denoted $T = T(\mathbf{X})$, where $\mathbf{X}$ represents the data. After being evaluated for the sample data $\mathbf{x}$, the result is called an observed test statistic and is written in lowercase, $t = T(\mathbf{x})$.
- Null distribution is the probability distribution of the test statistic when the null hypothesis is true.

1.1 Video Cola Example

```r
OD <- matrix(data = c(28, 6, 19, 7), nrow = 2, byrow = TRUE)
dimnames(OD) <- list(Location = c("East", "West"), Drink = c("Cola", "Orange"))
ODT <- as.table(OD)
ODT
```

```
        Drink
Location Cola Orange
    East   28      6
    West   19      7
```

```r
ODTDF <- as.data.frame(ODT)
DDF <- as.tbl(vcdExtra::expand.dft(ODTDF))
DDF
```

```
# A tibble: 60 x 2
   Location  Drink
     <fctr> <fctr>
 1     East   Cola
 2     East   Cola
 3     East   Cola
 4     East   Cola
 5     East   Cola
```

```
 6      East   Cola
 7      East   Cola
 8      East   Cola
 9      East   Cola
10      East   Cola
# ... with 50 more rows
```

```r
xtabs(~Location + Drink, data = DDF)
```

```
        Drink
Location Cola Orange
    East   28      6
    West   19      7
```

```r
DDF %>%
  group_by(Location) %>%
  summarize(Pcola = mean(Drink == "Cola"), Porange = mean(Drink == "Orange"))
```

```
# A tibble: 2 x 3
  Location     Pcola   Porange
    <fctr>     <dbl>     <dbl>
1     East 0.8235294 0.1764706
2     West 0.7307692 0.2692308
```

```r
# shuffle the data
T1 <- xtabs(~Location + Drink, data = DDF)
set.seed(13)
T2 <- xtabs(~Location + sample(Drink), data = DDF)
T2
```

```
        sample(Drink)
Location Cola Orange
    East   27      7
    West   20      6
```

```r
prop.table(T2, 1)
```

```
        sample(Drink)
Location      Cola    Orange
    East 0.7941176 0.2058824
    West 0.7692308 0.2307692
```

2 Test

2.1 Equality of proportion for cola preference between coasts.

2.1.1 Classical Approach (assumptions not satisfied)

$H_O : p_{\text{West}} - p_{\text{East}} = 0$ versus $H_A : p_{\text{West}} - p_{\text{East}} \neq 0$

Test Statistic: $\hat{p}_{\text{West}} - \hat{p}_{\text{East}}$ - What do we know about the Test Statistic? If we had a sufficiently large sample size (which we really do not) then we could claim that

$$\hat{p}_{\text{West}} - \hat{p}_{\text{East}} \overset{\bullet}{\sim} \mathcal{N}\left(\mu_{\hat{p}_{\text{West}} - \hat{p}_{\text{East}}}, \sigma_{\hat{p}_{\text{West}} - \hat{p}_{\text{East}}}\right),$$

and write a standardized test statistic as:

$$Z = \frac{(\hat{p}_{\text{West}} - \hat{p}_{\text{East}}) - 0}{\sigma_{\hat{p}_{\text{West}} - \hat{p}_{\text{East}}}} = \frac{(\hat{p}_{\text{West}} - \hat{p}_{\text{East}})}{\sqrt{\hat{p}_p(1-\hat{p}_p)\left(\frac{1}{n_W} + \frac{1}{n_E}\right)}}$$

$$\hat{p}_p = \frac{x+y}{n_x+n_y} = \frac{6+7}{34+26} = 0.2166667$$

$$z_{obs} = \frac{0.7307692 - 0.8235294}{\sqrt{0.21666 \cdot (1 - 0.21666) \cdot \left(\frac{1}{34} + \frac{1}{26}\right)}} = -0.8642567$$

The corresponding *p-value* is $P(Z \leq z_{obs}) \cdot 2 = 0.3874469$

```
T1
```

```
        Drink
Location Cola Orange
    East   28      6
    West   19      7
```

```
phats <- prop.table(T1, 1)
phats
```

```
        Drink
Location      Cola    Orange
    East 0.8235294 0.1764706
    West 0.7307692 0.2692308
```

```
phat_p <- (6 + 7)/(34 + 26)
phat_p
```

```
[1] 0.2166667
```

```
phats_d <- phats[2, 1] - phats[1, 1]
phats_d
```

```
[1] -0.09276018
```

```
zobs <- (phats_d)/(sqrt(phat_p*(1 - phat_p)*(1/34 + 1/26)))
zobs
```

```
[1] -0.8642567
```

```
pvalue <- pnorm(zobs)*2
pvalue
```

```
[1] 0.3874469
```

Fail to find evidence to suggest the proportion of people on the west coast that prefer cola is any different from the proportion of people on the east coast that prefer cola.

2.1.2 Randomization Approach

```
phats_d
```

```
[1] -0.09276018
```

```
sims <- 10^3 - 1
ts <- numeric(sims)
set.seed(13)
for(i in 1:sims){
  ps <- prop.table(xtabs(~Location + sample(Drink), data = DDF), 1)
  ts[i] <- ps[2, 1] - ps[1, 1]
}
pvalue <- 2*(sum(ts <= phats_d) + 1)/(sims + 1)
pvalue
```

```
[1] 0.616
```

2.1.3 Independence between `Location` and `Drink` χ^2

H_0 : `Location` and `Drink` are independent. H_A : `Location` and `Drink` are dependent.

- Test Statistic is now: $\chi^2 = \sum \frac{(O-E)^2}{E}$

```
chisq.test(T1, correct = FALSE)
```

```

	Pearson's Chi-squared test

data:  T1
X-squared = 0.74694, df = 1, p-value = 0.3874
```

- **Note** that $z_{obs}^2 = (-0.8642567^2) = 0.7469396 = \chi_{obs}^2 = 0.7469396$.

2.1.4 Randomiztion Approach (2)

```
obs_stat <- chisq.test(T1, correct = FALSE)$stat
obs_stat
```

```
X-squared
0.7469396
```

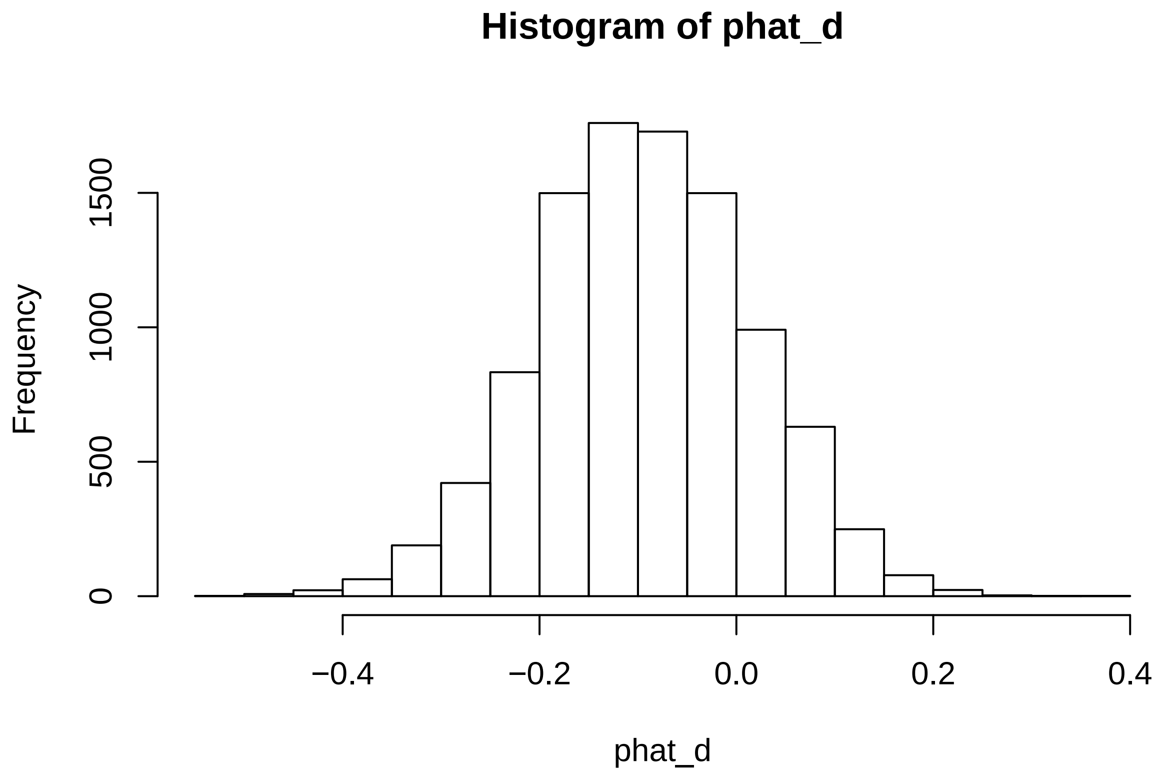
```
sims <- 10^3 - 1
ts <- numeric(sims)
set.seed(13)
for(i in 1:sims){
  ts[i] <- chisq.test(xtabs(~Location + sample(Drink), data = DDF), correct = FALSE)$stat
}
pvalue <- (sum(ts >= obs_stat) + 1)/(sims + 1)
pvalue
```

```
[1] 0.552
```

3 Bootstrapping

```
sims <- 10^4 - 1
phat_d <- numeric(sims)
for(i in 1:sims){
```

```r
  pe <- mean(sample(0:1, size = 34, replace = TRUE, prob = c(6/34, 28/34)))
  pw <- mean(sample(0:1, size = 26, replace = TRUE, prob = c(7/26, 19/26)))
  phat_d[i] <- pw - pe
}
hist(phat_d)
```




```r
quantile(phat_d, prob = c(0.025, 0.975))
```

```
      2.5%      97.5%
-0.3054299  0.1199095
```

```r
# Or
phats_d + c(-1, 1)*qnorm(.975)*sd(phat_d)
```

```
[1] -0.3049119  0.1193915
```