Lab 8

Apurva Shah

2022-05-29

Contents


```
## Date last run: 2022-05-29
```

```
## Hello World!
```

Examples

Requires library xtable.

Super Bowl Coin Tosses

```
## Read in our data
sbdf <- read.table("SuperBowl_coinTosses.tsv", sep="\t", header=TRUE)

head(sbdf, n=6)
```

```
##   SuperBowl             Matchup CoinToss Coin.TossWinner GameWinner
## 1         1   Chiefs vs Packers    Heads         Packers    Packers
## 2         2  Packers vs Raiders    Tails         Raiders    Packers
## 3         3       Colts vs Jets    Heads            Jets       Jets
```

```
## 4         4   Vikings vs Chiefs    Tails         Vikings     Chiefs
## 5         5    Colts vs Cowboys    Tails         Cowboys      Colts
## 6         6 Cowboys vs Dolphins    Heads        Dolphins    Cowboys
##   CoinTossWinnerEqualGameWinner.
## 1                            Yes
## 2                             No
## 3                            Yes
## 4                             No
## 5                             No
## 6                             No
```

Imagine that these 56 observations were randomly realized from an infinite population of Super Bowl coin tosses.

Let's test the claim that Super Bowl coin toss is not fair; that head and tails are not equally likely, at significance $\alpha = 0.05$.

So,

$H_0 : p_H = 1/2$

$H_a : p_H \neq 1/2$

at

$\alpha = 0.05$

We'll approximate the sampling distribution of our proportion of heads test statistic estimator with the standard normal distribution.

Let's start by creating our "critical region", aka, our "rejection region".

```
left_tail_cutoff <- qnorm(0.025)
left_tail_cutoff
```

```
## [1] -1.959964
```

```
right_tail_cutoff <- qnorm(0.975)
right_tail_cutoff
```

```
## [1] 1.959964
```

Our critical region is $(-\infty, -1.96]$ OR $[1.96, \infty)$

Now, calculate our test statistic

```
n <- nrow(sbdf)

xheads <- as.integer(sbdf[, "CoinToss"] %in% "Heads")

p_H0 <- 1/2

phat <- mean(xheads)
phat
```

```
## [1] 0.4821429
```

```
### OR

phat <- sum(xheads) / length(xheads)
phat
```

```
## [1] 0.4821429
```

```
var_phat <- p_H0 * (1 - p_H0)
var_phat
```

```
## [1] 0.25
```

```
SE <- sqrt(var_phat / n)
SE
```

```
## [1] 0.06681531
```

```
z_test <-  (phat - p_H0) / SE
z_test
```

```
## [1] -0.2672612
```

Using the normal approximation, our test statistic $z_{\text{test}} = -0.26726$ does not fall into our rejection region, so we fail to reject the hypothesis that Super Bowl coin tosses are fair.

Here's a more detailed conclusion: we fail to reject the hypothesis that within our imaginary, infinite target population of all possible Super Bowl coin tosses that could have occurred in the first 56 Super Bowls, the head-tails result is fair.

WNBA

Consider the WNBA seasons 2015-2021.

Suppose someone claims that the proportion of times a player scores 9 or more points in games in which they start is more than 50%.

Our population is starting player-game outcomes over 2015-2021.

We actually have that data, so we can directly calculate this proportion and accept of reject this claim with certainty.

Let's pretend we don't.

We'll use a sample of 100 starting player-game outcomes that was obtained by random selection, and test this claim at $\alpha = 0.01$.

$H_0 : p_{x \geq 9} = 1/2$

$H_a : p_{x \geq 9} > 1/2$

Let's start by creating our rejection region.

```r
right_tail_cutoff <- qnorm(0.99)
right_tail_cutoff
```

```
## [1] 2.326348
```

Our critical region is $[2.3263, \infty)$

```r
ydf_sample <- read.table("WNBA_starterGame_sample.tsv", sep="\t", header=TRUE)

xscore_sample <- as.integer( ydf_sample[ , "pts"] >= 9 )
```

```r
pH0 <- 1/2

n <- length(xscore_sample)
n
```

```
## [1] 100
```

```r
p_hat <- mean(xscore_sample)
p_hat
```

```
## [1] 0.68
```

```r
p_hat_var <- pH0 * (1 - pH0)
p_hat_var
```

```
## [1] 0.25
```

```r
SE <- sqrt( p_hat_var / n )
SE
```

```
## [1] 0.05
```

```r
z_test <- (p_hat - pH0) / SE
z_test
```

```
## [1] 3.6
```

Using the normal approximation, our test statistic $z_{\text{test}} = 3.6$ falls into our rejection region, so we conclude the occurrence of WNBA starters scoring 9 or more points is not 50%, but rather more.

Let's calculate the actual proportion.

```
ydf <- read.table("WNBA_playerGame.tsv", sep="\t", header=TRUE)

dim(ydf)
```

```
## [1] 26364    21
```

```
xmask_starters <- ydf[ , "strtr" ] %in% c(1,2,3,4,5)
ydf_starter <- ydf[ xmask_starters,  ]

dim(ydf_starter)
```

```
## [1] 13430    21
```

```
xscore <- as.integer( ydf_starter[ , "pts"] >= 9 )

xtrue_prop <- mean(xscore)

xtrue_prop
```

```
## [1] 0.6238273
```

It turns out our alternative is in fact true.

What's Alpha?

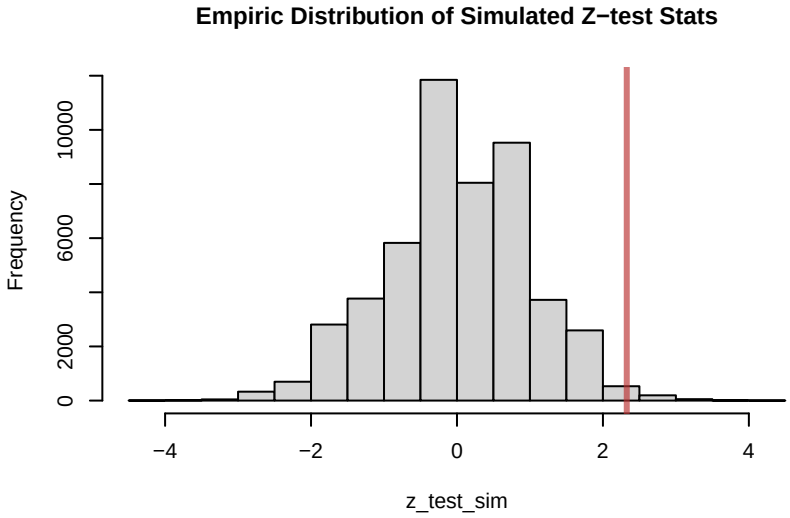
Suppose our null is true.

Let's simulate 50000 random samples and keep track of the test results of each simulation.

```
set.seed(777)

nn <- 50000 ### number of simulations

n <- 100

z_test_sim <- numeric(nn)

for(ii in 1:nn) {

  xrnd_ndx <- sample( I(1:nrow(ydf_starter)), size=n )

  ydf_starter_sample <- ydf_starter[ xrnd_ndx, ]

  p_hat_sim <- mean(as.integer( ydf_starter_sample[ , "pts"] >= 9 ))

  z_test_sim[ ii ] <- (p_hat_sim - xtrue_prop) / sqrt( xtrue_prop * (1 - xtrue_prop) / n )
```

```
}

xp_in_crit_region <- mean(z_test_sim > right_tail_cutoff)
xp_in_crit_region
```

```
## [1] 0.00934
```



Home Court Advantage

Although most sports leagues, including the WNBA, played at least partial seasons during the COVID-19 pandemic, many of these games were played in mostly empty arenas. This was especially the case in 2020.

Suppose someone says that the home court advantage in the WNBA during the pandemic seasons 2020 and 2021 was less than that of the previous seasons, 2015-2019.

We can measure "home team advantage" as the proportion of games won by the home team, and directly calculate this advantage for both groups of seasons.

```
sdf <- read.table("WNBA_game.tsv", sep="\t", header=TRUE)

xmask_before <- sdf[ , "season" ] %in% c(2015, 2016, 2017, 2018, 2019)
xmask_during <- sdf[ , "season" ] %in% c(2020, 2021)

n_before <- sum(xmask_before)
n_before
```

```
## [1] 1019
```

```
n_during <- sum(xmask_during)
n_during
```

```
## [1] 324
```

```
HW_before <- as.integer(sdf[ xmask_before, "HTscore" ] > sdf[ xmask_before, "VTscore" ])
p_hatHW_before <- mean(HW_before)
p_hatHW_before
```

```
## [1] 0.5839058
```

```
HW_during <- as.integer(sdf[ xmask_during, "HTscore" ] > sdf[ xmask_during, "VTscore" ])
p_hatHW_during <- mean(HW_during)
p_hatHW_during
```

```
## [1] 0.5216049
```

The home team advantage during the WNBA 2015-2019 seasons — before the pandemic — was 0.58391, whereas during the pandemic, it was 0.5216.

We can, of course, view this as an inferential question. We can view the actual games in our data set to be a randomly materialized subset from an abstract population of all possible games that could have resulted.

$H_0 : p_{\text{HWbefore}} = p_{\text{HWduring}}$

$H_a : p_{\text{HWbefore}} > p_{\text{HWduring}}$

Let's set our significance level to $\alpha = 0.05$.

```r
p_hat_pooled <- mean( c(HW_before, HW_during) )
p_hat_pooled
```

```
## [1] 0.5688757
```

```r
SE_pooled <- sqrt( p_hat_pooled * (1 - p_hat_pooled) * ( (1/n_before) + (1/n_during) ) )
SE_pooled
```

```
## [1] 0.03158554
```

```r
z_test <- ( p_hatHW_before - p_hatHW_during ) / SE_pooled
z_test
```

```
## [1] 1.972448
```

```r
p_val <- 1 - pnorm(z_test)
```

Our calculated p-value is 0.02428 which is less than our significance level of 0.05.

We therefore conclude the claim is correct, that home court advantage in the WNBA during the pandemic seasons 2020 and 2021 was reduced from that of prior seasons, 2015-2019.

Your Work

Make sure to edit the "author" information in the YAML header near the top to include your name and UID.

Complete/answer the following.

1 — Calculate the p-value for our Super Bowl coin tosses and comment on this value. How large is our target population?

```
# sbdf <- read.table("SuperBowl_coinTosses.tsv", sep="\t", header=TRUE)

print("The P-Value is .394735. As the P value is greater than .005 this is not significant. We accept t
```

```
## [1] "The P-Value is .394735. As the P value is greater than .005 this is not significant. We accept
```

2 — Show that we have met the requirements for using the normal approximation to the binomial model in our hypothesis test of starting player-game occurrence of 9-or-more point scoring. How large is our target population?

```
sbdf <- read.table("/Users/apurvashah/Documents/GitHub/stats10/lab9/yourLab/SuperBowl_coinTosses.tsv", s
left_tail_cutoff <- qnorm(0.025)
left_tail_cutoff
```

```
## [1] -1.959964
```

```
right_tail_cutoff <- qnorm(0.975)
right_tail_cutoff
```

```
## [1] 1.959964
```

```
n <- nrow(sbdf)
xheads <- as.integer(sbdf[, "CoinToss"] %in% "Heads")
p_H0 <- 1/2
phat <- mean(xheads)
phat
```

```
## [1] 0.4821429
```

```
### OR
phat <- sum(xheads) / length(xheads)
var_phat <- p_H0 * (1 - p_H0)
SE <- sqrt(var_phat / n)
z_test <-  (phat - p_H0) / SE

n*phat
```

```
## [1] 27
```

```
n*(1-phat)
```

```
## [1] 29
```

```
# As phat is from the binomial distribution, and both of these are greater than 10, these have met the
```

3 — Imagine, once again, we do not have access to the player-game WNBA population. Use a randomly generated sample of size 144 to test the claim that the proportion of bench players scoring 5 or more points in a game in which they play is greater than 1/3. Make sure to first check that we have met the requirements for using the normal approximation to the binomial model.

```
### here's a head start
ydf <- read.table("/Users/apurvashah/Documents/GitHub/stats10/lab9/yourLab/WNBA_playerGame.tsv", sep="\t
xmask_bench <- as.integer(ydf[ , "strtr" ]) > 5
ydf_bench <- ydf[ xmask_bench,  ]
dim(ydf_bench)
```

```
## [1] 12934    21
```

```
set.seed(777)
my_sample_df <- ydf_bench[ sample(1:nrow(ydf_bench), size=144), ]
xscore <- mean(my_sample_df[ , "pts"] >= 5)
xtrue_prop <- mean(xscore)

nn <- 144 ### number of simulations
n <- 100

pH0 <- 1/3
n <- length(xscore_sample)
p_hat <- mean(xscore_sample)
p_hat_var <- pH0 * (1 - pH0)
SE <- sqrt( p_hat_var / n )
z_test <- (p_hat - pH0) / SE
z_test
```

```
## [1] 7.353911
```

```
# Using the normal approximation, our test statistic ztest = 7.353911 falls into our rejection region,


#
# z_test_sim <- numeric(nn)
# for(ii in 1:nn) {
#   xrnd_ndx <- sample( I(1:nrow(ydf_bench)), size=n )
#   ydf_starter_sample <- ydf_bench[ xrnd_ndx, ]
#   p_hat_sim <- mean(as.integer( ydf_starter_sample[ , "pts"] >= 5 ))
#   z_test_sim[ ii ] <- (p_hat_sim - xtrue_prop) / sqrt( xtrue_prop * (1 - xtrue_prop) / n )
# }
#
# xp_in_crit_region <- mean(z_test_sim > right_tail_cutoff)
# xp_in_crit_region

## mean(my_sample_df[ , "tpm"] >= 2)
```

4 — Repeat Question 3, except test the claim that the proportion of bench players hitting 2 or more three-pointers in a game in which they play is less than 1/5.

```
ydf <- read.table("/Users/apurvashah/Documents/GitHub/stats10/lab9/yourLab/WNBA_playerGame.tsv", sep="\t
xmask_bench <- as.integer(ydf[ , "strtr" ]) > 5
ydf_bench <- ydf[ xmask_bench,  ]
dim(ydf_bench)
```

```
## [1] 12934    21
```

```
set.seed(777)
my_sample_df <- ydf_bench[ sample(1:nrow(ydf_bench), size=144), ]
xscore <- mean(my_sample_df[ , "tpm"] >= 2)
xtrue_prop <- mean(xscore)

nn <- 144 ### number of simulations
n <- 100

pH0 <- 1/5
n <- length(xscore_sample)
p_hat <- mean(xscore_sample)
p_hat_var <- pH0 * (1 - pH0)
SE <- sqrt( p_hat_var / n )
z_test <- (p_hat-pH0) / SE
z_test
```

```
## [1] 12
```

```
# Using the normal approximation, our test statistic ztest = 12 falls into our rejection region, so we
```

5 — In the Examples above, in the Section "What's Alpha?", we ran a number of simulations where our null was true. The proportion of times our calculated (simulated) test statistic fell into the critical region was 0.00934. Why is this value close to the significance level we set?

The alpha is the level of significance that we are testing. This value is close to the significance level that we set because the significance level determines the bounds of the curve that we are testing for. It would land in that area under the curve at the level of the alpha.

Furthermore, according to the central limit theorem, if the same size is large enough, the sample mean will follow a normal distribution around our true population, which is why the z score and probability from the test are near to significance level. When the significance level is reflected, we always reject the null, even if it is true; we want the significance level to be small, and the crucial area must be observed.