CAAP Statistics - Lec13

R Session5

Jul 26, 2022

Review

- Point estimates and sampling variability
  - What is sampling distribution
  - Central Limit Theorem
- Confidence intervals for a proportion
  - Interpreting the confidence interval
- Hypothesis testing for a proportion
  - Null hypothesis vs. Alternative hypothesis
  - Decision Error(Type I error, Type II error)

Learning Objectives

- Point estimates and sampling distribution
- Test Statistics
- Hypothesis testing and p-value

Load packages

```
library(openintro)
library(tidyverse)
library(ggplot2)
```

`yrbss` Data: Youth Risk Behaviour Surveillance System(YRBSS)

```
  head(yrbss)
## # A tibble: 6 × 13
##     age gender grade hispanic race     height weight helmet_12m text_while_driv…
##   <int> <chr>  <chr> <chr>    <chr>      <dbl>  <dbl> <chr>      <chr>
## 1    14 female 9     not      Black o…  NA       NA   never      0
## 2    14 female 9     not      Black o…  NA       NA   never      <NA>
## 3    15 female 9     hispanic Native …   1.73   84.4 never      30
## 4    15 female 9     not      Black o…   1.6    55.8 never      0
## 5    15 female 9     not      Black o…   1.5    46.7 did not r… did not drive
## 6    15 female 9     not      Black o…   1.57   67.1 did not r… did not drive
## # … with 4 more variables: physically_active_7d <int>,
## #   hours_tv_per_school_day <chr>, strength_training_7d <int>,
## #   school_night_hours_sleep <chr>
```

Define our quantity of Interest!

We want to know the "**proportion**" of students never wearing helmet while biking.

```
yrbss_upd = yrbss %>%
  mutate(helmet = ifelse(helmet_12m =="never",0,1))%>%
  drop_na()
(param = mean(yrbss_upd$helmet))
```

```
## [1] 0.4798228
```

Try sampling!

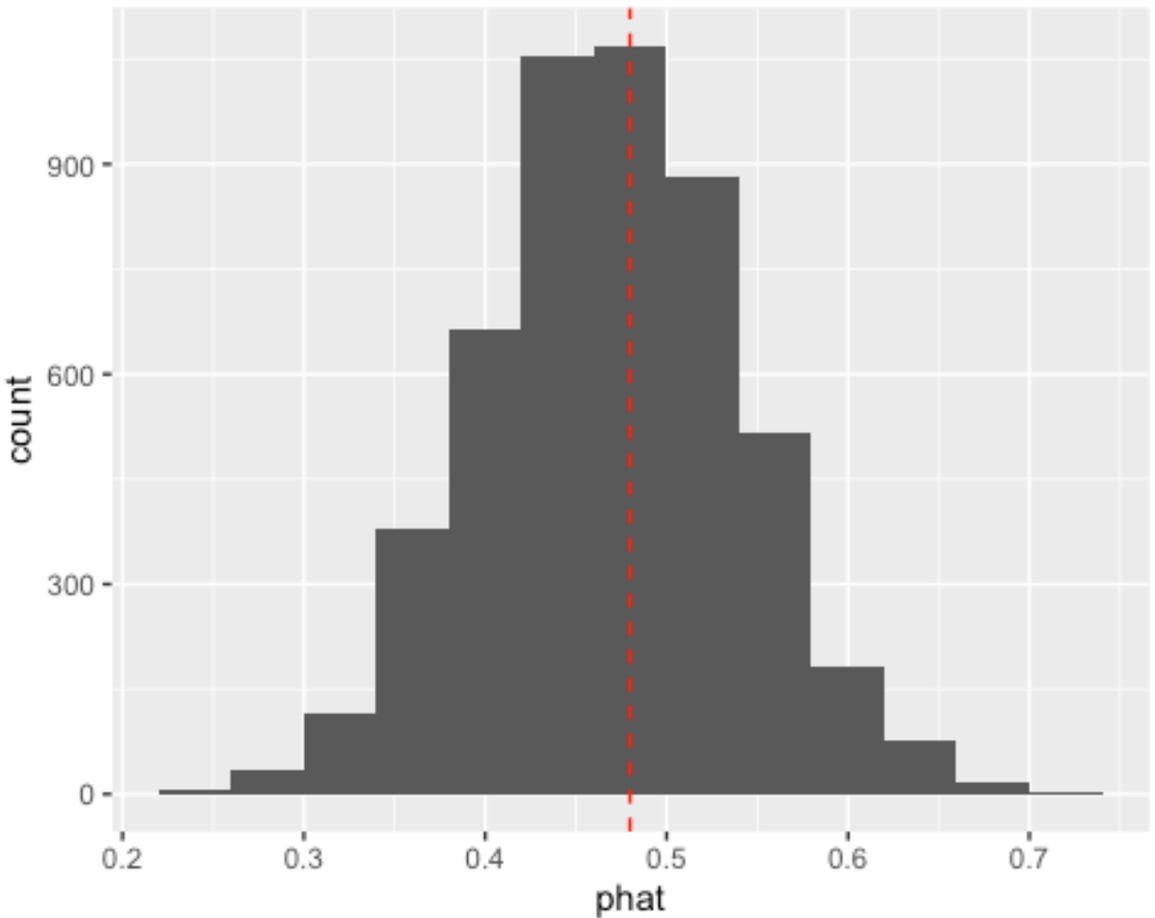
```
set.seed(12345)
mean(sample(yrbss_upd$helmet, size = 100, replace = TRUE))
```

```
## [1] 0.42
```

```
(phat = mean(sample(yrbss_upd$helmet, size = 100, replace = FALSE)))
```

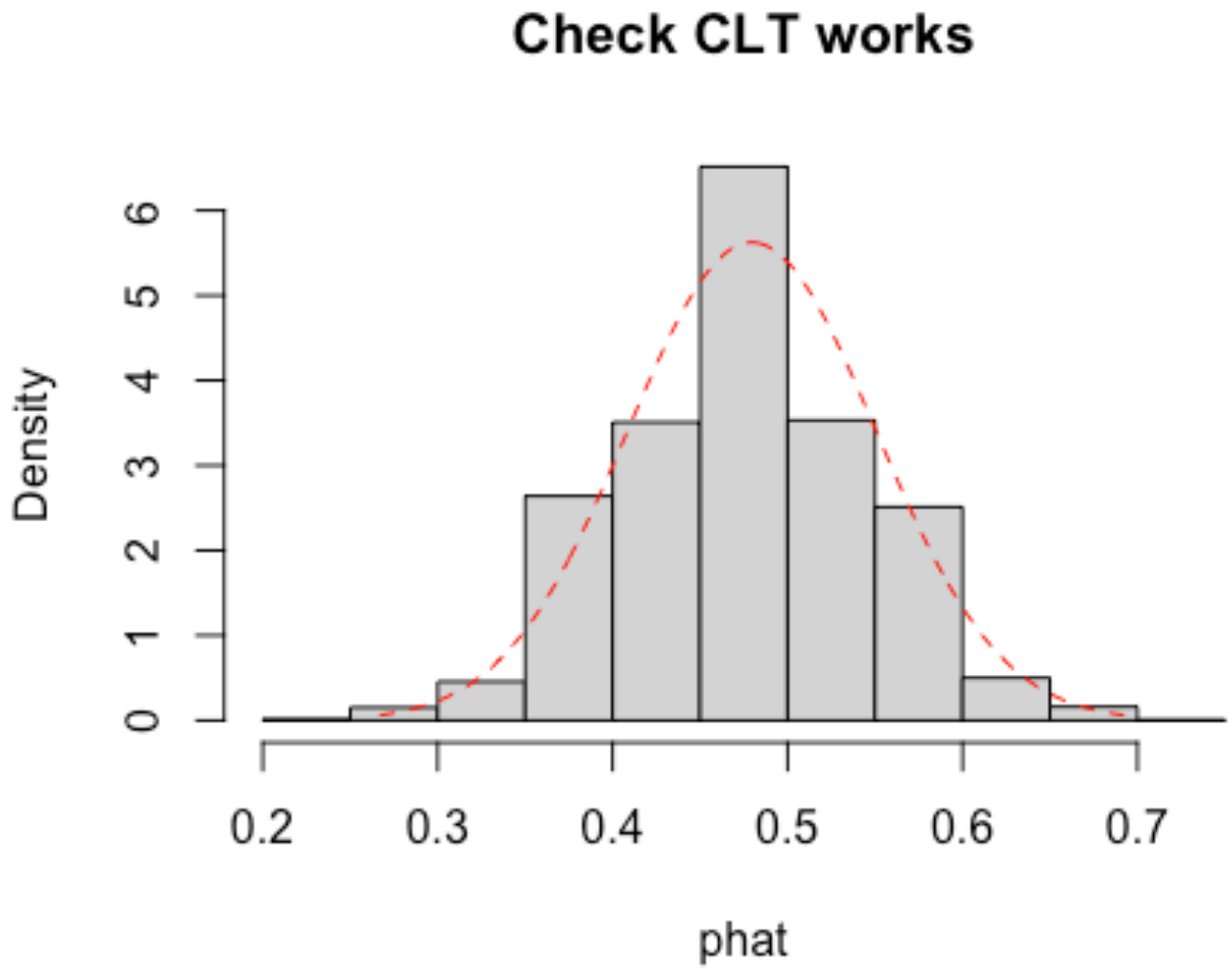
```
## [1] 0.48
```

```
sample_50 = data.frame(replicate(n= 5000, mean(sample(yrbss_upd$helmet, size = 50))))
colnames(sample_50) = "phat"
sample_50 %>%
  ggplot(aes(x=phat))+
  geom_histogram(binwidth = 0.04)+
  geom_vline(xintercept = param, color = "red", lty = 2)
```



Check if CLT works!

```
x = seq(-3,3, length = 500)
mu = mean(sample_50$phat)
sigma = sd(sample_50$phat)
hist(sample_50$phat, freq = FALSE, xlab="phat", main="Check CLT works")
lines(x*sigma+mu, dnorm(x*sigma+mu, mean = mu, sd = sigma),col="red",lty = 2)
```

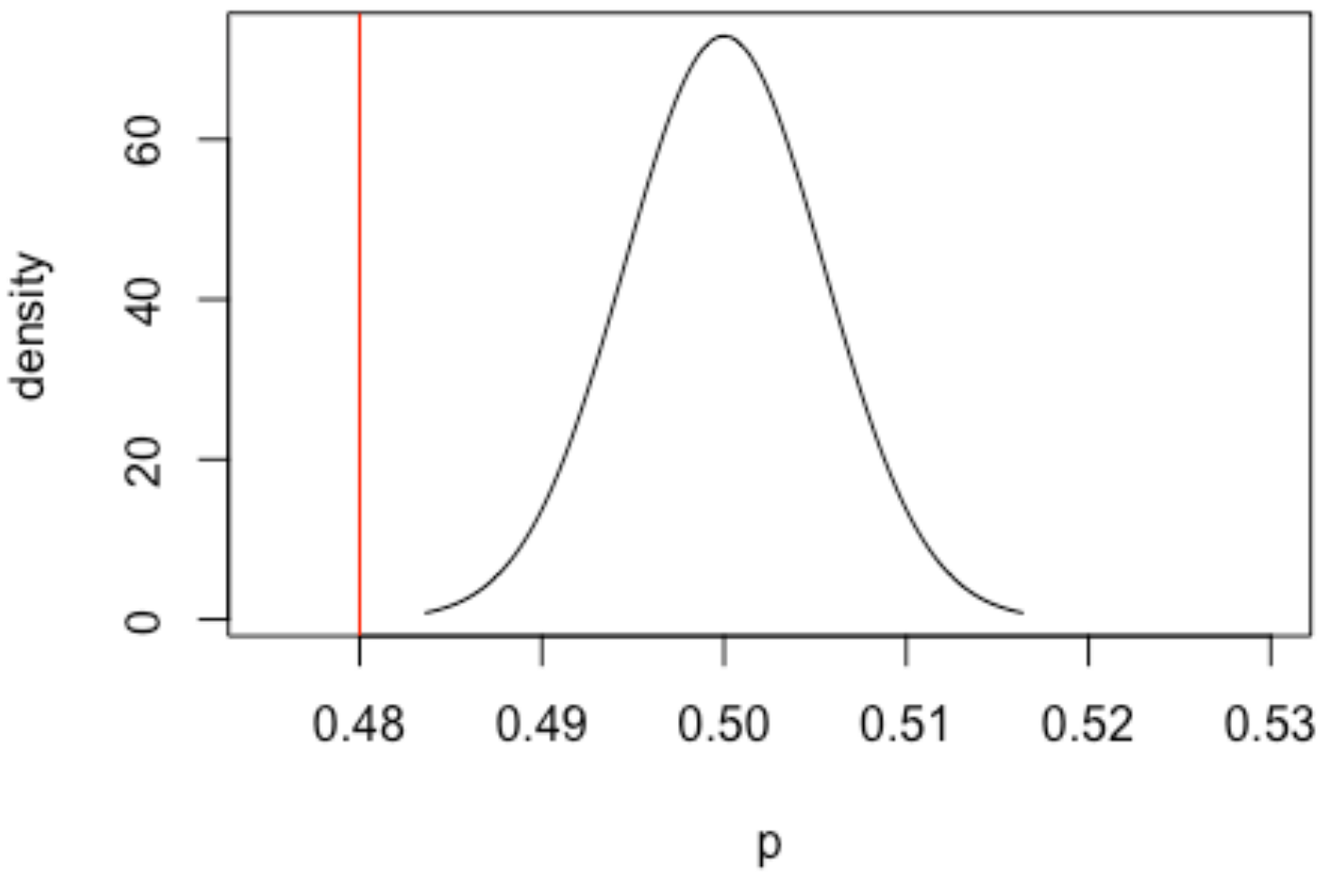


Hypothesis Testing

Does the majority of student wear a helmet?

- $H_0: p = 0.5$
- $H_1: p > 0.5$

```
se = sqrt((0.5*0.5)/nrow(yrbss_upd))
plot(x*se+0.5, dnorm(x*se+0.5,mean = 0.5, sd = se),type="l", xlab = "p",ylab="density",
xlim = c(0.475, 0.53))
abline(v = phat, col="red")
```



Calculate p-value

```
(zscore = (phat-0.5)/se)
## [1] -3.655352
2*pnorm(zscore) # H_1 : p \neq 0.5
## [1] 0.0002568291
1-pnorm(zscore) # H_1 : p > 0.5
## [1] 0.9998716
pnorm(zscore) # H_1: p < 0.5
## [1] 0.0001284146
```