STAT 3400 - Homework #7

Alex Ojemann

Due March 16, 2023

Problem 22.1

Suppose instead that your friend claimed that Venti drinks have less than 10 grams of protein, on average. Complete a hypothesis test of this claim.

$$H_0 : \mu = 10$$
$$H_A : \mu < 10$$

```
starbucks <- read.csv('data/starbucks_data.csv') %>%
  filter(Size=='Venti') %>%
  select(`Protein..g.`)
mean(starbucks$`Protein..g.`)
```

```
## [1] 9.055556
```

```
pnorm(q=mean(starbucks$`Protein..g.`),
      mean=10,
      sd=sd(starbucks$`Protein..g.`)/sqrt(nrow(starbucks)),
      lower.tail=TRUE)
```

```
## [1] 0.08497988
```

So, if the true mean grams of protein of Venti size drinks is 10, then there is only a 8.5% chance we would have observed a sample mean of 9.06 (or less). Our level of significance was 10%. Since our p-value is less than our level of significance we *reject* the null hypothesis in favor of the alternate hypothesis. It appears our friend was correct. Venti size drinks at Starbucks do average fewer than 10 grams of protein.

Problem 23.1

Suppose instead that our friend claimed that less than half of Venti drinks are under 40 grams of sugar. Complete a hypothesis test of this claim.

- Hypotheses

$$H_0 : p = 0.5$$
$$H_A : p < 0.5$$

- Level of significance

alpha = 0.1

- Test statistic

```
starbucks <- read.csv('data/starbucks_data.csv') %>%
  filter(Size=='Venti') %>%
  select(`Sugars..g.`)

mean(starbucks$`Sugars..g.`<40)
```

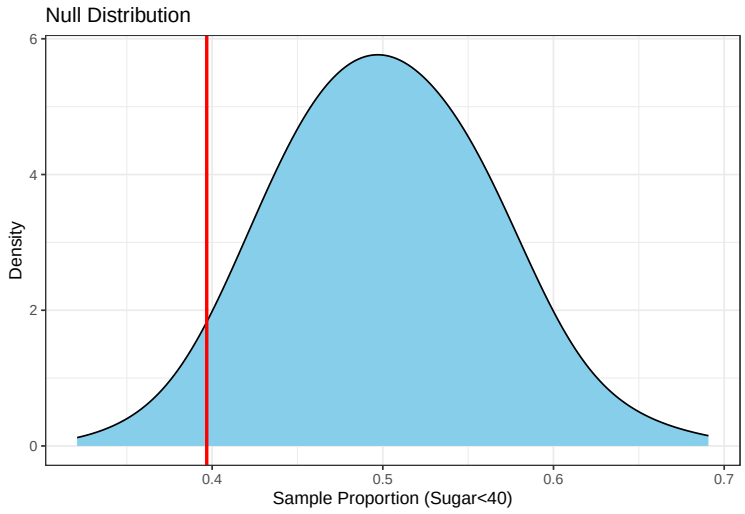
```
## [1] 0.3968254
```

- Null distribution

```
#calculate standard error
sterr <- sqrt((mean(starbucks$`Sugars..g.`<40)*(1-mean(starbucks$`Sugars..g.`<40)))/nrow(starbucks))

#simulate null distribution by drawing from Normal distribution
set.seed(653)
null_dist <- data.frame(cals=rnorm(n=1000,mean=0.5,sd=sterr))

#plot null distribution along with test statistic
ggplot(data=null_dist,aes(x=cals)) +
  geom_density(fill='sky blue',adjust=2) +
  geom_vline(xintercept=mean(starbucks$`Sugars..g.`<40),color='red',size=1) +
  labs(title='Null Distribution',
       x='Sample Proportion (Sugar<40)',
       y='Density') +
  theme_bw()
```



- p-value

```
#calculate p-value
pnorm(q=mean(starbucks$`Sugars..g.`<40),
      mean=0.5,
      sd=sterr,
      lower.tail=TRUE)
```

```
## [1] 0.04707804
```

- Conclusion

So, if the true proportion of Venti size drinks with more than 40 grams of sugar is 50%, then there is a 0.047 chance we would have observed a sample proportion of 40% (or less). Our level of significance was 0.10. Since our p-value is less than our level of significance we *reject* the null hypothesis. There is statistically significant evidence to support our friend's claim. It does appear that more than 50% of Venti drinks have more than 40 grams of sugar.

Problem 23.2

Suppose instead that our friend claimed the proportion of soy milk drinks with more than 8 grams of protein is less than the same proportion for nonfat milk drinks. Complete a hypothesis test of this claim.

- Hypotheses

$$H_0 : p_1 - p_2 = 0$$
$$H_A : p_1 - p_2 < 0$$

- Level of significance

alpha = 0.05

- Test statistic

```
#import and filter data
starbucks <- read.csv('data/starbucks_data.csv') %>%
  filter(Beverage_prep %in% c('Nonfat Milk','Soymilk')) %>%
  select(Beverage_prep,`Protein..g.`)

#summarize group size and proportion by milk prep
coffee_sum <- starbucks %>%
  group_by(Beverage_prep) %>%
  summarize(count=n(),
            prop=mean(`Protein..g.`>8)) %>%
  ungroup()

#calculate difference in proportions
test_stat <- coffee_sum[2,3]-coffee_sum[1,3]
test_stat
```

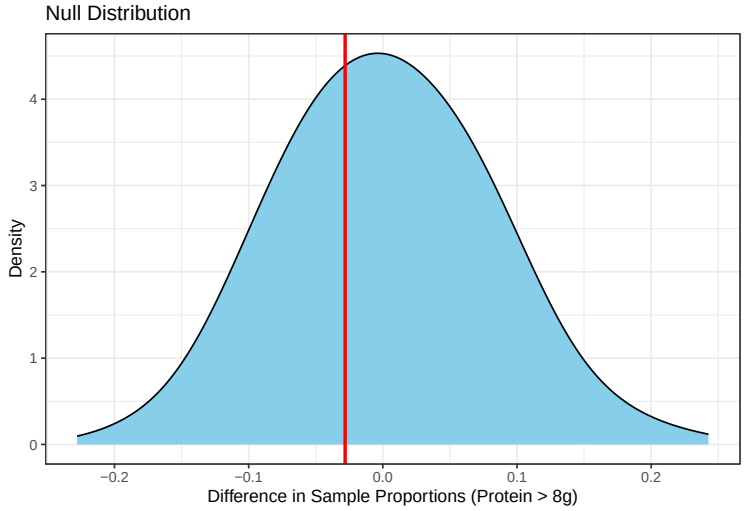
```
##          prop
## 1 -0.02811245
```

- Null distribution

```
#calculate standard error
sterr2 <- sqrt((coffee_sum[2,3]*(1-coffee_sum[2,3])/coffee_sum[2,2])+
               (coffee_sum[1,3])*(1-coffee_sum[1,3])/coffee_sum[1,2])

#simulate null distribution by drawing from Normal distribution
set.seed(653)
null_dist2 <- data.frame(cals=rnorm(n=1000,mean=0,sd=sterr2$prop))

#plot null distribution along with test statistic
ggplot(data=null_dist2,aes(x=cals)) +
  geom_density(fill='sky blue',adjust=2) +
  geom_vline(xintercept=test_stat$prop,color='red',size=1) +
  labs(title='Null Distribution',
       x='Difference in Sample Proportions (Protein > 8g)',
       y='Density') +
  theme_bw()
```



- p-value

```
#calculate p-value
pnorm(q=test_stat$prop,
      mean=0,
      sd=sterr2$prop,
      lower.tail=TRUE)
```

```
## [1] 0.3599688
```

- Conclusion

If there is truly no difference between the proportions of soy and nonfat milk drinks, then there is a 0.36 chance we would have observed a difference of 2.8 percentage-points (or greater). Our level of significance was 0.05. Since our p-value is greater than our level of significance we *fail to reject* the null hypothesis. There is no statistically significant evidence to support our friend's claim. It does *not* appear there is any difference between soy and nonfat milk drinks in terms of exceeding 8 grams of protein.

Problem 24.1

Construct a 95% confidence interval for the true proportion of NBA teams that have fewer than 8 offensive rebounds in a game. Then answer the following questions:

```
nba <- read.csv('data/nba_data.csv') %>%
  transmute(orb=ifelse(OREB<8,1,0))
#compute sample proportion
point <- mean(nba$orb)
point
```

```
## [1] 0.2302302
```

```
#compute quantile of standard normal distribution
critical <- qnorm(p=0.025,mean=0,sd=1,lower.tail=FALSE)
critical
```

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

```
#compute standard error of sample proportion
sterr <- sqrt(point*(1-point)/nrow(nba))
sterr
```

```
## [1] 0.01331922
```

```
#construct confidence interval
lower <- point-critical*sterr
upper <- point+critical*sterr
lower
```

```
## [1] 0.204125
```

```
upper
```

```
## [1] 0.2563354
```

- What is the correct interpretation of this interval in the context of the problem?

We are 95% confident that the true proportion of times a team grabs fewer that 8 offensive rebounds in a game is between 0.204 and 0.256.

- TRUE or FALSE: If we collect a different sample of games, we are guaranteed to get the same confidence bounds.

False

- If we decrease the sample size, the interval width will increase.
- If we increase the confidence level, the interval width will increase.

Problem 24.2

Returning to your data from the previous practice, construct a 95% upper confidence bound for the proportion of teams that achieve fewer than 8 offensive rebounds in a game. Then answer the following questions:

```
nba <- read.csv('data/nba_data.csv') %>%
  transmute(orb=ifelse(OREB<8,1,0))
#compute sample proportion
point <- mean(nba$orb)
point
```

```
## [1] 0.2302302
```

```
#compute quantile of standard normal distribution
critical <- qnorm(p=0.05,mean=0,sd=1,lower.tail=FALSE)
critical
```

```
## [1] 1.644854
```

```
#compute standard error of sample proportion
sterr <- sqrt(point*(1-point)/nrow(nba))
sterr
```

```
## [1] 0.01331922
```

```
#construct confidence interval
upper <- point+critical*sterr
upper
```

```
## [1] 0.2521384
```

- What is the correct interpretation of this bound in the context of the problem?

We are 95% confident that the true proportion of times a team grabs fewer that 8 offensive rebounds in a game is below 0.252.

- TRUE or FALSE: The one-sided upper bound will always be less than the two-sided upper bound.

True

- For a fixed sample size, the standard error is largest when $\hat{p}$ equals 0.5.