Assignment 8

Due Friday, October 30, 11:59 PM CT

Kyle yeo

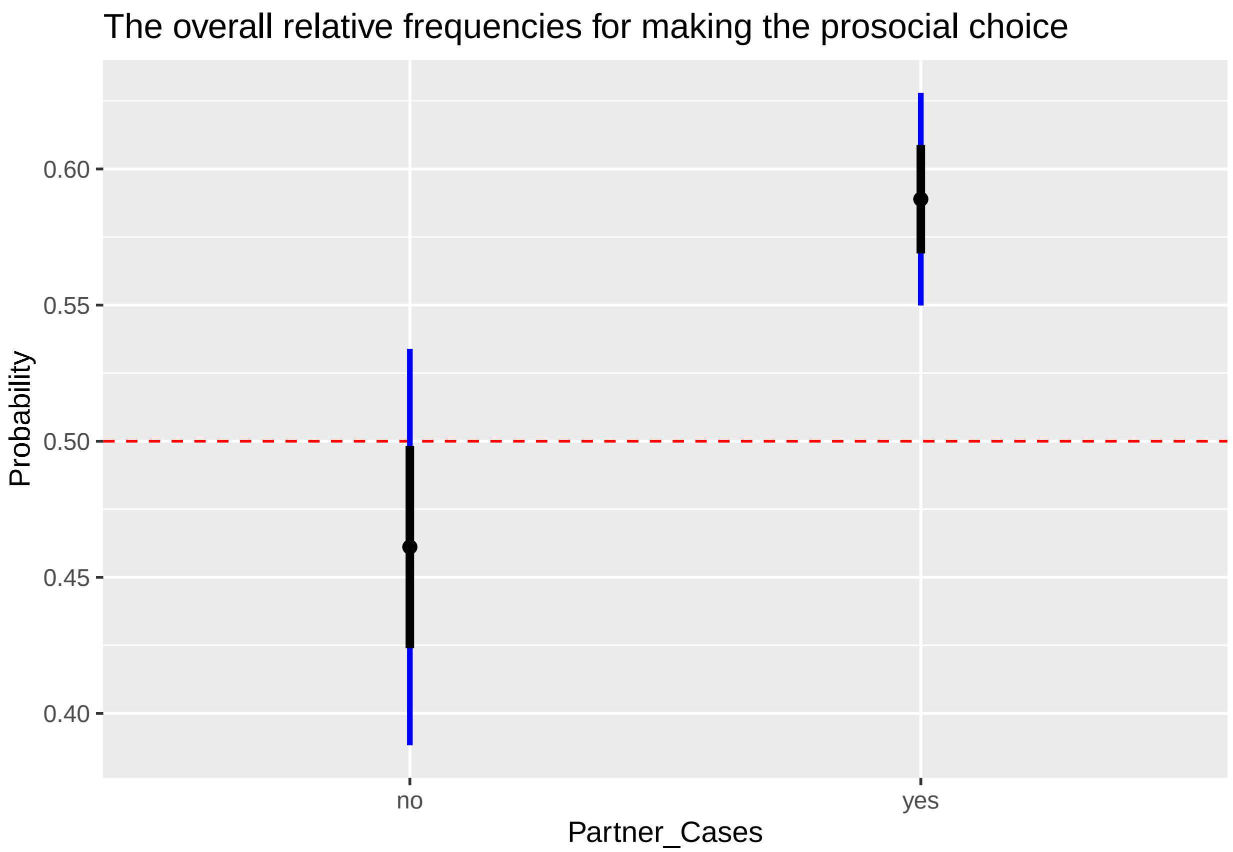
Problems

1

> Read in the `chimpanzee.csv` data file. Make a plot that displays the overall relative frequencies for making the prosocial choice; do this separately for the trials when a partner is present and when there is no partner present. (That is, plot a point estimate of the sample proportion for the trials with and without a parter.)
> Use a thin blue line segment to visualize a 95% confidence interval, a slightly thicker black segment to visualize the interval one standard error above and below the point estimate, and a point at the point estimate. Add a horizontal red dashed line at p = 0.5. Label axes appropriately and add an informative title to the plot.

Hint: Your final plot should have two vertical lines (with the layered segments noted in the question), a point in the center of each vertical line, and a horizontal line.

```r
chimp <- read_csv("C:/stat_240/data/chimpanzee.csv")

chimp1 <- chimp %>%
  mutate(partner_present = ifelse(partner == "none", "no", "yes"),
         p_hat = prosocial/(prosocial+selfish),
         n = prosocial+selfish) %>%
  group_by(partner_present) %>%
  summarise(p_hat = mean(p_hat), n=sum(n)) %>%
  mutate(ci_95_left = p_hat - 1.96*sqrt(p_hat*(1-p_hat)/n),
         ci_95_right = p_hat + 1.96*sqrt(p_hat*(1-p_hat)/n),
         se_left = p_hat - sqrt(p_hat*(1-p_hat)/n),
         se_right = p_hat + sqrt(p_hat*(1-p_hat)/n))


ggplot(chimp1, aes(x=partner_present, y=p_hat)) +
  geom_point(size=2) +
  geom_hline(yintercept = 0.5, color = "red", linetype = "dashed") +
  geom_segment(data = chimp1, mapping = aes(x=partner_present, xend = partner_present, y = ci_95_left, 
  geom_segment(data = chimp1, mapping = aes(x= partner_present, xend = partner_present, y = se_left, ye
  xlab("Partner_Cases") +
  ylab("Probability") +
  ggtitle("The overall relative frequencies for making the prosocial choice")
```



2

Consider Chimpanzee actor F in the setting with a partner present. Compute 99%, 95%, 90%, and 80% Wald confidence intervals for p, the probability of selecting the prosocial token. Print out all four confidence intervals.

```r
binom_se <-  function(n,p)
{
  return ( sqrt( p*(1-p)/n) )
}

binom_ci <- function(est,se,conf)
{
  z <- qnorm(1 - (1 - conf)/2)
  me <- z * se
  ci <- est + c(-1,1)*me
  return(ci)
}



chimp2 <- chimp %>%
  mutate(partner_present = ifelse(partner == "none", "no", "yes"),
         p_hat = prosocial/(prosocial+selfish),
         n = prosocial+selfish) %>%
  filter(actor == "F" & partner_present == "yes") %>%
  group_by(actor)%>%
  summarise(p_hat = mean(p_hat), n = sum(n))




se_wald = binom_se(chimp2$n,chimp2$p_hat)

binom_ci(chimp2$p_hat,se_wald,0.99)
```

```
## [1] 0.3865982 0.6578462
```

```
binom_ci(chimp2$p_hat,se_wald,0.95)
```

```
## [1] 0.4190251 0.6254193
```

```
binom_ci(chimp2$p_hat,se_wald,0.90)
```

```
## [1] 0.4356165 0.6088280
```

```
binom_ci(chimp2$p_hat,se_wald,0.80)
```

```
## [1] 0.4547453 0.5896992
```

3

> Summarize the full chimpanzee data set with a data frame that has one row for each actor chimpanzee (A-G), and columns for the variables listed below. Print out the final data frame. Note: the variables below are for the trials *with* a partner.

- `n`, the number of trials with a partner
- `prosocial`, the number of prosocial choices with a partner
- `selfish`, the number of selfish choices with a partner
- `p_hat`, the observed proportion of prosocial choices in trials with a partner
- `se_wald`, the estimated standard error using `p_hat` and `n`
- `a_wald`, the lower boundary of the Wald 90% confidence interval
- `b_wald`, the upper boundary of the Wald 90% confidence interval
- `p_tilde`, the Agresti-Coull point estimate of `p`
- `se_agresti`, the estimated standard error from the Agresti-Coull method
- `a_agresti`, the lower boundary of the Agresti-Coull 90% confidence interval
- `b_agresti`, the upper boundary of the Agresti-Coull 90% confidence interval

```
chimp3 <- chimp %>%
  mutate(partner_present = ifelse(partner == "none", "no", "yes")) %>%
  filter(partner_present != "no") %>%
  group_by(actor) %>%
  summarise(prosocial = sum(prosocial),
            selfish = sum(selfish),
            n = sum(prosocial+selfish),
            p_hat = prosocial/(prosocial+selfish),
            se_wald = sqrt(p_hat*(1-p_hat)/n),
            a_wald = p_hat-1.645*sqrt(p_hat*(1-p_hat)/n),
            b_wald =p_hat+ 1.645*sqrt(p_hat*(1-p_hat)/n),
            p_tidle = (prosocial+2)/(prosocial+selfish+4),
            se_agresti = sqrt(p_tidle*(1-p_tidle)/(n+4)),
            a_agresti = p_tidle - 1.645*se_agresti,
            b_agresti = p_tidle + 1.645*se_agresti
            )

chimp3
```

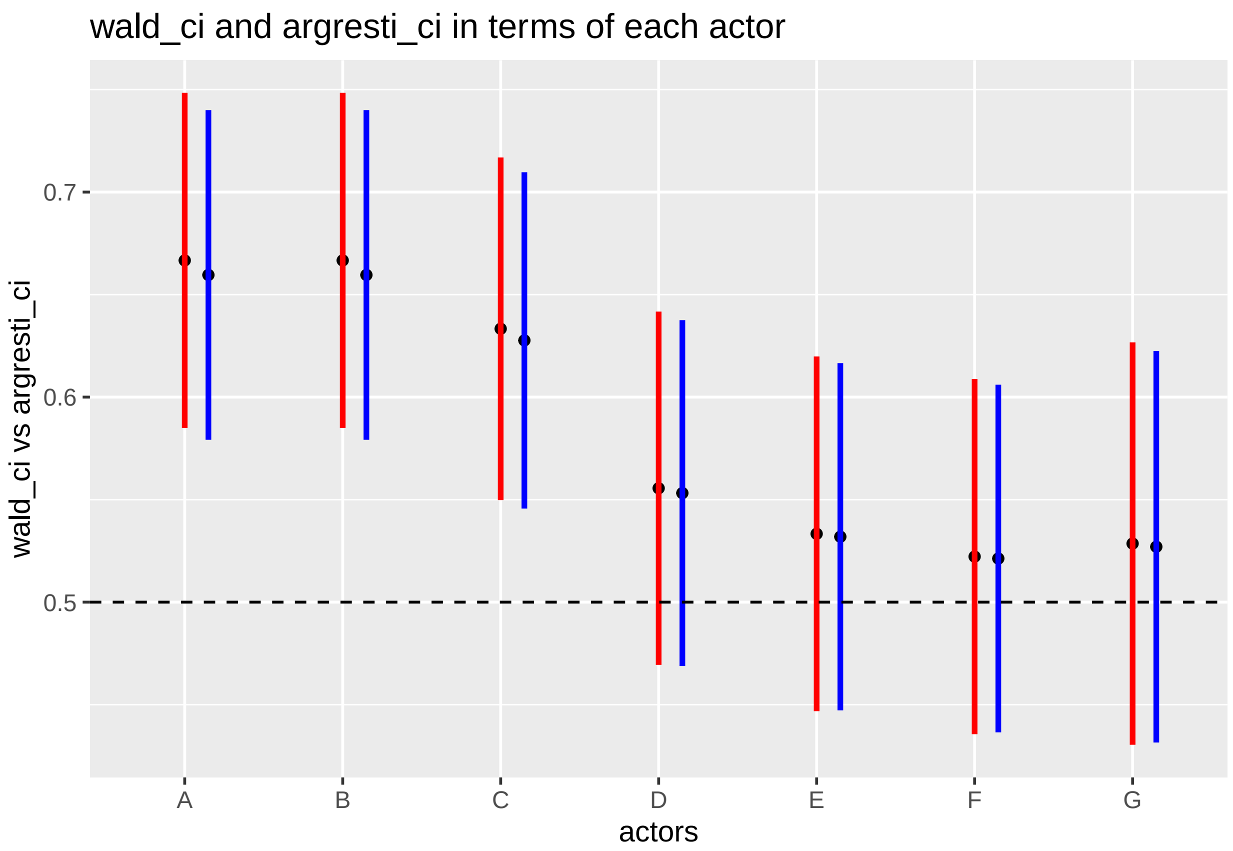
```
## # A tibble: 7 x 12
##   actor prosocial selfish     n p_hat se_wald a_wald b_wald p_tidle se_agresti
##   <chr>     <dbl>   <dbl> <dbl> <dbl>   <dbl>  <dbl>  <dbl>   <dbl>      <dbl>
## 1 A            60      30    90 0.667  0.0497  0.585  0.748   0.660     0.0489
## 2 B            60      30    90 0.667  0.0497  0.585  0.748   0.660     0.0489
## 3 C            57      33    90 0.633  0.0508  0.550  0.717   0.628     0.0499
## 4 D            50      40    90 0.556  0.0524  0.469  0.642   0.553     0.0513
## 5 E            48      42    90 0.533  0.0526  0.447  0.620   0.532     0.0515
## 6 F            47      43    90 0.522  0.0527  0.436  0.609   0.521     0.0515
## 7 G            37      33    70 0.529  0.0597  0.430  0.627   0.527     0.0580
## # ... with 2 more variables: a_agresti <dbl>, b_agresti <dbl>
```

4

> Using the table from Question 3, make a plot with two line segments for each actor chimpanzee, one displaying the Wald 90% confidence interval and one for the Agresti 90% confidence interval. Add a point representing the point estimate to each interval. Plot the line segments for each actor close to each other for easy comparison. Add a horizontal black dashed line at $p = 0.5$. Label axes appropriately and add an informative title to the plot.

Hint: The `actor` values are strings, which end up getting plotted at 1, 2, . . . , 7 on the x-axis (but labeled as the actor's letter A:G). In order to plot the Wald and Agresti confidence intervals for each actor near each other (and not overlapping), you may consider using `as.integer(as.factor(actor))` plus some small number (e.g., 0.15) to move the x-axis values for one of the intervals slightly. The result is for one interval (e.g., Wald) to be plot at the integer values (1:7), and the other interval (e.g., Agresti-Coull) to be plotted at 1.15, 2.15, . . . , 7.15.

```
ggplot()+
  geom_point(chimp3, mapping = aes(x = actor, y = p_hat))+
  geom_point(chimp3, mapping = aes(x = as.integer(as.factor(actor))+0.15, y = p_tidle))+
  geom_segment(data = chimp3, mapping = aes(x = actor, xend= actor, y = a_wald, yend = b_wald), color =
  geom_segment(data = chimp3, mapping = aes(x = as.integer(as.factor(actor))+0.15, xend= as.integer(as.
  geom_hline(yintercept = 0.5, color = "black", linetype = "dashed") +
  ylab("wald_ci vs argresti_ci")+
  xlab("actors")+
  ggtitle("wald_ci and argresti_ci in terms of each actor")
```



5

Repeat Problem 3 for the data on the trials without partners present. Note that only six of the seven chimpanzees had trials without partners.

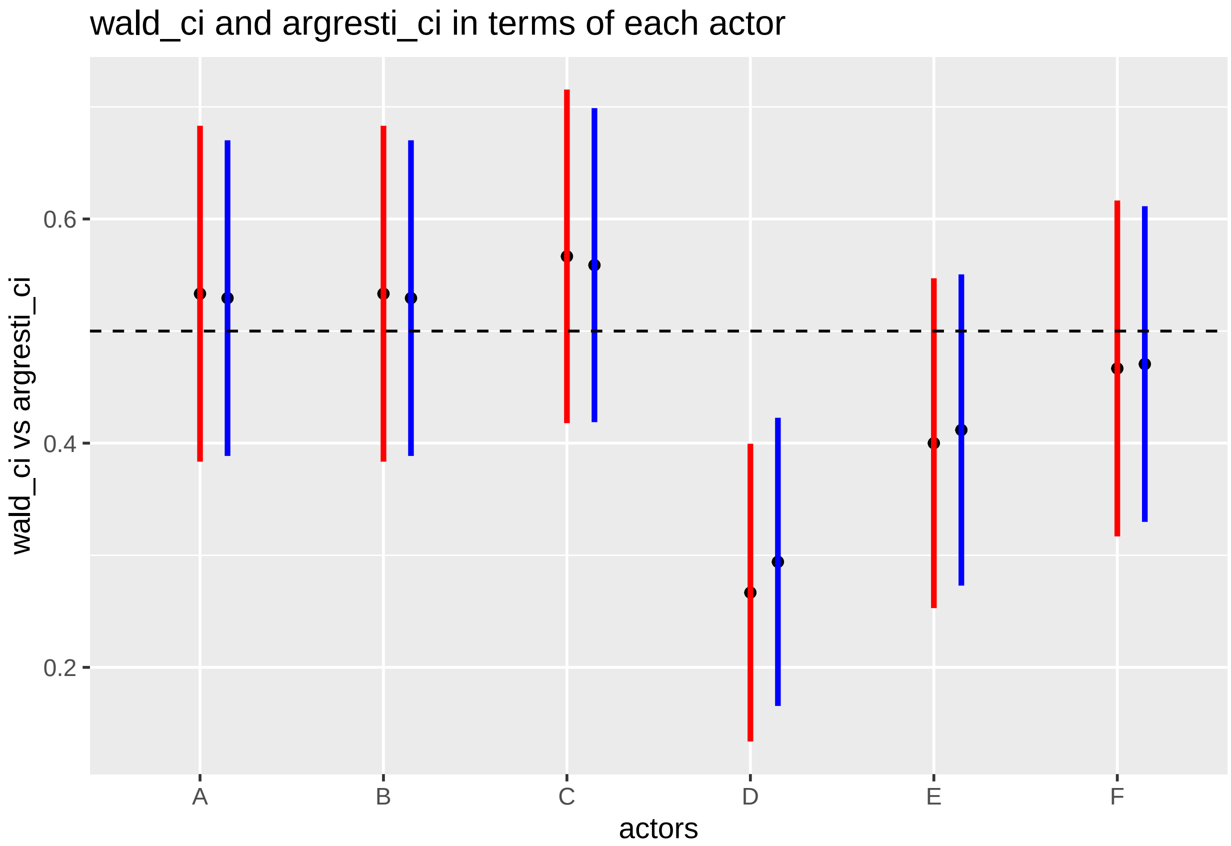
```
chimp5 <- chimp %>%
  mutate(partner_present = ifelse(partner == "none", "no", "yes")) %>%
  filter(partner_present == "no") %>%
  group_by(actor) %>%
  summarise(prosocial = sum(prosocial),
            selfish = sum(selfish),
            n = sum(prosocial+selfish),
            p_hat = prosocial/(prosocial+selfish),
            se_wald = sqrt(p_hat*(1-p_hat)/n),
            a_wald = p_hat-1.645*sqrt(p_hat*(1-p_hat)/n),
            b_wald =p_hat+ 1.645*sqrt(p_hat*(1-p_hat)/n),
            p_tidle = (prosocial+2)/(prosocial+selfish+4),
            se_agresti = sqrt(p_tidle*(1-p_tidle)/(n+4)),
            a_agresti = p_tidle - 1.645*se_agresti,
            b_agresti = p_tidle + 1.645*se_agresti
            )

chimp5
```

```
## # A tibble: 6 x 12
##   actor prosocial selfish     n p_hat se_wald a_wald b_wald p_tidle se_agresti
##   <chr>     <dbl>   <dbl> <dbl> <dbl>   <dbl>  <dbl>  <dbl>   <dbl>      <dbl>
## 1 A             16       14     30 0.533  0.0911  0.384  0.683   0.529      0.0856
## 2 B             16       14     30 0.533  0.0911  0.384  0.683   0.529      0.0856
## 3 C             17       13     30 0.567  0.0905  0.418  0.715   0.559      0.0852
## 4 D              8       22     30 0.267  0.0807  0.134  0.399   0.294      0.0781
## 5 E             12       18     30 0.4    0.0894  0.253  0.547   0.412      0.0844
## 6 F             14       16     30 0.467  0.0911  0.317  0.616   0.471      0.0856
## # ... with 2 more variables: a_agresti <dbl>, b_agresti <dbl>
```

6

Repeat Problem 4 for the data for the trials without partners (using your data frame from question 5).

```r
ggplot()+
  geom_point(chimp5, mapping = aes(x = actor, y = p_hat))+
  geom_point(chimp5, mapping = aes(x = as.integer(as.factor(actor))+0.15, y = p_tidle))+
  geom_segment(data = chimp5, mapping = aes(x = actor, xend= actor, y = a_wald, yend = b_wald), color =
  geom_segment(data = chimp5, mapping = aes(x = as.integer(as.factor(actor))+0.15, xend= as.integer(as.
  geom_hline(yintercept = 0.5, color = "black", linetype = "dashed") +
  ylab("wald_ci vs argresti_ci")+
  xlab("actors")+
  ggtitle("wald_ci and argresti_ci in terms of each actor")
```



7

Suppose we computed a 90% confidence interval for the proportion of times one of the actor chimpanzees, say Chimpanzee A, selected the prosocial token to be [0.585, 0.748]. Can we say

> that there is a 90% probability that the interval [0.585, 0.748] contains the true proportion for selecting the prosocial token? Briefly explain your answer.

A confidence interval refers to the probability that a population parameter will fall between a set of values for a certain proportion of times, thus according to this definition, we cannot say that there is a 90% probability that the interval contains the true proportion for selecting the prosocial token. (90% means 90% of the repeated sampling, not 90% for individual interval)

8

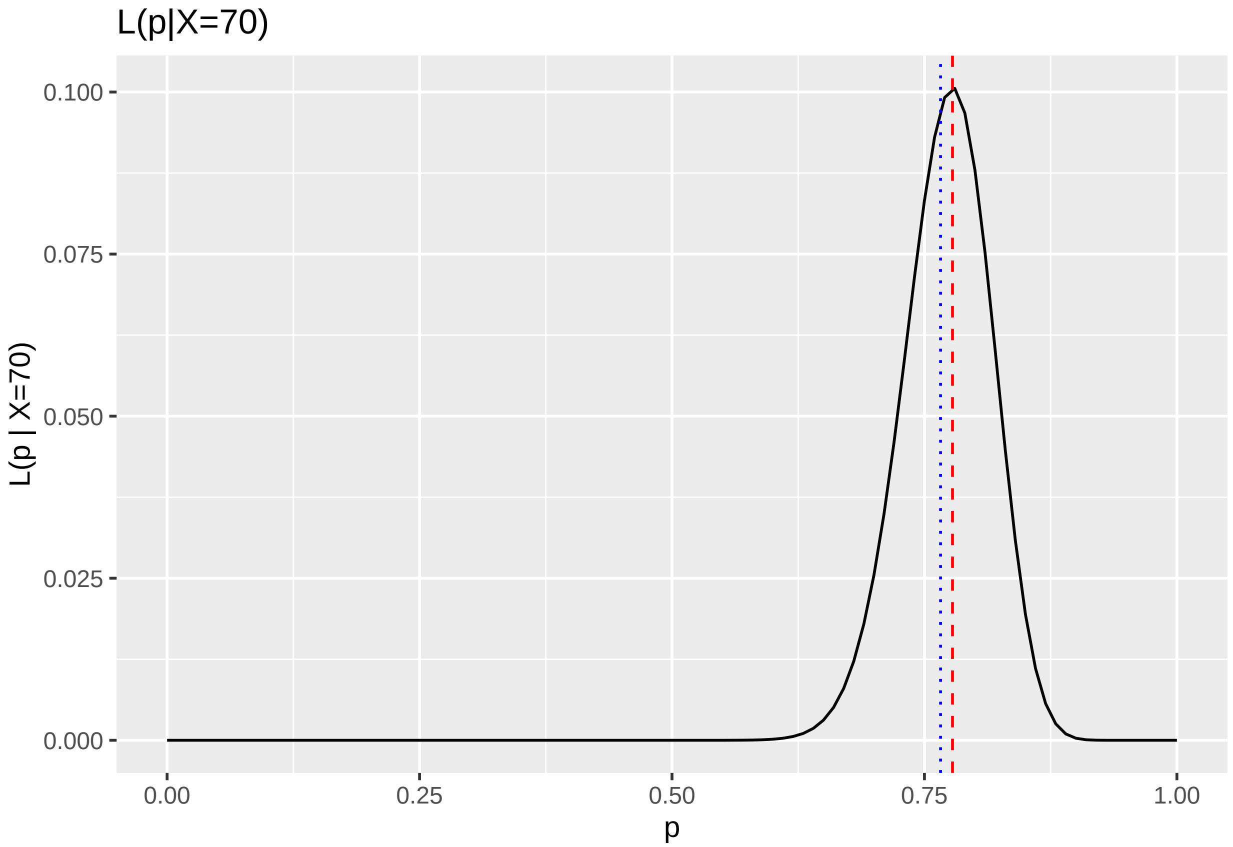
> Suppose we carried out 10,000 new experiments for Chimpanzee A in the setting where a partner was present, each with $n = 90$ trials, and created a 90% confidence interval from each one. (So you end up with 10,000 confidence intervals.) Approximately how many of those 10,000 confidence intervals do you expect to contain the true proportion for Chimpanzee A selecting the prosocial token?

9000

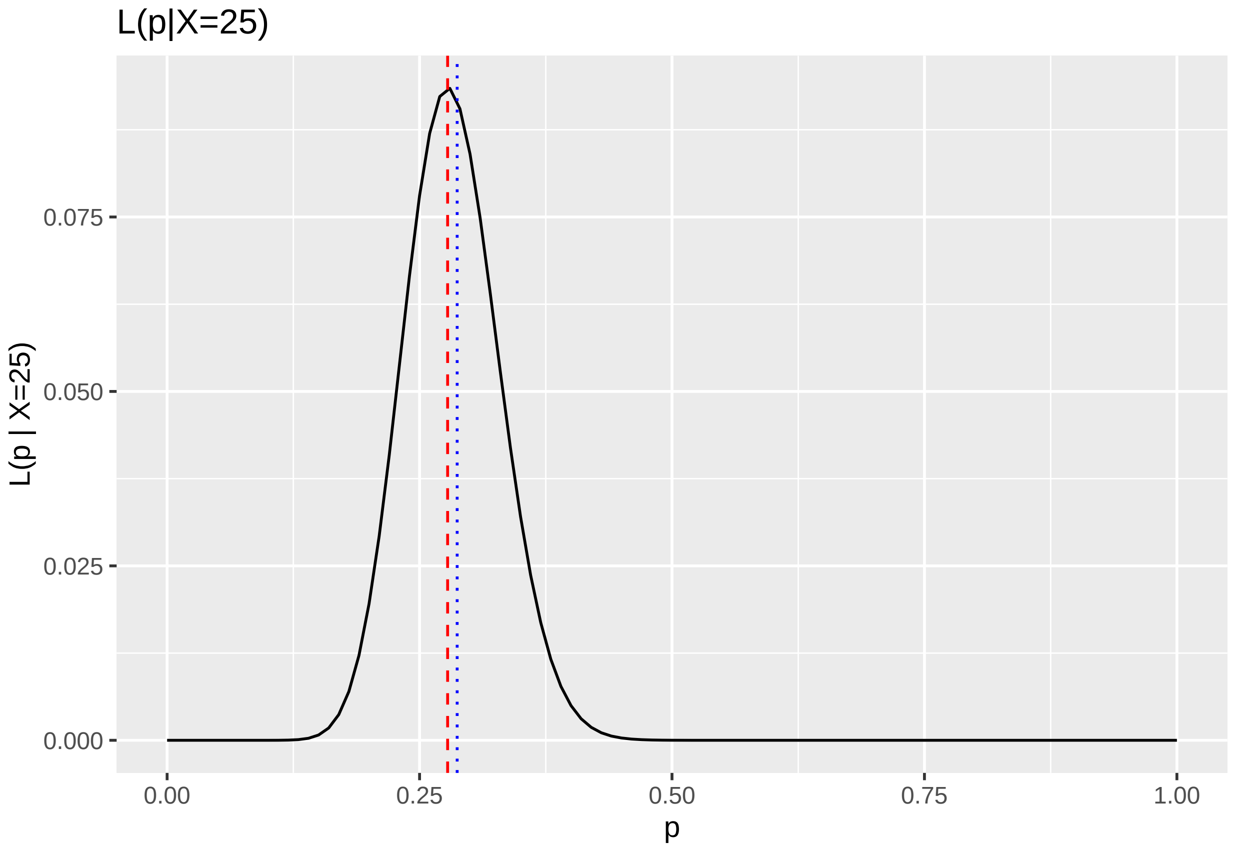
9

> Consider a Binomial random variable X ~ Binomial(90, p). Create a plot of the likelihood function L(p | X=x) if you observe x = 70, that is plot L(p | X=70). Add a red vertical dashed line at the maximum likelihood estimate, and a blue vertical dotted line at the Agresti-Coull estimate. Create a second plot of the likelihood if you observe x = 25, that is, plot L(p | X=25). Add meaningful axis labels and title.
> For each of the two plots, what is the relationship between the two point estimates (e.g., is the Wald greater than the Agresti-Coull estimate? Or vice versa?)? Why is this happening?

```r
n <- 90
x <- 70
df1 <- tibble(pseq=seq(0,1,by=.01), L=dbinom(x,n,pseq))

ggplot(df1, aes(pseq, L)) +
  geom_line() +
  xlab("p") +
  ylab("L(p | X=70)") +
  ggtitle(paste0("L(p|X=",x,")")) +
  geom_vline(xintercept=x/n, color="red",linetype="dashed") + #wald
  geom_vline(xintercept= (x+2)/(n+4), color="blue", linetype = "dotted") #agresti
```



```
n <- 90
x <- 25
df1 <- tibble(pseq=seq(0,1,by=.01), L=dbinom(x,n,pseq))

ggplot(df1, aes(pseq, L)) +
  geom_line() +
  xlab("p") +
  ylab("L(p | X=25)") +
  ggtitle(paste0("L(p|X=",x,")")) +
  geom_vline(xintercept=x/n, color="red",linetype="dashed") + #wald
  geom_vline(xintercept= (x+2)/(n+4), color="blue", linetype = "dotted") #agresti
```



When a plot of the likelihood function is left_skewed(when x = 70), the Wald is greater than the Agresti-Coull estimate, while when a plot is right_skewed(when x = 25), the Agresti-Coull is greater than the Wald. This type of pattern has shown since regardless of the value of X, maximum value of the likelihood is at our p_hat = x/n.