Assignment 9

Due Friday, November 6, 11:59 PM CT

Kyle Yeo

Problems

1

> Read in the `chimpanzee.csv` data file. Consider only those trials with a partner. Make an assumption that there is a universal p_{partner} during which any chimpanzee would make a prosocial choice in a single trial under the experimental conditions we have been examining. Assume that all trials are independent. Under these assumptions, write down a statistical model for X_1, the total number of prosocial choices made with a partner present in this experiment. Test the hypothesis that $p_{\text{partner}} = 0.5$ versus the two-sided alternative that it does not. Report a p-value. Create a graph that shows the sampling distribution of X_1 under the null hypothesis and indicates (with different colors and/or lines) how the p-value relates to the graph. Interpret the results of the hypothesis test in context.

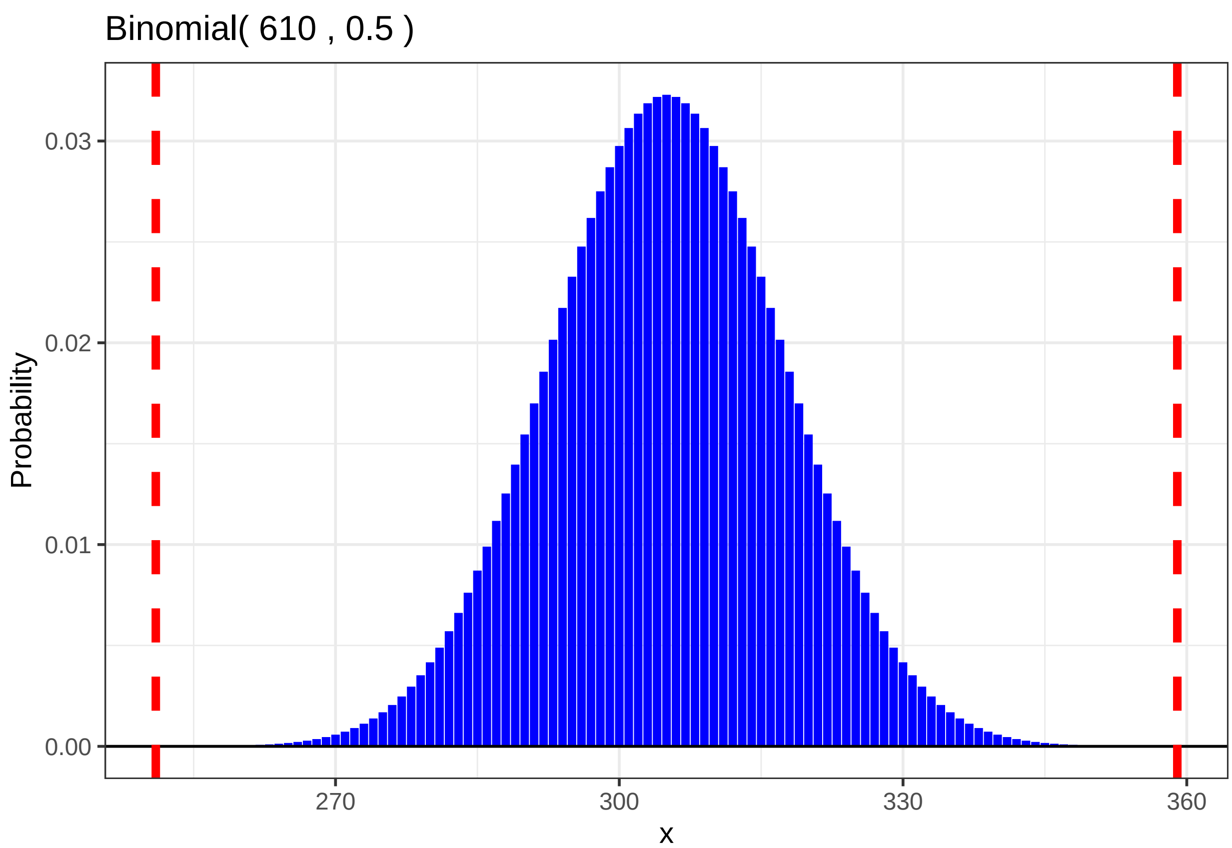
```
data1 <- read_csv("C:/stat_240/data/chimpanzee.csv") %>%
  mutate(with_partner = case_when(partner == "none" ~ FALSE, TRUE ~ TRUE)) %>%
  filter(with_partner == TRUE) %>%
  group_by(with_partner) %>%
  summarise(prosocial = sum(prosocial),
            selfish = sum(selfish),
            n = prosocial + selfish,
            p = prosocial / n)




n = 610
x = 359
p0 = 0.5
tol = 1e-8

all_x = 0:n
extreme_x = all_x[dbinom(all_x, n, p0) < dbinom(x, n, p0) + tol]
pvalue = sum( dbinom(extreme_x, n, p0))
pvalue
```

```
## [1] 1.405895e-05
```

```
pvalue2 = pbinom(251,n,p0) + (1 - pbinom(358,n,p0))
pvalue2
```

```
## [1] 1.405895e-05
```

```
gbinom(610, 0.5, scale = TRUE, size= 1.5) +
  geom_vline(aes(xintercept=359), color = "red", size=1.5, linetype = "dashed") +
  geom_vline(aes(xintercept=251), color = "red", size=1.5, linetype = "dashed") +
  theme_bw()
```



There is strong evidence that chimpanzees make the prosocial with a partner present choice more than half the time in the experimental settings (p=0.000014, two-sided binomial test). Since the p_value is less than 0.01, we can say that this is highly statistically significant. More specifically, the null hypothesis might be false and the true p is closer to the observed 359/610.

2

> Repeat the previous problem, but use the data for all trials without a partner for an assumed universal parameter $p_{\text{no partner}}$, using a statistical model for X_2, the total number of prosocial choices made without a partner present in this experiment.

```
data2 <- read_csv("C:/stat_240/data/chimpanzee.csv") %>%
  mutate(with_partner = case_when(partner == "none" ~ FALSE, TRUE ~ TRUE)) %>%
  filter(with_partner != TRUE) %>%
  group_by(with_partner) %>%
  summarise(prosocial = sum(prosocial),
            selfish = sum(selfish),
            n = prosocial+selfish,
            p = prosocial/n)




n = 180
x = 97
p0 = 0.5
tol = 1e-8
```

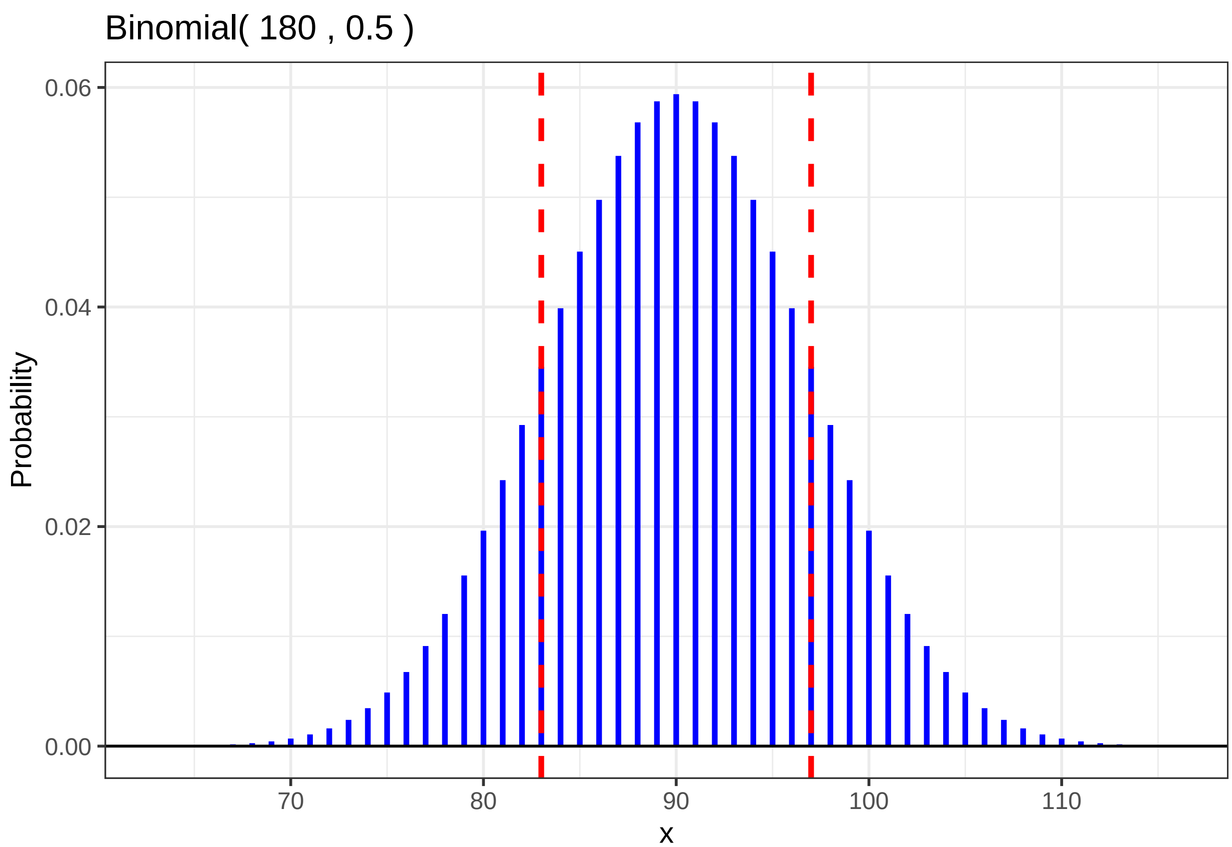
```r
all_x = 0:n
extreme_x = all_x[dbinom(all_x, n, p0) < dbinom(x, n, p0) + tol]
pvalue = sum( dbinom(extreme_x, n, p0))
pvalue
```

```
## [1] 0.3325791
```

```r
pvalue2 = pbinom(83,n,p0) + (1 - pbinom(96,n,p0))
pvalue2
```

```
## [1] 0.3325791
```

```r
gbinom(180,0.5,scale=TRUE, size = 1) +
  geom_vline(xintercept=83,color="red",
             linetype="dashed", size = 1) +
  geom_vline(xintercept=97,color="red",
             linetype="dashed", size = 1) +
  theme_bw()
```



In this case, the null hypothesis is significant and should not be rejected, consistent with the choosing randomly assumption.

3

Hypothesis tests may also be used to compare population proportions. Here, we wish to test the null hypothesis that $p_{\text{partner}} = p_{\text{no partner}}$ versus the alternative that they are different. Notice that this hypothesis statement differs from the previous two in that there is no specific value for the proportions to be equal to if the null hypothesis is true. This problem will lead you through a randomization approach to test the hypothesis.

(a) Let p be the unknown shared probability of making the prosocial choice in a single trial if the null hypothesis is true. Write down statistical models for X_1 and X_2 defined in the previous problems under this hypothesis.

X_1 ~ Binomial(610,p) X_2 ~ Binomial(180,p2)

(b) Under the null hypothesis, what is a statistical model for $X = X_1 + X_2$? Use the combined data from the experiment with all trials with and without a partner to find the maximum likelihood estimate for p assuming the null hypothesis is true.

```
data3 <- read_csv("C:/stat_240/data/chimpanzee.csv") %>%
  mutate(with_partner = case_when(partner == "none" ~ FALSE, TRUE ~ TRUE)) %>%
  summarise(prosocial = sum(prosocial),
            selfish = sum(selfish),
            n = prosocial + selfish,
            phat = prosocial/n)
data3
```

```
## # A tibble: 1 x 4
##   prosocial selfish     n  phat
##       <dbl>   <dbl> <dbl> <dbl>
## 1       442     348   790 0.559
```

The maximum likelihood estimate is $p = 0.5595$

(c) Use simulation to conduct the experiment $B = 10,000$ times using the value for p estimated in the previous problem. This results in B simulated values X_1^* and X_2^* from the assumed statistical model. For each corresponding replicate of the simulation, calculate a test statistic which is the difference in sample proportions. This collection of simulated proportion differences is a simulation-based estimate of the sampling distribution of the test statistic. Find the mean and standard deviation of this distribution.

```
N <-  10000 ## number of repetitions
df_e <- tibble(
  p_hat_1 = rbinom(N,610,442/790) / 610,
  p_hat_2 = rbinom(N,180,442/790) / 180,
  diff = p_hat_1 - p_hat_2,
  extreme = abs(diff) >= abs((83/180)-(359/610))## compare simulated differences to our observed diffe

p_value_e <- mean(df_e$extreme)
sd = sd(df_e$extreme)
p_value_e
```

```
## [1] 0.002
```

```
sd
```

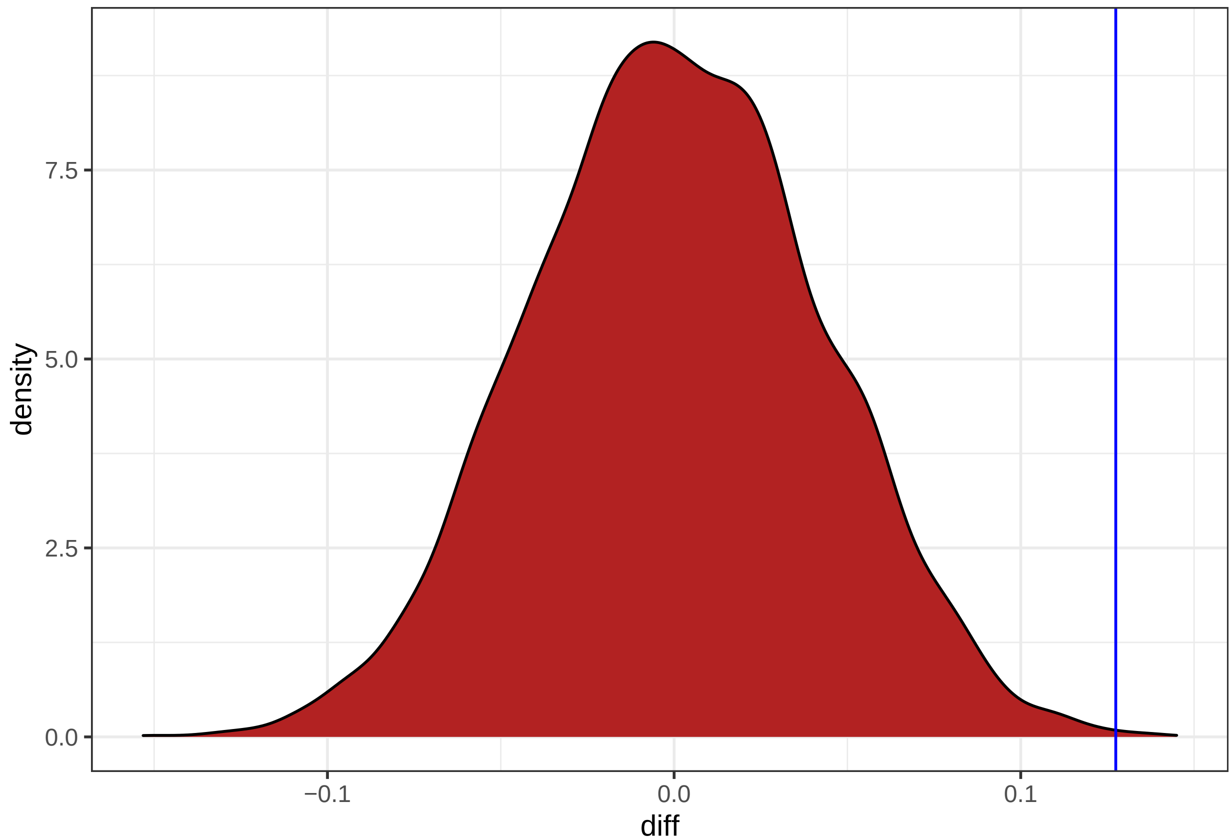
```
## [1] 0.04467885
```

(d) What should the value of the the mean of the sampling distribution approach if we let B approach infinity? What special two-word name is given to the standard deviation of this sampling distribution?

If we let B approach infinity, then the value of the mean of the sampling distribution approach 0 and the special two-word name is standard error.

(e) Display the distribution of the simulated sampling distribution. Add to this graph a vertical line which is the realized test statistic from the actual data.

```
ggplot(df_e, aes(x=diff)) +
  geom_density(fill = "firebrick") +
  geom_vline(xintercept = (359/610)-(83/180),color='blue')+
  theme_bw()
```



(f) Calculate the p-value for this hypothesis test. You may either directly report the proportion of extreme simulated proportion differences or make an approximation based on the shape of the sampling distribution to compute an area under an appropriate density curve.

```
diff = (359/610)-(83/180)
simulated_diff = (rbinom(10000,610,442/790) / 610)-(rbinom(10000,180,442/790) / 180)
mean(abs(simulated_diff)>=diff)
```

```
## [1] 0.0023
```

(g) Summarize the results of the hypothesis test in context. (In context means you should be discussing what the results say about the probabilities of chimpanzees making prosocial choices with or without partners, and not about statistical significance or rejecting hypotheses.)

The low p value suggests that there should be discrepancy about the probabilities between with/without partner in terms of making prosocial choices.

Likelihood ratio test

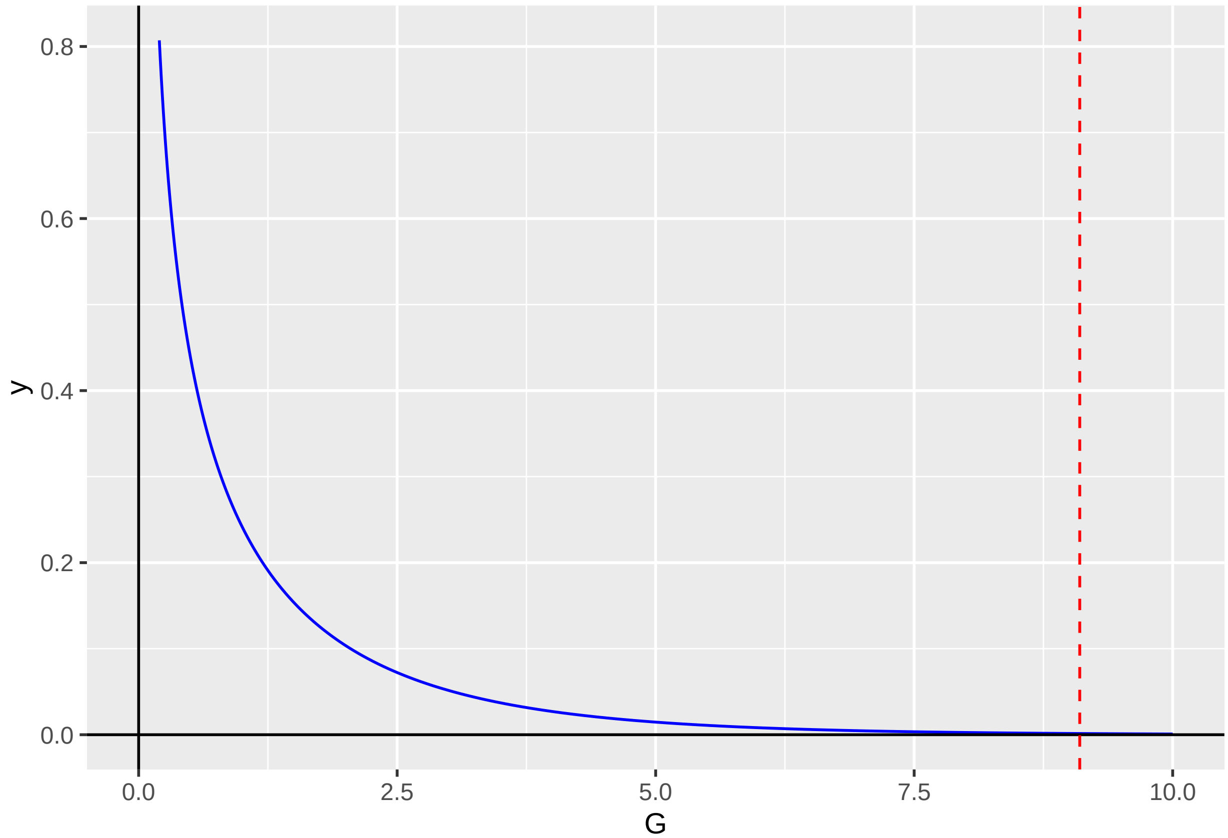
```r
extra <- read_csv("C:/stat_240/data/chimpanzee.csv") %>%
  mutate(with_partner = case_when(partner == "none" ~ FALSE, TRUE ~ TRUE)) %>%
  group_by(with_partner) %>%
  summarise(prosocial = sum(prosocial),
            selfish = sum(selfish),
            n = prosocial + selfish,
            p = prosocial / n)




extra2 = extra %>%
  mutate(p0 = sum(prosocial) / sum(n),
         logl_0 = dbinom(prosocial, n, p0, log = TRUE),
         logl_1 = dbinom(prosocial, n, p, log = TRUE))

extra
```

```
## # A tibble: 2 x 5
##   with_partner prosocial selfish     n     p
##   <lgl>            <dbl>   <dbl> <dbl> <dbl>
## 1 FALSE               83      97   180 0.461
## 2 TRUE               359     251   610 0.589
```

```r
extra_sum = extra2 %>%
  summarize( logl_0 = sum(logl_0),
             logl_1 = sum(logl_1),
             G = 2 * (logl_1 - logl_0),
             pval = 1 - pchisq(G,1))

extra_sum
```

```
## # A tibble: 1 x 4
##   logl_0 logl_1     G    pval
##    <dbl>  <dbl> <dbl>   <dbl>
## 1  -10.8  -6.24  9.10 0.00255
```

```r
ggplot() +
  geom_chisq_fill(df = 1, a = extra_sum$G, b = 10) +
  geom_chisq_density(df = 1, a = 0.2, b = 10) +
  geom_hline(yintercept = 0) +
  geom_vline(xintercept = extra_sum$G, linetype = "dashed", color = "red") +
  geom_vline(xintercept = 0) +
  xlab("G")
```



The p-value from the likelihood ratio test is the area to the right of G=9.10 under a chi-square random variable with one degree of freedom which is about 0.0026.

4

> Write three criticisms of the assumptions made for the previous three hypothesis tests where reality may differ from the assumptions, possibly leading to misleading conclusions.

There are only seven chimpanzees in the experiment and all live in the same social group. These chimpanzees may not be representative of all chimpanzees, including those in other research stations or in the wild. Trials within the same session may not be independent. Different chimpanzees may exhibit different behavior in the long run. The probability of making the prosocial choice might differ over time within sessions. The probability of making the prosocial choice might differ between sessions.