# U1P1_Version2 due 10/2/20

John Kaspers

9/12/2020

## Goals of computations:

## 1) Read in data

## 2) Manipulate data so that table is readable (with confidence intervals)

## 3) Create confidence intervals and add them to table

## 4) Perform log_lik test on three groupings of 9 weeks each of data from table

## Making sure Rmd can knit

#1 # Reading in the data

```
mydata = read.csv("http://dept.stat.lsa.umich.edu/~bbh/s485/data/UMCovid_SPSU+4_2020.csv")
mydata
```

```
##    week_of tests positivity
## 1   3/8/20    36      0.194
## 2  3/15/20   273      0.125
## 3  3/22/20   179      0.223
## 4  3/29/20   120      0.167
## 5   4/5/20   106      0.085
## 6  4/12/20    86      0.047
## 7  4/19/20   124      0.040
## 8  4/26/20   131      0.008
## 9   5/3/20   163      0.025
## 10 5/10/20   162      0.006
## 11 5/17/20   159      0.019
## 12 5/24/20   156      0.006
## 13 5/31/20   204      0.005
## 14  6/7/20   339      0.015
## 15 6/14/20   431      0.049
```

```
## 16 6/21/20   348       0.029
## 17 6/28/20   398       0.025
## 18  7/5/20   572       0.023
## 19 7/12/20   525       0.029
## 20 7/19/20   454       0.046
## 21 7/26/20   664       0.035
## 22  8/2/20   871       0.018
## 23  8/9/20   865       0.010
## 24 8/16/20   853       0.012
## 25 8/23/20  1497       0.025
## 26 8/30/20  1717       0.011
## 27  9/6/20  2240       0.019
## 28 9/13/20  3687       0.013
## 29 9/20/20  3387       0.017
## 30 9/27/20     7       0.143
```

#2 Manipulating data

# Next I must create a third numerical variable in mydata for the number of Covid-19 tests that came out positive. This value will be p__hat multiplied by n, or V2 multiplied by X

```
mydata$X_positive <- mydata$tests*mydata$positivity
mydata
```

```
##    week_of tests positivity X_positive
## 1   3/8/20    36      0.194      6.984
## 2  3/15/20   273      0.125     34.125
## 3  3/22/20   179      0.223     39.917
## 4  3/29/20   120      0.167     20.040
## 5   4/5/20   106      0.085      9.010
## 6  4/12/20    86      0.047      4.042
## 7  4/19/20   124      0.040      4.960
## 8  4/26/20   131      0.008      1.048
## 9   5/3/20   163      0.025      4.075
## 10 5/10/20   162      0.006      0.972
## 11 5/17/20   159      0.019      3.021
## 12 5/24/20   156      0.006      0.936
## 13 5/31/20   204      0.005      1.020
## 14  6/7/20   339      0.015      5.085
## 15 6/14/20   431      0.049     21.119
## 16 6/21/20   348      0.029     10.092
## 17 6/28/20   398      0.025      9.950
## 18  7/5/20   572      0.023     13.156
## 19 7/12/20   525      0.029     15.225
## 20 7/19/20   454      0.046     20.884
## 21 7/26/20   664      0.035     23.240
## 22  8/2/20   871      0.018     15.678
## 23  8/9/20   865      0.010      8.650
## 24 8/16/20   853      0.012     10.236
```

```
## 25 8/23/20  1497      0.025     37.425
## 26 8/30/20  1717      0.011     18.887
## 27  9/6/20  2240      0.019     42.560
## 28 9/13/20  3687      0.013     47.931
## 29 9/20/20  3387      0.017     57.579
## 30 9/27/20     7      0.143      1.001
```

## Creating Confidence Intervals -> Agresti - Coull

```
#lower bound
lower <- function(X, n){
    k <- qnorm(0.975)
    X_tilda <- X + k^2/2
    n_tilda <- n + k^2
    p_tilda <- X_tilda/n_tilda
    q_tilda <- 1 - p_tilda


  p_tilda - k*sqrt(p_tilda * q_tilda)*n_tilda^(-1/2)
}

#Finding lower bound and adding to table
mydata$lower.bound <- round(lower(mydata$X_positive, mydata$tests), digits = 3)

#upper bound
upper <- function(X, n){
    k <- qnorm(0.975)
    X_tilda <- X + k^2/2
    n_tilda <- n + k^2
    p_tilda <- X_tilda/n_tilda
    q_tilda <- 1 - p_tilda


  p_tilda + k*sqrt(p_tilda * q_tilda)*n_tilda^(-1/2)
}
#Finding upper bound and adding to table
mydata$upper.bound <- round(upper(mydata$X_positive, mydata$tests), digits = 3)
```

## Log-Likelihood Computations:

```
#First I must divide up the data into three periods, so I will have 3 periods each of length 9
#I must then take the mean of each period's test-positivity

##########
#Period 1
##########
p_First_9 <- mean(mydata$positivity[1:9])
#For the number of tests taken in period 1
sum(mydata$tests[1:9])
```

```
## [1] 1218
```

```r
First_9_weeks <- (mean(mydata$positivity[1:9]) * sum(mydata$tests[1:9]))
First_9_weeks <- round(First_9_weeks)

#########
#Period 2
#########
p_Middle_9 <- mean(mydata$positivity[10:18])
sum(mydata$tests[10:18])
```

```
## [1] 2769
```

```r
Middle_9_weeks<- (mean(mydata$positivity[10:18]) * sum(mydata$tests[10:18]))
Middle_9_weeks <- round(Middle_9_weeks)

#########
#Period 3
########
p_Last_9<- mean(mydata$positivity[19:27])
sum(mydata$tests[19:27])
```

```
## [1] 9686
```

```r
Last_9_weeks <- (mean(mydata$positivity[19:27]) * sum(mydata$tests[19:27]))
Last_9_weeks <- round(Last_9_weeks)

#Now, I can consider taking the likelihood estimates of the underlying probabilities of each period
# I must also then compare H_0 and H_a
```

H_0: p_First_9 = p_Middle_9 = p_Last_9

H_A: p_First_9 > p_Middle_9 > p_Last_9

```r
p_First_9
```

```
## [1] 0.1015556
```

```r
p_Middle_9
```

```
## [1] 0.01966667
```

```r
p_Last_9
```

```
## [1] 0.02277778
```

```r
# Log-likelihood function
#####################
n_1 = c(sum(mydata$tests[1:9]),sum(mydata$tests[10:18]), sum(mydata$tests[19:27]))
p_u = c(p_First_9, p_Middle_9, p_Last_9)
```

```r
log_lik = function(p) {
  sum(dbinom(round(n_1*p_u), size=n_1, prob=p, log=T))
}
var = 2*(log_lik(p_u) - log_lik(sum(round(n_1*p_u))/sum(n_1)))
pchisq(var, 2, lower=F)
```

```
## [1] 1.443793e-36
```

### Likelhood Ratio Test and the overall positivity rate

```r
overall_test_positivity <- c(p_First_9, p_Middle_9, p_Last_9)
mean_overall_test_positivity <- mean(overall_test_positivity)

log_lik(mean_overall_test_positivity)
```

```
## [1] -153.5796
```

```r
var_2 = 2*(log_lik(mean_overall_test_positivity) - log_lik(sum(round(n_1*p_u))/sum(n_1)))
pchisq(var_2, 2, lower=F)
```

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