Stat 516 HW 5

Dongyang Wang

11/14/2021

Q1

```
#rm(list=ls())
set.seed(42)
# 1.2
mean(c(0, 1, 3, 2, 5, 1, 0, 5, 3, 2, 0, 3, 5, 1, 2, 0, 7, 1, 3, 0, 0, 4, 2, 3, 5, 6, 1, 2, 3,
1, 1, 2, 1, 0, 4, 2, 4, 1, 4, 0, 0, 2, 1, 0, 3, 5, 0, 3, 0, 1))
```

```
## [1] 2.1
```

```
# 1.3
Tvalue <- sqrt(50)*2*7.4
pvalue <- 2*(1 -  pnorm(Tvalue, 0, 1) )
pvalue
```

```
## [1] 0
```

```
# 1.4
# normal
z975 <- qnorm(0.975,0,1)
CI1 = 8.4+ z975 * sqrt(4/50)
CI2 = 8.4- z975 * sqrt(4/50)
c(CI2,CI1)
```

```
## [1] 7.845638 8.954362
```

```
# bootstrap
sample <- c(0, 1, 3, 2, 5, 1, 0, 5, 3, 2, 0, 3, 5, 1, 2, 0, 7, 1, 3, 0, 0, 4, 2, 3, 5, 6, 1, 2, 3,
1, 1, 2, 1, 0, 4, 2, 4, 1, 4, 0, 0, 2, 1, 0, 3, 5, 0, 3, 0, 1)
mle <- c()
for(i in 1:200){
  sample1 <- sample(sample, 50,replace =  T)
  mle[i] = mean(sample1)/0.25
}
mle
```

```
##   [1]  8.72  8.48  8.48  8.08  8.32  5.76  7.04 10.24  8.00  8.64  7.04  8.80
##  [13]  8.72  7.84  6.32  7.52 10.24  7.52  8.32  7.76  8.40  6.96  9.28  9.36
##  [25]  9.44  7.44  8.80  8.80  9.52  8.08  9.12  8.64  8.88  8.40  8.00  7.52
##  [37]  7.04  8.08  9.12  6.96  8.88  7.20  6.16  9.12  9.28  8.40 11.28  7.36
##  [49]  7.68  9.28  7.12  7.92  8.16  8.24  8.08  9.36  9.52 10.80 10.32  9.36
##  [61]  8.40  8.56  9.28  8.40  8.08  8.40  9.60  7.52  9.60  8.88  7.84  8.40
##  [73]  8.00  7.20  7.92  8.24  7.28 10.00  6.96 10.48  6.56  8.96  8.48  8.56
##  [85]  7.76  9.12  8.96  7.60  9.84  9.84  8.48  8.88  9.20  9.52  8.64  9.28
##  [97]  9.52  8.88  8.48  7.76  7.76  8.40  9.36  9.44  6.64  8.96  7.12  7.44
## [109]  7.12  9.36  8.08  8.48  7.12  6.72  8.88  9.52  7.44  8.56  8.56 10.40
## [121]  7.44  8.80  7.52  7.92  8.72  8.56  8.88  8.64  6.80  8.80  8.88  8.64
## [133]  7.52  6.88  6.80  9.28  8.56  6.88  8.00  8.80  9.68  9.44  7.20 11.04
## [145]  8.00  9.52  7.68  8.72  8.80  9.84 10.48  7.44  8.24  8.80  8.40  9.28
## [157]  9.20  8.16  7.28  8.56  8.88  6.64  8.08  9.44  8.64  8.88  8.40 10.96
## [169]  7.76  9.44  9.04  9.60  9.60 11.28  9.68 11.28  8.16 10.00  9.12  9.04
## [181]  7.04  8.80  6.96  8.56  8.08 10.32  7.52  9.28  8.48 10.56  7.84  9.68
## [193]  7.84  6.64  9.52  9.44  8.00  7.84 10.08  8.56
```

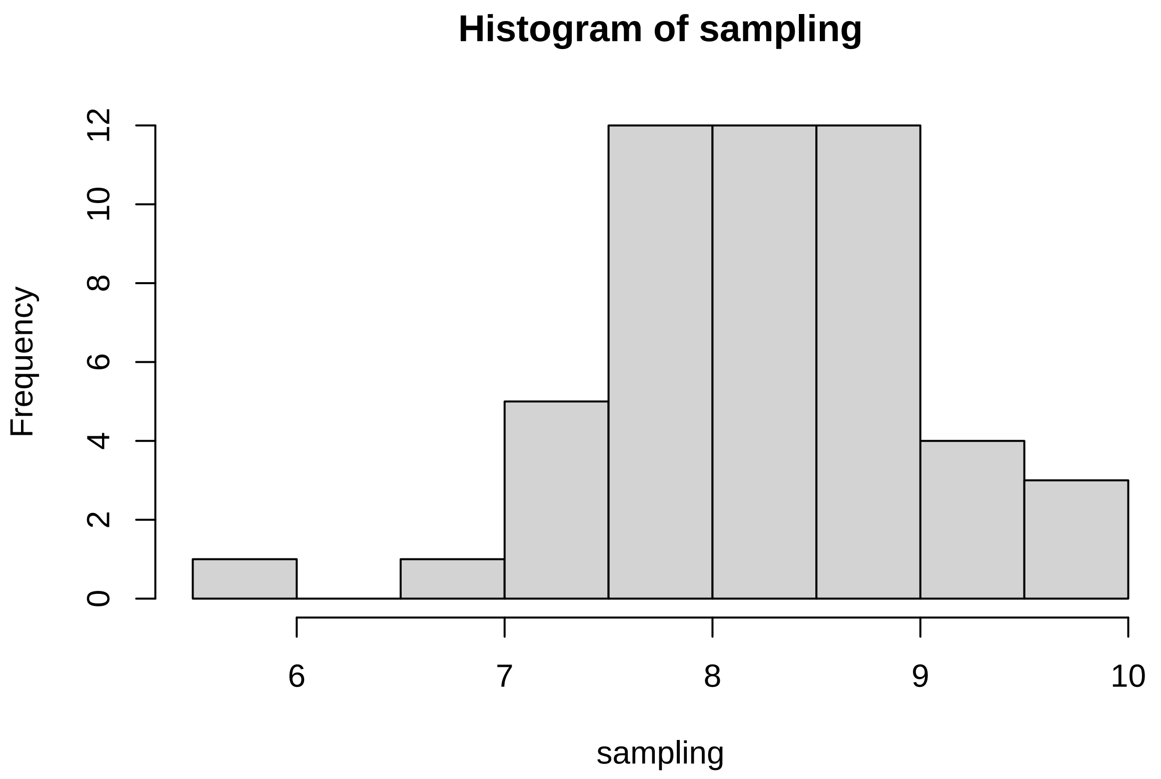
```
mle <- sort(mle)
mle
```

```
##   [1]  5.76  6.16  6.32  6.56  6.64  6.64  6.64  6.72  6.80  6.80  6.88  6.88
##  [13]  6.96  6.96  6.96  6.96  7.04  7.04  7.04  7.04  7.12  7.12  7.12  7.12
##  [25]  7.20  7.20  7.20  7.28  7.28  7.36  7.44  7.44  7.44  7.44  7.44  7.52
##  [37]  7.52  7.52  7.52  7.52  7.52  7.52  7.60  7.68  7.68  7.76  7.76  7.76
##  [49]  7.76  7.76  7.84  7.84  7.84  7.84  7.84  7.92  7.92  7.92  8.00  8.00
##  [61]  8.00  8.00  8.00  8.00  8.08  8.08  8.08  8.08  8.08  8.08  8.08  8.08
##  [73]  8.16  8.16  8.16  8.24  8.24  8.24  8.32  8.32  8.40  8.40  8.40  8.40
##  [85]  8.40  8.40  8.40  8.40  8.40  8.40  8.48  8.48  8.48  8.48  8.48  8.48
##  [97]  8.48  8.56  8.56  8.56  8.56  8.56  8.56  8.56  8.56  8.56  8.64  8.64
## [109]  8.64  8.64  8.64  8.64  8.72  8.72  8.72  8.72  8.80  8.80  8.80  8.80
## [121]  8.80  8.80  8.80  8.80  8.80  8.88  8.88  8.88  8.88  8.88  8.88  8.88
## [133]  8.88  8.88  8.88  8.96  8.96  8.96  9.04  9.04  9.12  9.12  9.12  9.12
## [145]  9.12  9.20  9.20  9.28  9.28  9.28  9.28  9.28  9.28  9.28  9.28  9.36
## [157]  9.36  9.36  9.36  9.36  9.44  9.44  9.44  9.44  9.44  9.44  9.52  9.52
## [169]  9.52  9.52  9.52  9.52  9.52  9.60  9.60  9.60  9.60  9.68  9.68  9.68
## [181]  9.84  9.84  9.84 10.00 10.00 10.08 10.24 10.24 10.32 10.32 10.40 10.48
## [193] 10.48 10.56 10.80 10.96 11.04 11.28 11.28 11.28
```

```
c(mle[5],mle[195])
```

```
## [1]  6.64 10.80
```

```
# 1.5
set.seed(42)
shape = sum(sample) + 0.84 - 1
rate = 50*0.25 + 0.27
sampling = rgamma(50,shape, rate)
hist(sampling)
```



```
mean(sampling)
```

```
## [1] 8.242868
```

```
#install.packages('bayestestR')
library(bayestestR)
ci(sampling, method = "HDI")
```

```
## 95% HDI: [6.82, 9.87]
```

Q2

```
set.seed(42)
precipitation <- read.delim('snoqualmie_falls.txt', header = FALSE, sep ='')

# For simplicity, treat all February months as 28 days
total_days <- 31+28+31
precipitation <- precipitation[, 1:total_days]

for(i in 1:nrow(precipitation)){
  for(j in 1:ncol(precipitation)){
    if ( precipitation[i,j] >0){
      precipitation[i,j] <- 1
    }
  }
}

year <- c(1948:1983)

p00 = 0
```

```
p01 = 0
p10 = 0
p11 = 0
for(i in 1:nrow(precipitation)){
  for(j in 2:ncol(precipitation)){
    if(precipitation[i,j-1] == 0){
      if(precipitation[i,j] ==0){
        p00 = p00 +1
      }
      else{
        p01 = p01 +1
      }
    }
    if(precipitation[i,j-1] == 1){
      if(precipitation[i,j] ==0){
        p10 = p10 +1
      }
      else{
        p11 = p11 +1
      }
    }
  }
}
#not rain to not rain
p00
```

```
## [1] 625
```

```
#not rain to rain
p01
```

```
## [1] 423
```

```
#rain to not rain
p10
```

```
## [1] 423
```

```
#rain to rain
p11
```

```
## [1] 1733
```

```
tpm <- matrix(c(p00/(p00+p01), p01/(p00+p01), p10/(p10+p11), p11/(p10+p11)),
              nrow = 2, ncol = 2,byrow=TRUE)
tpm
```

```
##           [,1]      [,2]
## [1,] 0.5963740 0.4036260
## [2,] 0.1961967 0.8038033
```

(a)

```
# 1

# tpm
tpm
```

```
##           [,1]      [,2]
## [1,] 0.5963740 0.4036260
## [2,] 0.1961967 0.8038033
```

```
# mle
p12 = p01
p21 = p10
mle1 <- tpm[1,2]
mle2 <- tpm[2,1]

# stationary distribution
tpm1 <- tpm
for (i in 1:5){
  tpm1 <- tpm1 %*% tpm1
  print(tpm1)
}
```

```
##           [,1]      [,2]
## [1,] 0.4348521 0.5651479
## [2,] 0.2747101 0.7252899
##           [,1]      [,2]
## [1,] 0.3443482 0.6556518
## [2,] 0.3187027 0.6812973
##           [,1]      [,2]
## [1,] 0.3275337 0.6724663
## [2,] 0.3268760 0.6731240
##           [,1]      [,2]
## [1,] 0.3270914 0.6729086
## [2,] 0.3270910 0.6729090
##           [,1]      [,2]
## [1,] 0.3270911 0.6729089
## [2,] 0.3270911 0.6729089
```

```
# critical value
cv <- qnorm(0.975,0,1)

# n
n <- 90*36

# For mle1
denom1 <- sqrt(mle1 * (1-mle1) / (n * tpm1[1,1]))
CI11 <- mle1- denom1 * cv
CI12 <- mle1+ denom1 * cv
mle1
```

```
## [1] 0.403626
```

```
c(CI11, CI12)
```

```
## [1] 0.3740873 0.4331646
```

```
# For mle2
denom2 <- sqrt(mle2 * (1-mle2) / (n * tpm1[2,2]))
CI21 <- mle2- denom2 * cv
CI22 <- mle2+ denom2 * cv
mle2
```

```
## [1] 0.1961967
```

```
c(CI21, CI22)
```

```
## [1] 0.1795273 0.2128660
```

So, mle1 is 0.403626 with confidence interval 0.3740873, 0.4331646. So, mle2 is 0.1961967 with confidence interval 0.1795273, 0.2128660.

(b)

With independent uniform priors, we know that the posteriors follows Beta distributions. This first has with parameters $n_1, n - n_1$. Namely, the pdf can be written as $q_1^{n_1-1}(1-q_1)^{n-n_1-1}$. Another one is $q_2^{n_2-1}(1-q_2)^{n-n_2-1}$, follows $Beta(n_2, n-n_2)$. Here $p_{12} = q_1$ and $p_{21} = q_2$.

```
# medians
qbeta(0.5, p12, n - p12)
```

```
## [1] 0.1304795
```

```
qbeta(0.5, p21, n - p21)
```

```
## [1] 0.1304795
```

```
# CI
ci(distribution_beta(n, p12, n-p12), method = "HDI")
```

```
## 95% HDI: [0.12, 0.14]
```

```
ci(distribution_beta(n, p21, n - p21), method = "HDI")
```

```
## 95% HDI: [0.12, 0.14]
```

The medians are the same:0.1304795, and the credible intervals are the same: [0.12, 0.14].

(c)

```
# Frequentist

part1 <- p01 * log(p01*n/((p01 +p00) * (p01+p11)))
part2 <- p00 * log(p00*n/((p01 +p00) * (p00+p10)))
part3 <- p10 * log(p10*n/((p10 +p11) * (p10+p00)))
part4 <- p11 * log(p11*n/((p11 +p10) * (p01+p11)))

t_n1 <- 2*(part1 + part2 + part3 +part4)

# degree of freedom = total observations - # columns
dof = 2^2 -2

qchisq(0.95, dof)
```

```
## [1] 5.991465
```

Since the critical value is 5.991465 but our statistic is way larger, we can reject the null hypothesis and claim that the raining data is actually dependent, i.e., rain today implies a higher chance for raining tomorrow.

Since the likelihood has been calculated, we select a uniform prior which makes it easy to calculate the BF.

```
# Bayesian
rain_days <- sum(precipitation)
ratio1 <- rain_days/n
ratio2 <- 1 - ratio1

numeratorBAY <- ratio1^(rain_days/2)* ratio2^((n-rain_days)/2) *
```

```
  ratio1^(rain_days/2)* ratio2^((n-rain_days)/2)
denominatorBAY <- tpm[1,1]^(p00) * tpm[1,2]^(p01)  * tpm[2,1]^(p10) * tpm[2,2]^(p11)

bayesianf <- numeratorBAY/denominatorBAY

log(ratio1^(rain_days/2))*2 + log(ratio2^((n-rain_days)/2) )*2
```

```
## [1] -2047.443
```

```
log(tpm[1,1]^(p00)) + log(tpm[1,2]^(p01)) + log(tpm[2,1]^(p10)) +log(tpm[2,2]^(p11))
```

```
## [1] -1774.23
```

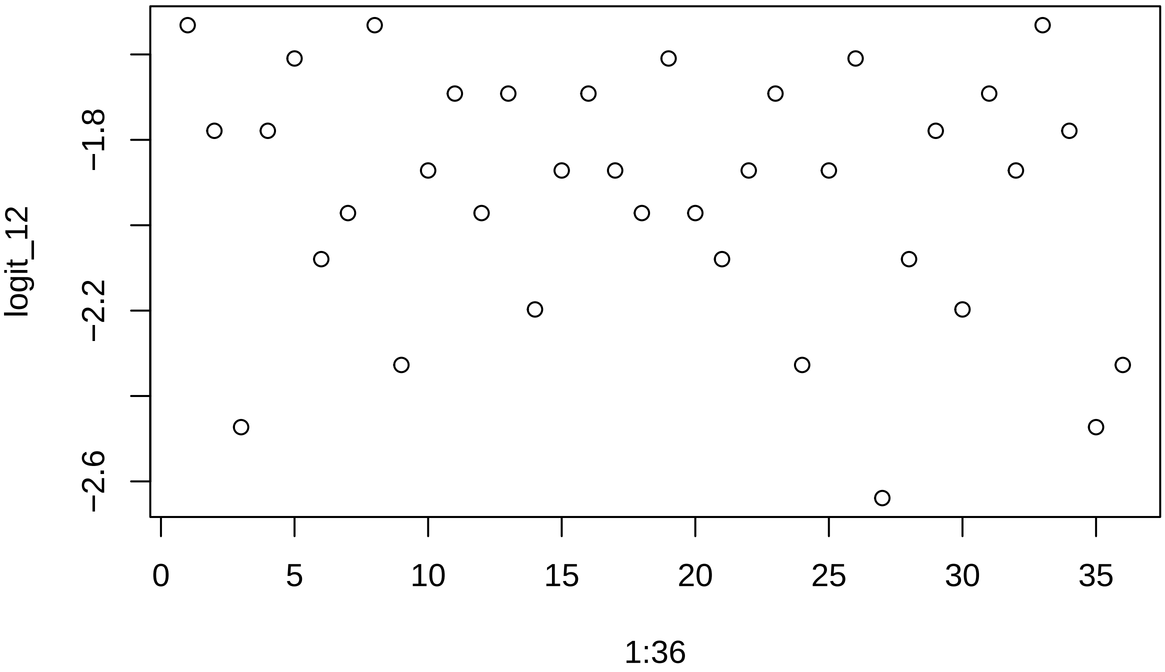
This calculation does not provide the answer, however, we can obtain the number of zeros by multiplying numeratorBAY and denominatorBAY. Using log, we can find that the numerator is smaller. Therefore, we reject the null hypothesis and claim that the data is not independent.

(d)

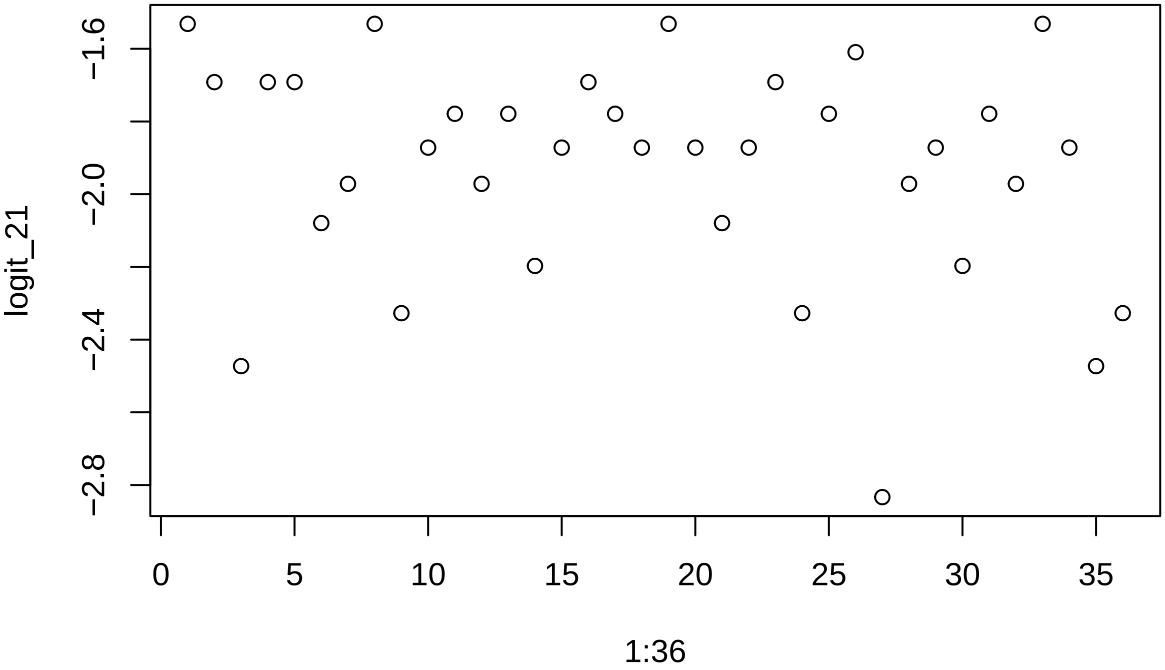
```
# 1
logit_12 = c()
for (i in 1:nrow(precipitation)){
  index = 0
  for(j in 2:ncol(precipitation)){
    if(precipitation[i,j-1] == 0){
      if(precipitation[i,j] ==1){
        index = index + 1
      }
    }
  }
  proportion1 = index/total_days
  logit_12[i] = log(proportion1/(1-proportion1))
}

logit_21 = c()
for (i in 1:nrow(precipitation)){
  index = 0
  for(j in 2:ncol(precipitation)){
    if(precipitation[i,j-1] == 1){
      if(precipitation[i,j] ==0){
        index = index + 1
      }
    }
  }
  proportion1 = index/total_days
  logit_21[i] = log(proportion1/(1-proportion1))
}


plot(1:36, logit_12)
```



```r
plot(1:36, logit_21)
```



As shown above, the two logits are plotted vs year.

Assuming a Dirichlet distribution for the prior, we can easily compute the Bayesian Factor with the likelihood by following the lecture notes.

```r
statistic1 <- c()

for (i in 1:nrow(precipitation)){
  index = 0
  for(j in 2:ncol(precipitation)){
    if(precipitation[i,j-1] == 0){
      if(precipitation[i,j] ==0){
        index = index + 1

      }
    }
  }
  statistic1[i] = (index/total_days)^index/(index+1)
}

statistic2 <- c()
for (i in 1:nrow(precipitation)){
  index = 0
  for(j in 2:ncol(precipitation)){
    if(precipitation[i,j-1] == 0){
      if(precipitation[i,j] ==1){
        index = index + 1
      }
    }
  }
  statistic2[i] = (index/total_days)^index/(index+1)
}

statistic3 <- c()
for (i in 1:nrow(precipitation)){
  index = 0
  for(j in 2:ncol(precipitation)){
    if(precipitation[i,j-1] == 1){
      if(precipitation[i,j] ==0){
        index = index + 1
      }
    }
  }
  statistic3[i] = (index/total_days)^index/(index+1)
}

statistic4<- c()
for (i in 1:nrow(precipitation)){
  index = 0
  for(j in 2:ncol(precipitation)){
    if(precipitation[i,j-1] == 1){
      if(precipitation[i,j] ==1){
        index = index + 1
      }
    }
  }
  statistic4[i] = (index/total_days)^index/(index+1)
}

# LRT
qchisq(0.95, 36*2)
```

```
## [1] 92.80827
```

```r
# Vector numerator
numerator <- statistic1 * statistic2 *statistic3 * statistic4

statistic1 = prod(statistic1)
statistic2 = prod(statistic2)
statistic3 = prod(statistic3)
statistic4 = prod(statistic4)

numerator <- statistic1 * statistic2 *statistic3 * statistic4

denominator <- tpm[1,1]^(p00+1)/p00 * tpm[1,2]^(p01+1)/p01  * tpm[2,1]^(p10+1)/p10 * tpm[2,2]^(p11+1)/p

bf <- numerator/denominator
bf
```

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

```r
# Investigate denominator terms
tpm[1,1]^(p00+1)/p00
```

```
## [1] 4.773621e-144
```

```r
tpm[1,2]^(p01+1)/p01
```

```
## [1] 2.036065e-170
```

```r
tpm[2,1]^(p10+1)/p10
```

```
## [1] 2.984713e-303
```

```r
tpm[2,2]^(p11+1)/p11
```

```
## [1] 1.954184e-168
```

```r
zero1 <- 36*36
zero2 <- 144+170+303+168
zero1 > zero2
```

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

Here, we have calculated the likelihood's integration and their product. The result is proportional to the likelihood and therefore for the LRT, we know that the ratio is small and apparently smaller than the threshold 92.80827, so we reject the null that all years have a same tpm.

For the Bayesian approach, the numerator and the denominator of the BF are too small to calculate, but we can estimate them by looking at the number of zeros before the significant digits. I approximate the number of zeros in the numerator by using the product of the smallest number of zeros and the number of years. Apparently, there are more zeros in the numerator in the BF. Therefore, BF < 1 and we can conclude that we reject the null hypothesis. Therefore, each year varies in terms of tpm.