Computer Practical 1

Statistics 2

Dom Hutchinson

2 Binomial Maximum Likelihood Estimators

Let $Y \sim \text{Binomial}(n, p)$. The maximum likelihood estimate for p is $\hat{p}(Y) = \frac{Y}{n}$. $\hat{p}$ is an unbiased estimator.

Question 1

Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$. Then

$$
\begin{aligned}
\text{Var}(\hat{p}) &= \text{Var}\left(\frac{Y}{n}\right) \\
&= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}.n\text{Var}\left(X_1\right) \\
&= \frac{1}{n}p(1-p)
\end{aligned}
$$

Question 2

```
n<-13; p<-.31
sample_size<-100; trials=1000
var_hat=(1/n)*p*(1-p)
cat("var_hat:",var_hat,"\n")
```

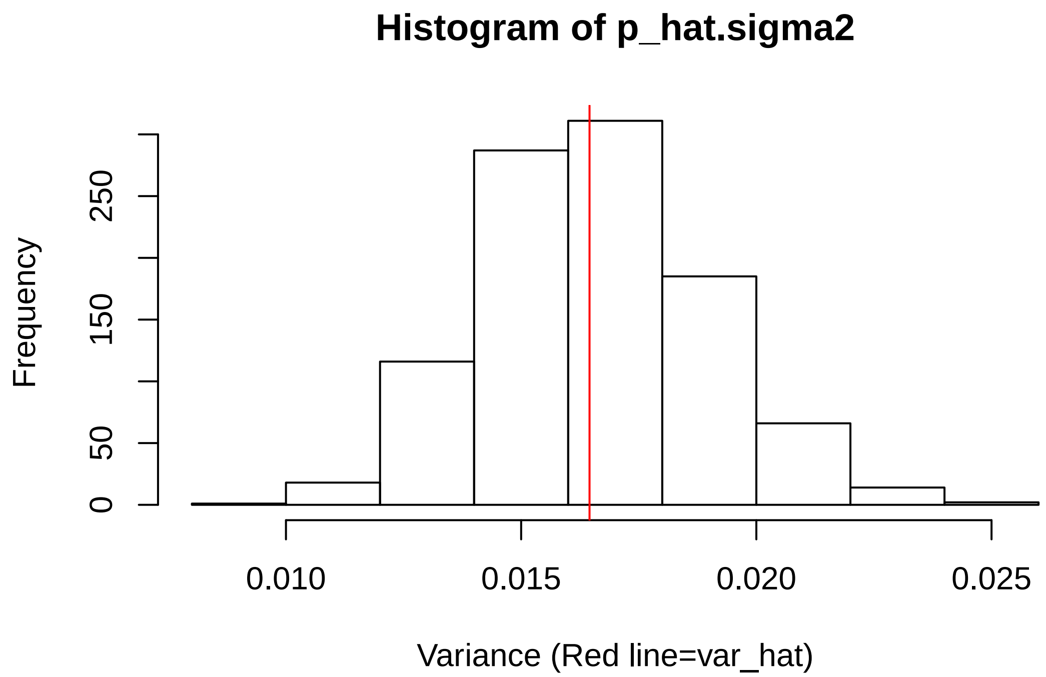
```
## var_hat: 0.01645385
```

```
phat<-function(Y) {
  Y/n
}

x_values<-rbinom(n=sample_size*trials,size=n,prob=p)
x_samples<-matrix(x_values,nrow=sample_size)
p_hat.samples<-apply(x_samples,1,phat)
p_hat.sigma2=apply(p_hat.samples,1,var)
cat("observered_var:",mean(p_hat.sigma2))
```

```
## observered_var: 0.01659042
```

```
hist(p_hat.sigma2,breaks=10,xlab="Variance (Red line=var_hat)")
abline(v=var_hat,col="red")
```



3 Clinic Data

```
year.data<-read.csv("year_data.csv")
knitr::kable(year.data)
```

year	births	deaths	clinic
1841	3036	237	1
1842	3287	518	1
1843	3060	274	1
1844	3157	260	1
1845	3492	241	1
1846	4010	459	1
1841	2442	86	2
1842	2659	202	2
1843	2739	164	2
1844	2956	68	2
1845	3241	66	2
1846	3754	105	2

Let $Y_i \sim \text{Binomial}(n_i, p_i)$ model the number of deaths in clinic i where n_i is the total number of births in clinic i & p_i is the mortality rate for clinic i. Assume Y_1 & Y_2 are independent.

```
n1 <- sum(year.data[year.data$clinic==1,]$births) # number of births in clinic 1
y1 <- sum(year.data[year.data$clinic==1,]$deaths) # number of deaths in clinic 1
cat("Number of births in clinic 1:",prettyNum(n1,big.mark=","),"\nNumber of deaths in clinic 1:",pretty
```

```
## Number of births in clinic 1: 20,042
```

```
## Number of deaths in clinic 1: 1,989
```

```r
n2 <- sum(year.data[year.data$clinic==2,]$births) # number of births in clinic 2
y2 <- sum(year.data[year.data$clinic==2,]$deaths) # number of deaths in clinic 2
cat("Number of births in clinic 2:",prettyNum(n2,big.mark=","),"\nNumber of deaths in clinic 2:",prettyN
```

```
## Number of births in clinic 2: 17,791
## Number of deaths in clinic 2: 691
```

Question 3

```r
p1_hat=y1/n1
p2_hat=y2/n2

cat("p1_hat:",p1_hat,"\np2_hat:",p2_hat)
```

```
## p1_hat: 0.09924159
## p2_hat: 0.03883986
```

Quetion 4

Assume that $p = p_1 = p_2$ and define $W := \hat{p}_1(Y_1) - \hat{p}_2(Y_2)$. Then

$$
\begin{aligned}
\mathbb{E}(W) &= \mathbb{E}(\hat{p}_1(Y_1) - \hat{p}_2(Y_2)) \\
&= \mathbb{E}(\hat{p}_1(Y_1)) - \mathbb{E}(\hat{p}_2(Y_2)) \\
&= p_1 - p_2 \\
&= p - p \\
&= 0 \\
\mathrm{Var}(W) &= \mathrm{Var}(\hat{p}_1(Y_1) - \hat{p}_2(Y_2)) \\
&= \mathrm{Var}(\hat{p}_1(Y_1)) + \mathrm{Var}(\hat{p}_2(Y_2)) \\
&= \tfrac{1}{n_1} p_1(1-p_1) + \tfrac{1}{n_2} p_2(1-p_2) \\
&= \tfrac{1}{n_1} p(1-p) + \tfrac{1}{n_2} p(1-p) \\
&= \tfrac{n_1+n_2}{n_1 n_2} p(1-p)
\end{aligned}
$$

Question 5

Suppose $p = p_1 = p_2$. We have $\hat{p} = \frac{1989+691}{20042+17791} = \frac{2680}{37833} = 0.0708376$.

$$
\begin{aligned}
\mathbb{P}(|W - \mu_W| \geq \hat{p}_1(y_1) - \hat{p}_1(y_2)) &= \mathbb{P}(|W - | \geq 0.0992416 - 0.0388399) \\
&= \mathbb{P}(|W| \geq 0.0604017) \\
&\leq \frac{\sigma_W^2}{0.0604017^2} \text{ by Chebyshev's Inequality} \\
&= \frac{1}{0.0604017^2} \times \frac{n_1+n_2}{n_1 n_2} \hat{p}(1-\hat{p}) \\
&= \frac{1}{0.0604017^2} \times \frac{20042+17791}{20042 \times 17791} \times 0.0708376 \times 0.9291624 \\
&= 0.0019142
\end{aligned}
$$

Thus it is very unlikely to observe these two mortality rates, assuming the underlying rate is the same.

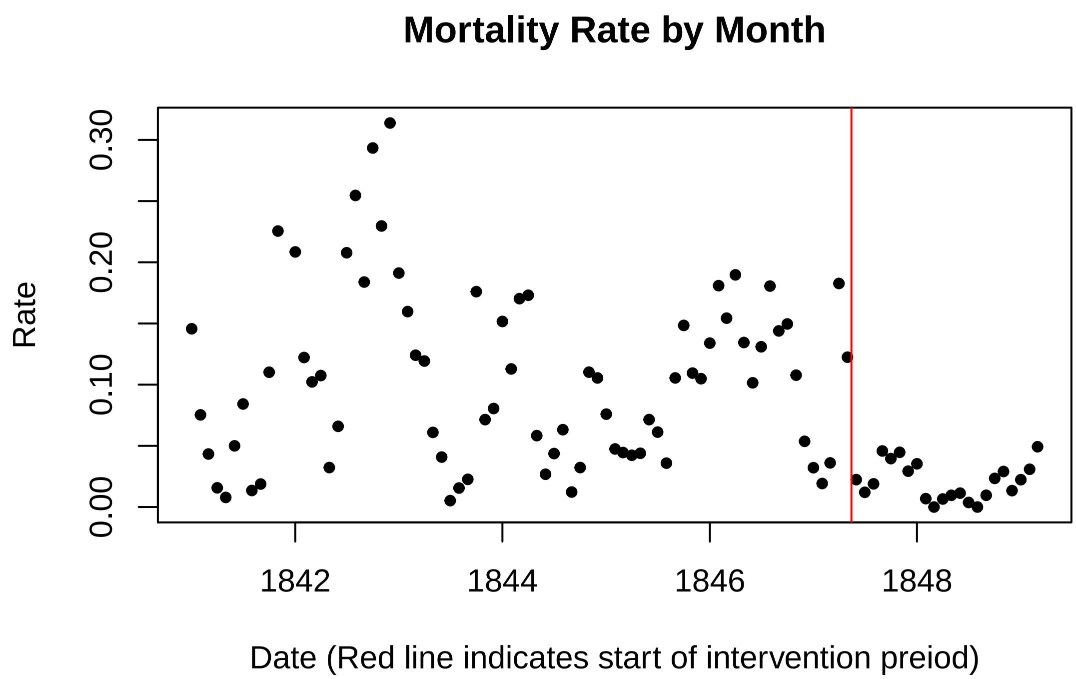
4 Intervention: Chlorine Hand Washing

```r
month.data<-read.csv("month_data.csv")
month.data<-month.data[!is.na(month.data$births),]
```

```
month.data$rate<-month.data$deaths/month.data$births
month.data$date<-as.Date(month.data$date)

intervention.date<-as.Date("1847- 5-15")

plot(month.data$date, month.data$rate, pch=20, main="Mortality Rate by Month",
     xlab="Date (Red line indicates start of intervention preiod)", ylab="Rate")
abline(v=intervention.date, col="red")
```




```
before.intervention<-month.data[month.data$date< intervention.date,]
n1<-sum(before.intervention$births)
y1<-sum(before.intervention$deaths)
cat("Number of births before intervention:",prettyNum(n1,big.mark=","),
    "\nNumber of deaths before intervention:",prettyNum(y1,big.mark=","),"\n\n")
```

```
## Number of births before intervention: 19,571
## Number of deaths before intervention: 2,060
```

```
after.intervention <-month.data[month.data$date>=intervention.date,]
n2<-sum(after.intervention$births)
y2<-sum(after.intervention$deaths)
cat("Number of births after intervention: ",prettyNum(n2,big.mark=","),
    "\nNumber of deaths after intervention: ",prettyNum(y2,big.mark=","))
```

```
## Number of births after intervention:  6,595
## Number of deaths after intervention:  142
```

Question 6

```
p1_hat<-y1/n1
p2_hat<-y2/n2

cat("p1_hat:",p1_hat,"\np2_hat:",p2_hat)
```

```
## p1_hat: 0.1052578
## p2_hat: 0.02153146
```

Define random variable $W = \hat{p}_1(Y_1) - \hat{p}_2(Y_2)$.

Suppose $p = p_1 = p_2$. We have $\hat{p} = \frac{y_1+y_2}{n_1+n_2} - \frac{2060+142}{19571+6595} = \frac{2202}{26166} = 0.084155$.

$$\begin{aligned}
\mathbb{P}(|W - \mu_W| \geq \hat{p}_1(y_1) - \hat{p}_1(y_2)) &= \mathbb{P}(|W - 0| \geq 0.1052578 - 0.0215315) \\
&= \mathbb{P}(|W| \geq 0.0837263) \\
&\leq \frac{\sigma_W^2}{0.0837263^2} \text{ by Chebyshev's Inequality} \\
&= \frac{1}{0.0837263^2} \times \frac{n_1+n_2}{n_1 n_2}\hat{p}(1-\hat{p}) \\
&= \frac{1}{0.0837263^2} \times \frac{19571+6595}{19571 \times 6595} \times 0.084155 \times 0.915845 \\
&= 0.0022289
\end{aligned}$$

Thus. it is very unlikely to observe these two mortality rates, assuming the underlying rate is the same.

5 A First Logistic Regression

```
x1<-c(1,0)
x2<-c(1,1)

sigma<-function(z) {
  1/(1+exp(-z))
}
```

Question 7

$$\begin{array}{rrcl}
 & L(\theta) & \propto & \prod_{i=1}^{2} f_{Y_i}(y_i; n_i, x_i, \theta) \\
 & & = & f(y_1; n_1, x_1, \theta) f(y_2; n_2, x_2, \theta) \\
 & & = & \binom{n_1}{y_1} g(x_1,\theta)^{y_1}(1-g(x_1,\theta))^{n_1-y_1}\binom{n_2}{y_2} g(x_2,\theta)^{y_2}(1-g(x_2,\theta))^{n_2-y_2} \\
 & & = & \binom{n_1}{y_1} \sigma(\theta_1)^{y_1}(1-\sigma(\theta_1))^{n_1-y_1}\binom{n_2}{y_2} \sigma(\theta_1+\theta_2)^{y_2}(1-\sigma(\theta_1+\theta_2))^{n_2-y_2} \\
 & & \propto & \sigma(\theta_1)^{y_1}(1-\sigma(\theta_1))^{n_1-y_1}\sigma(\theta_1+\theta_2)^{y_2}(1-\sigma(\theta_1+\theta_2))^{n_2-y_2} \\
\implies & \ell(\theta) & = & c + y_1 \ln(\sigma(\theta_1)) + (n_1-y_1)\ln(1-\sigma(\theta_1)) \\
 & & + & y_2 \ln(\sigma(\theta_1+\theta_2)) + (n_2-y_2)\ln(1-\sigma(\theta_1+\theta_2))
\end{array}$$

$$\begin{array}{rrcl}
 & \sigma(\theta_1) & = & \dfrac{1}{1+e^{-\theta_1}} \\
\implies & \ln(\sigma(\theta_1)) & = & -\ln(1+e^{-\theta_1}) \\
\& & \ln(1-\sigma(\theta_1)) & = & \ln\left(\frac{e^{-\theta_1}}{1+e^{-\theta_1}}\right) \\
 & & = & -\theta_1 - \ln(1+e^{-\theta_1})
\end{array}$$

$$\begin{array}{rrcl}
 & \sigma(\theta_1+\theta_2) & = & \dfrac{1}{1+e^{-(\theta_1+\theta_2)}} \\
\implies & \ln(\sigma(\theta_1+\theta_2)) & = & -\ln(1+e^{-(\theta_1+\theta_2)}) \\
\& & \ln(1-\sigma(\theta_1+\theta_2)) & = & \ln\left(\frac{e^{-(\theta_1+\theta_2)}}{1+e^{-(\theta_1+\theta_2)}}\right) \\
 & & = & -(\theta_1+\theta_2) - \ln(1+e^{-(\theta_1+\theta_2)})
\end{array}$$

$$\begin{array}{rrcl}
\implies & \ell(\theta) & = & c - y_1\ln(1+e^{\theta_1}) - (n_1-y_1)(\theta_1 + \ln(1+e^{-\theta_1})) \\
 & & - & y_2 \ln(1+e^{-(\theta_1+\theta_2)}) - (n_2-y_2)(\theta_1+\theta_2+\ln(1+e^{-(\theta_1+\theta_2)})) \\
 & & = & c - n_1\ln(1+e^{-\theta_1}) - n_2\ln(1+e^{-(\theta_1+\theta_2)}) - \theta_1(n_1-y_1) - (\theta_1+\theta_2)(n_2-y_2) \\
 & & = & c - 19571\ln(1+e^{-\theta_1}) - 6595\ln(1+e^{-(\theta_1+\theta_2)}) - \theta_1(19571-2060) - (\theta_1+\theta_2)(6595-2060) \\
 & & = & c - 19571\ln(1+e^{-\theta_1}) - 6595\ln(1+e^{-(\theta_1+\theta_2)}) - 17511\theta_1 - 4535(\theta_1+\theta_2) \\
 & & = & c - 19571\ln(1+e^{-\theta_1}) - 6595\ln(1+e^{-(\theta_1+\theta_2)}) - 22046\theta_1 - 4535\theta_2
\end{array}$$

```r
ell<-function(theta) {
  l<--19571*log(1+exp(-theta[1]))-6595*log(1+exp(-theta[2]))-22046*theta[1]-4535*theta[2]
  -l # In order to find maximum
}

result<-optim(c(.5,.5), ell)
cat("theta_hat:(",result$par[1],",",result$par[2],")")
```

```
## theta_hat:( -709.7827 , -0.7892119 )
```

This value of $\hat{\theta}_1$ shows that the probability of mortality before intervention was very low, which is corroborated by $\hat{p}_1$ being low. This value of $\hat{\theta}_2$ shows that the probability of mortality decreased after intervention, this is corroborated by $\hat{p}_2 < \hat{p}_1$.