Statistics 2: Computer Practical 1 Solutions by Joshua Acton

Question 1.

$$Y \sim \text{Binomial}(n, p)$$

The ML estimator of p is

$$\hat{p}(Y) = \frac{Y}{n}$$

and this is unbiased. Then the variance of $\hat{p}(Y)$ can be calculated:

$$\begin{aligned}
Var(\hat{p}(Y); p) &= \mathbb{E}(\hat{p}(Y)^2; p) - \mathbb{E}(\hat{p}(Y); p)^2 \\
&= \mathbb{E}\left(\frac{Y^2}{n^2}; p\right) - p^2 \\
&= \frac{1}{n^2}\mathbb{E}(Y^2; p) - p^2 \\
&= \frac{np(1-p) + n^2p^2 - n^2p^2}{n^2} \\
&= \frac{p(1-p)}{n}
\end{aligned}$$

Question 2.

The following code performs 1000 trials, each trial taking 100 samples from a Binomial$(13, 0.31)$ distribution, and computes the empirical variance of the ML estimate of p, $\hat{p}(Y)$, from each trial. It then plots a histogram of the empirical variances of $\hat{p}(Y)$, with a vertical dashed red line plotted at the value of the ML estimate of $Var(\hat{p}(Y))$ derived in question 1.

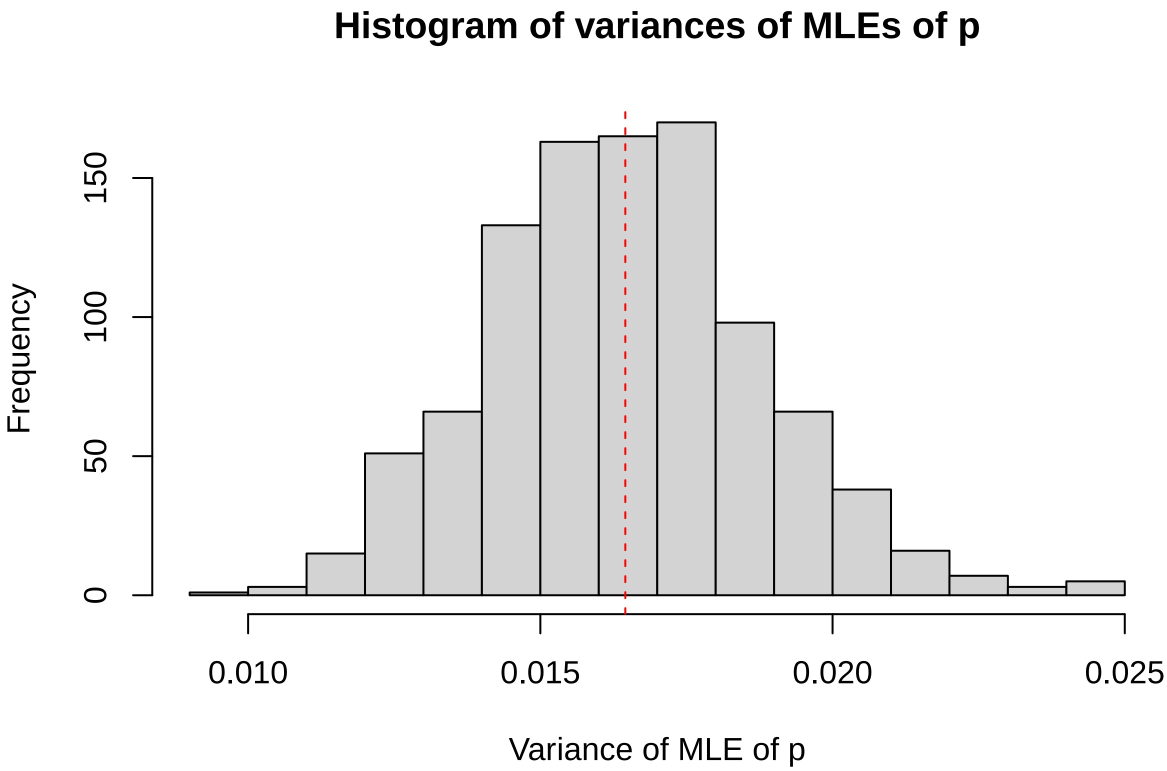
```r
trials <- 1000
samples <- 100
n <- 13
p <- 0.31

var.p.mles <- rep(0, trials)

for (i in 1:trials) {
  xs <- rbinom(samples, n, p)
  var.p.mles[i] <- var(xs/n)
}

var.mle <- p*(1-p)/n

hist(var.p.mles, breaks=20, main="Histogram of variances of MLEs of p",
     xlab="Variance of MLE of p", ylab="Frequency")
abline(v=var.mle, col="red", lty=2)
```



We can see from the histogram that the empirical values of the variance of $\hat{p}$ do center on the value obtained from the ML estimate of $Var(\hat{p}(Y))$ derived in question 1.

Question 3

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

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

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

p.hat.1 = y1/n1 # MLE of p for clinic 1
p.hat.2 = y2/n2 # MLE of p for clinic 2
```

```
## The MLE of the mortality rate from clinic 1 is 0.09924159
```

```
## The MLE of the mortality rate from clinic 2 is 0.03883986
```

Question 4

We have that

$$Y_1 \sim \text{Binomial}(n_1, p_1)$$

and

$$Y_2 \sim \text{Binomial}(n_2, p_2)$$

and assume that Y_1 and Y_2 are independent. Defining W by

$$W := \hat{p}_1(Y_1) - \hat{p}_2(Y_2),$$

and under the assumption that

$$p_1 = p_2 = p,$$

We have that

$$\begin{aligned}\mathbb{E}(W;p) &= \mathbb{E}(\hat{p}_1(Y_1);p) - \mathbb{E}(\hat{p}_2(Y_2);p) \\ &= p - p \\ &= 0\end{aligned}$$

which follows from linearity of expectation and the fact that the ML estimator of p for a binomial distribution is unbiased. Moreover,

$$\begin{aligned}Var(W) &= Var(\hat{p}_1(Y_1) - \hat{p}_2(Y_2);p) \\ &= Var(\hat{p}_1(Y_1);p) + Var(\hat{p}_2(Y_2);p) - 2Cov(\hat{p}_1(Y_1), \hat{p}_2(Y_2);p) \\ &= \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2} - 2(\mathbb{E}(\hat{p}_1(Y_1)\hat{p}_2(Y_2);p) - \mathbb{E}(\hat{p}_1(Y_1);p)\mathbb{E}(\hat{p}_2(Y_2);p)) \\ &= \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2} - 2(\mathbb{E}(\hat{p}_1(Y_1);p)\mathbb{E}(\hat{p}_2(Y_2);p) - \mathbb{E}(\hat{p}_1(Y_1);p)\mathbb{E}(\hat{p}_2(Y_2);p)) \\ &= \frac{n_2 p(1-p) + n_1 p(1-p)}{n_1 n_2}\end{aligned}$$

following on from the fact that Y_1 and Y_2 are independent.

Question 5

Chebyshev's inequality gives us that for a random variable $X, \mu = \mathbb{E}(X), \sigma^2 = Var(X), \forall\, k > 0$,

$$\mathbb{P}(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

Applying Chebyshev's inequality to W as defined above and using that $\mathbb{E}(W) = 0$ and $Var(W) = \frac{2p(1-p)}{n}$,

$$\begin{aligned}\mathbb{P}(|W - \mathbb{E}(W)| \geq k) &\leq \frac{Var(W)}{k^2} \\ \mathbb{P}(|W| \geq k) &\leq \frac{2p(1-p)}{nk^2}\end{aligned}$$

Since $\hat{p}_1(Y_1) \approx 0.10$ and $\hat{p}_2(Y_2) \approx 0.04$,

$$W = \hat{p}_1(Y_1) - \hat{p}_2(Y_2) \approx 0.06$$

Setting $k = 0.06$ and letting $n = n_1 + n_2 = 20042 + 17791 = 37833$,

$$\mathbb{P}(|W| \geq 0.06) \leq \frac{2p(1-p)}{37833 \cdot 0.06^2} = \frac{2p(1-p)}{136.1988}$$

Now we just need to find an upper bound for $2p(1-p),\ p \in [0,1]$.

$$\begin{aligned}2p(1-p) &= 2p - 2p^2 \\ \frac{d}{dp} 2p - 2p^2 &= 2 - 4p \\ 2 - 4p &= 0 \\ \implies p &= \frac{1}{2} \in [0,1] \\ \frac{d^2}{dp^2} 2p - 2p^2 &= -4 \\ &\leq 0\ \forall\, p \in [0,1]\end{aligned}$$

So we know that $p = \frac{1}{2}$ maximises $2p(1-p)$.

$$\begin{aligned}\mathbb{P}(|W| \geq 0.06) \leq \frac{2p(1-p)}{136.1988} &\leq \frac{2 \cdot \frac{1}{2}(1-\frac{1}{2})}{136.1988} \\ &= \frac{1}{272.3976} \\ &\approx 0.0037\end{aligned}$$

This provides a very small upper bound on the probability that the difference between the observed mortality rates could be so large (under the assumption that the underlying mortality rates are identical).

Question 6

We have that:

$$Y_1 \sim \text{Binomial}(n_1, p_1)$$

and

$$Y_2 \sim \text{Binomial}(n_2, p_2)$$

and assume that Y_1 and Y_2 are independent.

```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("1847-05-15")
before.intervention <- month.data[month.data$date < intervention.date,]
after.intervention <- month.data[month.data$date > intervention.date,]
n1 <- sum(before.intervention$births)
y1 <- sum(before.intervention$deaths)
n2 <- sum(after.intervention$births)
y2 <- sum(after.intervention$deaths)
p.hat.1 = y1/n1 # MLE of p before intervention
p.hat.2 = y2/n2 # MLE of p after intervention
```

```
## The MLE of the mortality rate before intervention is 0.1052578
```

```
## The MLE of the mortality rate after intervention is 0.02153146
```

As before, define the random variable W as

$$W := \hat{p}_1(Y_1) - \hat{p}_2(Y_2)$$

Under the assumption that

$$p_1 = p_2 = p,$$

Chebyshev's inequality allows us to obtain

$$\mathbb{P}(|W| \geq k) \leq \frac{2p(1-p)}{nk^2}$$

From the observations of clinic data before and after intervention,

$$\begin{aligned}\hat{p}_1(Y_1) &\approx 0.11 \\ \hat{p}_2(Y_2) &\approx 0.02 \\ W = \hat{p}_1(Y_1) - \hat{p}_2(Y_2) &\approx 0.09\end{aligned}$$

so let $k = 0.09$.
As before, $n = n_1 + n2 = 37833$. Substituting values into Chebyshev's inequality gives us

$$\mathbb{P}(|W| \geq 0.09) \leq \frac{2p(1-p)}{37833 \cdot 0.08^2} \leq \frac{1}{612.8946} \approx 0.0016$$

Again, under the assumption that the underlying mortality rates are identical, we are provided with a very small upper bound on the probability that the difference between the observed mortality rates could be as large as it is observed to be.

Question 7

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

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

ell <- function(theta) {
  log(choose(n1, y1)) + log(choose(n2, y2)) + y1*log(sigma(theta[1])) +
    y2*log(sigma(theta[1] + theta[2])) + (n1 - y1)*log(1-sigma(theta[1])) +
    (n2-y2)*log(1-sigma(theta[1] + theta[2]))
}

ell(c(0,0))
```

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

The full expression for the log-likelihood function returns values that are too large for R to handle due to the massive positive constants obtained from log(n1 C y1) and log(n2 C y2). However, since these are constant for all theta, they are irrelevant to maximising the function. Instead, we can write an expression for the log-likelihood function excluding these constants and maximise this.

```
ell.reduced <- function(theta) {
  y1*log(sigma(theta[1])) + y2*log(sigma(theta[1] + theta[2])) +
    (n1 - y1)*log(1-sigma(theta[1])) +
    (n2-y2)*log(1-sigma(theta[1] + theta[2]))
}

ell.reduced(c(0,0))
```

```
## [1] -18136.89
```

```
optim.out.0 <- optim(c(0,0), ell.reduced, control=list(fnscale=-1))$par
optim.out.neg1 <- optim(c(-1,-1), ell.reduced, control=list(fnscale=-1))$par
optim.out.neg10 <- optim(c(-10,-10), ell.reduced, control=list(fnscale=-1))$par
```

```
## The value of theta.hat with starting parameter (0,0) is -2.140973 -1.675298

## The value of theta.hat with starting parameter (-1,-1) is -2.139947 -1.675914

## The value of theta.hat with starting parameter (-10,-10) is -2.140442 -1.675932
```

From the information given, $\theta_1 < 0$ means that the mortality rate before intervention is less than 50%, and $\theta_2 < 0$ means that the mortality rate decreases after intervention. From the MLEs $\hat{p}_1$ and $\hat{p}_2$ we can see that both of these facts are likely. So we can expect reasonable values of $\hat{\theta}$ to be such that $\theta_1, \theta_2 < 0$. Maximising ell.reduced(θ) with a range of plausible starting values of θ, we obtain estimates $\hat{\theta}$ that seem both consistent with our prior beliefs and with each other.