Quiz Week 5 - Statistical Inference - Solutions

FIT5197 teaching team

Question 1

For sample $\vec{x}$ of size n distributed as $N(\mu, \sigma)$ the sum of squared errors (SSE) of mean estimate μ is given by

$$\text{SSE}(\mu) = \sum_{i=1}^{n} (x_i - \mu)^2$$

Demonstrate using differentiation that the SSE point estimate $\hat{\mu}$, corresponding to the value of μ that minimises the SSE, is equivalent to the sample mean

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Answer 1

Formally we can write this process as

$$\hat{\mu} = \text{argmin}_{\mu} \left\{ \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

where $\text{argmin}_x \{f(x)\}$ means find the value of x that minimises $f(x)$.

First, differentiate $\text{sse}(\mu)$ with respect to μ

$$\frac{d\,\text{sse}(\mu)}{d\mu} = \sum_{i=1}^{n} \frac{d}{d\mu}(x_i - \mu)^2 = -2 \sum_{i=1}^{n} (x_i - \mu) = -2 \sum_{i=1}^{n} x_i + 2n\mu$$

Then set the derivative to zero, and solve for μ, yielding:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

which is readily identified as the sample mean.

Note: sometimes people use the notation $N(\mu, \sigma^2)$ for the normal distribution if they want to define the second parameter as the variance instead of the standard deviation like we did above - it doesn't matter much since it is the same distribution regardless of the notation)

Question 2

An alternative measure of goodness-of-fit to estimate the mean μ and standard deviation σ of a normally distributed random variable is maximum likelihood estimation. For sample $\mathbf{y} = (y_1, \ldots, y_n)$ drawn from the normally distributed random variables, $Y_i \sim N(\mu, \sigma)$, the likelihood is given by

$$p(\mathbf{y}|\mu, \sigma) = \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mu - y_i)^2 \right)$$

(a) Why does this expression involve the product of functions of y_i? Why is the likelihood a measure of goodness-of-fit to estimate the mean μ and standard deviation σ?

(b) Using the fact that $e^{-a}e^{-b} = e^{-a-b}$ for arbitrary variables a and b, Show the negative log-likelihood has the following form:

$$L_{-}(\mathbf{y}|\mu, \sigma) = -\log p(\mathbf{y}|\mu, \sigma) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2 \tag{1}$$

(c) Is maximising the likelihood the same as minimising the negative log-likelihood? Why? What is the difference between maximising the log-likelihood and minimising the negative log-likelihood? Why bother minimise the negative log-likelihood, why not just use likelihood?

(d) Show the MLE estimate of the mean μ is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

(e) Show the MLE estimate of the standard deviation σ is

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2}$$

(f) Show that the bias of the MLE estimate of the mean, $\hat{\mu}$, given in question 2(d) above is equal to zero. Note that bias is given by $\boldsymbol{B}_\mu = E[\hat{\mu}] - \mu$

(g) Show that the variance of the MLE estimate of the mean, $\hat{\mu}$, given in question 2(d) above is equal to $\frac{\sigma^2}{n}$. Note the variance is given by $V[\hat{\mu}] = E[(\hat{\mu} - E[\hat{\mu}])^2]$.

Answer 2

(a) The likelihood expression can be written as the product of functions of y_i because the y_i are assumed to be independent. The likelihood is a measure of goodness-of-fit to estimate the mean μ and standard deviation σ because the likelihood is a function of both μ and σ that can be maximised with respect to these variables.

(b) For the normal distribution $N(\mu, \sigma)$ and given data $\mathbf{y} = (y_1, \ldots, y_n)$ the likelihood is %

$$
\begin{aligned}
p(\mathbf{y}|\mu, \sigma) &= \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mu - y_i)^2 \right) \\
&= \prod_{i=1}^{n} \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(\mu - y_i)^2 \right) \\
&= \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\mu - y_i)^2 \right)
\end{aligned}
$$

The second last line above comes from splitting up the products. The last line above is given by the fact that $e^{-a} e^{-b} = e^{-a-b}$.

The negative log-likelihood function is then:

$$
\begin{aligned}
L(\mathbf{y}|\mu,\sigma) &= -\log p(\mathbf{y}|\mu,\sigma) \\
&= -\log\left(\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mu-y_i)^2\right)\right) && (2)\\
&= -\log\left(\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\right)-\log\left(\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mu-y_i)^2\right)\right) && (3)\\
&= -\frac{n}{2}\log\left(\frac{1}{2\pi\sigma^2}\right)+\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mu-y_i)^2 && (4)\\
&= \frac{n}{2}\log(2\pi\sigma^2)+\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 && (5)
\end{aligned}
$$

where we get equation (3) above using $\log(ab) = \log(a) + \log(b)$, equation (4) using $\log(a^b) = b\log(a)$ and $\log(\exp(a)) = a$, and equation (5) using $\log(a/b) = -\log(b/a)$.

(c) First let's be clear on the naming. The likelihood is the multiplication of probabilities for independent random variables Y_i. The log-likelihood is the log of the likelihood and the negative log likelihood is the negative of the log of the likelihood. The likelihood, log-likelihood and negative log-likelihood are different formulations of the same estimation approach. For models/distributions where you can optimise these formulations to give an analytic solution, these different formulations will arrive at the same solution for the parameter estimates. In practice with complex models, optimisation needs to be done numerically and different formulations have different advantages. Now to the answers.

Yes maximising the likelihood is the same as minimising the negative log-likelihood. When we maximise the likelihood, $p(\mathbf{y}|\mathbf{\Theta})$ with respect to the parameters Θ we are trying to maximise the likelihood of the data we see given the parameters. Since likelihood is a joint pdf (e.g. for Gaussian) or joint pmf (for discrete random variables)) we can note that the values of the likelihood will be defined on the interval $[0, C]$ where C is some positive constant. When we take the log of the likelihood we are mapping values from the interval $[0, C]$ to the interval $(-\infty, \log(C)]$. Then when we take the negative of the log-likelihood we will be mapping from the interval $[0, C]$ to the interval $(\infty, -\log(C)]$. So we can see that minimising the negative log-likelihood will lead to values closer to $-\log(C)$. At the same time this will lead to values closer to C on the likelihood interval. If instead we maximise the likelihood, this leads to values closer to C, so we can see maximising the likelihood or minimising the negative likelihood will have similar effects.

Just as there is no real difference between maximising the likelihood and minimising the negative log-likelihood, maximising the log-likelihood has the same effect as minimising the negative log-likelihood. We minimise the negative log-likelihood for two reasons. First, taking the log of multiplied variables lets us write the multiplication as a sum and this can aid with analytic solutions. It also enables us to perform more accurate optimization when numerical methods are required because probabilities/variables can often be small and this creates numerical precision issues when you

multiply numerous small values together. Using the log of multiples of small values means we can use sums of the log of small values to solve problems instead which leads to less numerical issues. Second, optimiser algorithms typically minimize a function, so we use negative log-likelihood as minimising that is equivalent to maximising the log-likelihood or the likelihood itself.

(d) To minimise the negative log-likelihood for μ and σ we need to differentiate equation (4) with respect to μ and σ and find the values that set the (partial) derivatives to zero, i.e., we need to solve the simultaneous equations:

$$\begin{aligned}\partial L(\mathbf{y}|\mu,\sigma)/\partial\mu &= 0,\\ \partial L(\mathbf{y}|\mu,\sigma)/\partial\sigma &= 0.\end{aligned}$$

Let's find $\hat{\mu}$ by computing the partial derivative with respect to μ:

$$\begin{aligned}\frac{\partial L(\mathbf{y}|\mu,\sigma)}{\partial\mu} &= \frac{\partial}{\partial\mu}\left(\frac{n}{2}\log(2\pi\sigma^2)+\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right)\\ &= 0-\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu)\\ &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}y_i+\frac{n\mu}{\sigma^2}\end{aligned}$$

which is similar to our minimum squared error estimator.

In fact, setting this equation to zero and solving for μ yields

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

which is again, just the sample mean.

(e) Plugging $\hat{\mu}$ into $L(\mathbf{y}|\mu,\sigma)$ removes μ from the equation. Let's find $\hat{\sigma}$ by computing the partial derivative with respect to σ:

$$\begin{aligned}
\frac{\partial L(\mathbf{y}|\hat{\mu},\sigma)}{\partial \sigma} &= \frac{\partial}{\partial \sigma}\left(\frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\hat{\mu})^2\right) \\
&= \frac{\partial}{\partial \sigma}\frac{n}{2}\left[\log\sigma^2 + \log(2\pi)\right] + \sum_{i=1}^{n}(y_i-\hat{\mu})^2\frac{\partial}{\partial \sigma}\frac{1}{2\sigma^2} \qquad (6)\\
&= \frac{\partial}{\partial \sigma}\frac{n}{2}\left[\log\sigma^2\right] + \frac{\partial}{\partial \sigma}\frac{n}{2}\left[\log(2\pi)\right] - \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\hat{\mu})^2 \qquad (7)\\
&= \frac{\partial}{\partial \sigma}\frac{n}{2}\left[2\log\sigma\right] + \frac{\partial}{\partial \sigma}\frac{n}{2}\left[\log(2\pi)\right] - \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\hat{\mu})^2 \qquad (8)\\
&= \frac{n}{\sigma} + 0 - \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\hat{\mu})^2 \qquad (9)\\
&= \frac{n}{\sigma} - \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\hat{\mu})^2 \qquad (10)
\end{aligned}$$

where to get equation (6) we use $\log(ab) = \log a + \log b$ and $\frac{\partial}{\partial x}Kg(z)f(x) = Kg(z)\frac{\partial}{\partial x}f(x)$, to get equation (8) we use $\log(a^b) = b\log(a)$, and to get equation (9) we use $\frac{\partial}{\partial x}f(z) = 0$. (Note this is just a subset of the changes in each line - you really should be familiar with these and how to do differentiation from MAT9004.)

(f) In general the bias formula follows:

$$\mathbf{B}_\theta(\hat{\theta}) = \mathbb{E}[\hat{\theta}(Y)] - \theta$$

The bias of the MLE estimate of the mean $\hat{\mu}$ (calculated in (d) above) for the normal distribution is zero since

$$\begin{aligned}
\mathbb{E}[\hat{\mu}] &= \mathbb{E}\left[\frac{Y_1 + Y_2 + \cdots + Y_n}{n}\right] \\
&= \frac{\mathbb{E}[Y_1]}{n} + \frac{\mathbb{E}[Y_2]}{n} + \cdots + \frac{\mathbb{E}[Y_n]}{n} \\
&= \frac{n\mu}{n} = \mu
\end{aligned}$$

where $\mathbb{E}[Y_i] = \mu$ by our assumption that $Y_i \sim \mathcal{N}(\mu,\sigma)$ and we think of the samples y_i as being drawn from the random variable Y_i.\ \pause⇒ So, the MLE estimate of the mean $\hat{\mu}$ (a.k.a the sample mean) is an unbiased estimator of the population mean μ.

(g) The variance of $\hat{\mu}$ is given by $V[\hat{\mu}] = E[(\hat{\mu} - E[\hat{\mu}])^2]$.

The variance of the MLE estimate of the mean, $\hat{\mu}$, for the normal distribution is then

$$\begin{aligned}\mathbb{V}[\hat{\mu}] &= \mathbb{V}\left[\frac{Y_1}{n} + \frac{Y_2}{n} + \cdots + \frac{Y_n}{n}\right] \\ &= \mathbb{V}\left[\frac{Y_1}{n}\right] + \mathbb{V}\left[\frac{Y_2}{n}\right] + \cdots + \mathbb{V}\left[\frac{Y_n}{n}\right] \\ &= \frac{1}{n^2}(\mathbb{V}[Y_1] + \mathbb{V}[Y_2] + \cdots + \mathbb{V}[Y_n]) \\ &= \sigma^2/n\end{aligned}$$

where we use: (i) the independence of the random variables Y_i in step 2, (ii) the fact that $\mathbb{V}[kX] = k^2\mathbb{V}[X]$ in step 3, and (iii) the fact that $\mathbb{V}[Y_i] = \sigma^2$ (by assumption) in step 4.

$\Rightarrow$ So the larger n, the less variable the MLE estimate of the mean for the normal distribution, a.k.a. the sample mean, becomes.

Question 3

For a Poisson random variable X the probability mass function is given by:

$$P(X = x) = \begin{cases} exp(-\lambda)\frac{1}{x!}\lambda^x, & if\ x \in \mathbb{Z}_+ \\ 0, & if\ x \notin \mathbb{Z}_+ \end{cases}$$

where the parameter λ is the mean of the distribution. For sample $\mathbf{x} = (x_1, \ldots, x_n)$ drawn from the Poisson distributed random variables, $X_i \sim Pois(\lambda)$,

(a) Derive the likelihood function, $p(\mathbf{x}|\lambda)$.

(b) Derive the negative log-likelihood function, $L_{-}(\mathbf{y}|\lambda)$.

(c) Derive the MLE estimate of the mean, $\hat{\lambda}$.

(d) Derive the bias of the MLE estimate of the mean, $\hat{\lambda}$.

(e) Derive the variance of the MLE estimate of the mean, $\hat{\lambda}$.

(f) Is the MLE estimate of the mean, $\hat{\lambda}$, consistent?

Answer 3

(a) Assumptions:

We assume to observe n independent draws from a Poisson distribution. In more formal terms, we observe the first n terms of an IID sequence $x = x_1 \dots x_n$ of Poisson random variables. Thus, the probability mass function of a term of the sequence, x_i is;

$$P(X = x) = \begin{cases} exp(-\lambda)\frac{1}{x!}\lambda^x, & if\ x \in \mathbb{Z}_+ \\ 0, & if\ x \notin \mathbb{Z}_+ \end{cases}$$

where Z_+ is the set of all the realizations that have a positive probability of being observed and λ is the parameter of interest (for which we want to derive the MLE).

Likelihood Function:

The n observations are independent. As a consequence, the likelihood function is equal to the product of their probability mass functions:

$$p(x|\lambda) = \prod_{j=1}^{n} f_X(x_j|\lambda)$$

Furthermore, the observed values $x_1 \dots x_n$ necessarily belong to the Z_+. So, we have;

$$p(x|\lambda) = \prod_{j=1}^{n} f_X(x_j|\lambda)$$

$$p(x|\lambda) = \prod_{j=1}^{n} \exp(-\lambda)\ \frac{1}{x!}\ \lambda^{x_j}$$

(b) Negative log-likelihood function:

By taking the natural logarithm of the likelihood function derived above, we get the log-likelihood:

$$\ln p(x|\lambda) = L(x|\lambda) = \ln\Big(\prod_{j=1}^{n} \exp(-\lambda)\ \frac{1}{x_j!}\ \lambda^{x_j}\Big) = \sum_{j=1}^{n} \ln\Big(\exp(-\lambda)\ \frac{1}{x_j!}\ \lambda^{x_j}\Big)$$

$$= \sum_{j=1}^{n} \Big[\ \ln(\exp(-\lambda))\ -\ln(x_j!)\ +\ln(\lambda^{x_j})\Big]$$

$$= \sum_{j=1}^{n} \Big[\ -\lambda\ -\ln(x_j!)\ +x_j\ln(\lambda)\Big]$$

$$= -n\lambda - \sum_{j=1}^{n} \ln(x_j!)\ +\ln(\lambda)\sum_{j=1}^{n} x_j$$

By multiplying both sides with -1, we get negative log-likelihood;

$$-L(x|\lambda) = L_{-}(x|\lambda) = n\lambda + \sum_{j=1}^{n} \ln(x_j!)\ -\ln(\lambda)\sum_{j=1}^{n} x_j$$

(c) The MLE is the solution of the following minimisation problem ;

$$\hat{\lambda}_{ML} = argmin_{\lambda} L_{-}(x|\lambda)$$

Value of a variable that maximises/minimizes a function is found by taking derivative of the function and finding the value of the variable that makes the derivative zero. Hence, the first order condition for a maximum is

$$\frac{\partial}{\partial\lambda} L_{-}(x|\lambda) = 0$$

The first derivative of the negative log-likelihood with respect to the parameter *lambda* is

$$\frac{\partial}{\partial\lambda} L_{-}(x|\lambda) = \frac{\partial}{\partial\lambda}\Big(n\lambda + \sum_{j=1}^{n} \ln(x_j!)\ -\ln(\lambda)\sum_{j=1}^{n} x_j\Big)$$

$$= \frac{\partial}{\partial\lambda}(n\lambda) + \frac{\partial}{\partial\lambda}\sum_{j=1}^{n} \ln(x_j!)\ -\frac{\partial}{\partial\lambda}(\ln(\lambda)\sum_{j=1}^{n} x_j)$$

$$= n + 0 - \frac{1}{\lambda}\sum_{j=1}^{n} x_j$$

Impose that the first derivative be equal to zero, and get;

$$\hat{\lambda}_{ML} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

Therefore, the maximum likelihood estimator $\hat{\lambda}_{ML}$ is just the sample mean of the n observations in the sample. This makes intuitive sense because the expected value of a Poisson random variable is equal to its parameter λ, and the sample mean is an unbiased estimator of the expected value.

(d) Similar to what we have done in Q2, again we have the bias formula for λ

$$\mathbf{B}_\theta(\hat{\lambda_{ML}}) = \mathbb{E}[\hat{\lambda_{ML}}] - \lambda$$

And the mean of out maximum likelihood estimator $\hat{\lambda}_{ML}$ is

$$\begin{aligned}
\mathbb{E}[\hat{\lambda}_{ML}] &= \mathbb{E}[\frac{1}{n}\sum_{j=1}^{n} X_j] \\
&= \mathbb{E}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\
&= \frac{\mathbb{E}[X_1]}{n} + \frac{\mathbb{E}[X_2]}{n} + \cdots + \frac{\mathbb{E}[X_n]}{n} \\
&= \frac{n\lambda}{n} = \lambda
\end{aligned}$$

where $\mathbb{E}[X_i] = \lambda$ by our assumption that $X_i \sim Pois(\lambda)$ and we think of the samples x_i as being drawn from the random variable X_i.

$\Rightarrow$ So, the bias of our estimator equal to zero as $\mathbf{B}_\lambda(\hat{\lambda_{ML}}) = \lambda - \lambda = 0$

(e) Similarly, The variance of $\hat{\lambda}_{ML}$ is also given by $\mathbb{V}[\hat{\lambda}_{ML}] = \mathbb{E}[(\hat{\lambda}_{ML} - \mathbb{E}[\hat{\lambda}_{ML}])^2]$.

The variance of the MLE estimate of the mean, $\hat{\lambda_{ML}}$, for the Poisson distribution is then

$$\begin{aligned}\mathbb{V}[\hat{\lambda}_{ML}] &= \mathbb{V}\left[\frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n}\right] \\ &= \mathbb{V}\left[\frac{X_1}{n}\right] + \mathbb{V}\left[\frac{Y_2}{n}\right] + \cdots + \mathbb{V}\left[\frac{X_n}{n}\right] \\ &= \frac{1}{n^2}(\mathbb{V}[X_1] + \mathbb{V}[X_2] + \cdots + \mathbb{V}[X_n]) \\ &= \lambda/n\end{aligned}$$

where we use: (i) the independence of the random variables X_i (ii) the fact that $\mathbb{V}[kX] = k^2\mathbb{V}[X]$ and (iii) the fact that $\mathbb{V}[Y_i] = \lambda$ (by assumption)

(f) Formally speaking, an estimator $\hat{\theta_n}$ is said to be **consistent** if it converges in probability to the true value of the parameter as:

$$\lim_{n\to\infty} P(|\hat{\theta_n} - \theta| \geq \epsilon) = 0 \, for \, all \, \epsilon > 0.$$

We can show this condition is equal to the $\lim_{n\to\infty} MSE_{\hat{\lambda}_{ML}} \to 0$ in which $MSE_{\hat{\lambda}_{ML}} = \mathbb{E}[(\hat{\lambda}_{ML} - \lambda)^2]$ for our maximum likelihood estimator $\hat{\lambda}_{ML}$ by applying Chebyshev's inequality:

$$P(|\hat{\lambda}_{ML} - \lambda| \geq \epsilon) = P((\hat{\lambda}_{ML} - \lambda)^2 \geq \epsilon^2) \leq \frac{\mathbb{E}[(\hat{\lambda}_{ML} - \lambda)^2]}{\epsilon^2}$$

with the fact $\mathbb{E}[\hat{\lambda}_{ML} - \lambda] = \mathbb{E}[\hat{\lambda}_{ML}] - \lambda = 0$ and then we just need to show that $MSE = \mathbb{E}[(\hat{\lambda}_{ML} - \lambda)^2]$ goes to 0 as $n \to \infty$.

Remember that MSE can be written as the sum of $bias^2$ and $variance$:

$$\begin{aligned}\mathbb{E}[(\hat{\lambda}_{ML} - \lambda)^2] &= (\mathbb{E}[\hat{\lambda}_{ML}] - \lambda)^2 + \mathbb{E}[(\hat{\lambda}_{ML} - E[\hat{\lambda}_{ML}])^2] \\ &= \mathbf{B}_\theta^2(\hat{\lambda_{ML}}) + \mathbb{V}[\hat{\lambda}_{ML}]\end{aligned}$$

From (d) we know the bias $\mathbf{B}_\theta(\hat{\lambda_{ML}})$ equal to 0, therefore

$$\begin{aligned}\lim_{n\to\infty} MSE_{\hat{\lambda}_{ML}} &= \lim_{n\to\infty} \mathbb{V}[\hat{\lambda}_{ML}] \\ &= \lim_{n\to\infty} \frac{\lambda}{n} \to 0\end{aligned}$$

So our MLE estimate of mean $\hat{\lambda_{ML}}$ is a consistent estimator for true λ.

Question 4: R code hackers mega-mini challenge

For sample $\mathbf{x} = (x_1, \ldots, x_n)$ drawn from the Bernoulli distributed random variables, $X_i \sim Ber(\theta)$, it can be shown that the log-likelihood function obeys:

$$L_+(\mathbf{x}|\theta) = ylog(\theta) + (n - y)log(1 - \theta)$$

where $y = \sum_{i=1}^{n} x_i$. (For kicks you might derive this from the Bernoulli PMF.)

Using R, create a function for the log-likelihood that takes as input a Bernoulli random sample, $\mathbf{x}$, and the probability of success parameter θ. Now in R generate a Bernoulli random sample with $n = 20$ samples and $\theta = 0.5$. Then using the R built-in function called 'optimize' and your function for the log-likelihood compute the MLE estimate of the probability of success, $\hat{\theta}$. Explain your reasoning behind the inputs you entered into the 'optimize' function. See what happens to the sum of squares error of your estimates if you increase the number of samples n or vary the true value for θ. (Note this might require a bit of self study to get 'optimize' working, but don't fret, we will provide a solution).

In [1]:
```
#This function simply follows the equation for log-likelihood function for beroulli distribution.
log_likelihood_fn <- function (theta, x) {
    y <- sum(x)
    n <- length(x)

    term1 <- y * log(theta)
    term2 <- (n-y) * log(1 - theta)

    return (term1 + term2)
}
```

In [2]:
```
#Please note that Bernoulli random variable is a special case of binomial random variable.
#Therefore, we can try rbinom(N,1,p).
#This will generate N samples, with value 1 with probability p, value 0 with probability (1-p)

theta <- 0.5
n <- 20

x <- rbinom(n, 1, theta)
```

In [3]:
```
#Optimize gives estimate of an argument with respect to the function for maximum or minimum value.
#The first argument is a function which is to be maximised or minimised. Please note that the first argument of the
#function passed to optimize is estimated
#Second argument is interval in which we try to locate the variable
#Following arguments are the arguments(except the first one) passed to the function
#maximum = T or F tells whether function is to be maximized or minimized
theta_estimate <- optimize(f=log_likelihood_fn, interval=c(0,1), x=x, maximum=TRUE)$maximum
theta_estimate
```

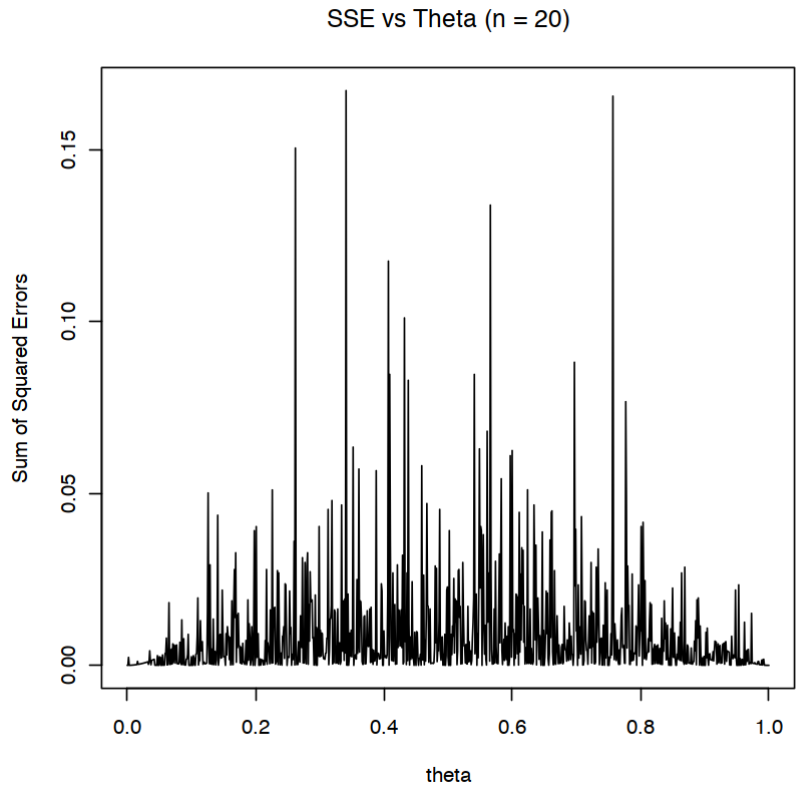
0.55000037099309

In [4]:
```
#We try to calculate sum of squared errors for our distribution w.r.t. parameter theta estimation.
#You will notice that if n is increased to a large number, theta is estimated very close to the true value.
#SSE gets very close to 0 as n is increased to a large number.
#If theta is changing, there is variation in SSE around which increases in the mid but low in the edges.
#SSE is very small when theta is 0 or 1 compared to mid values like 0.5
bernoulli_sse <- function(theta, theta_estimate){
    sse <- (theta-theta_estimate)^2
    return (sse)
}
bernoulli_sse(theta, theta_estimate)
```

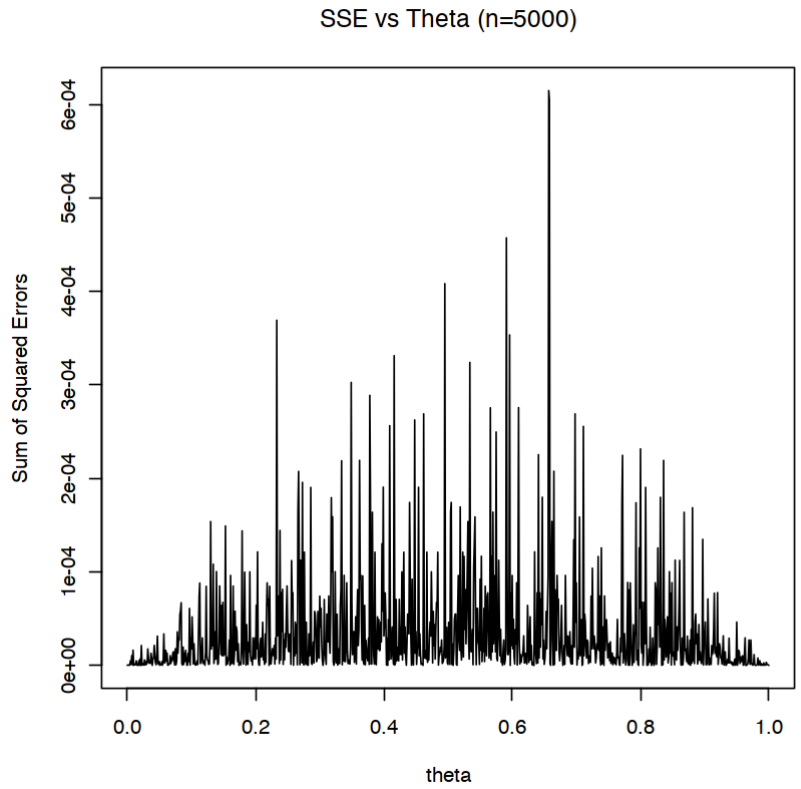
0.00250003709944668

In [5]:
```
sse_from_estimate <- function(theta, n){
    x <- rbinom(n, 1, theta)
    theta_estimate <- optimize(f=log_likelihood_fn, interval=c(0,1), x=x, maximum=TRUE)$maximum
    return (bernoulli_sse(theta, theta_estimate))
}
```

In [6]:
```
thetas <- seq(0, 1, 0.001)
n <- 20
sses <- unlist(lapply(thetas, sse_from_estimate, n=n), use.names=FALSE)
df <- data.frame(theta=thetas, sse=sses)
plot(sse ~ theta,
  data = df,
  type = "l",
  ylab = "Sum of Squared Errors",
  main = "SSE vs Theta (n = 20)")
```



In [7]:
```
thetas <- seq(0, 1, 0.001)
n <- 5000
sses <- unlist(lapply(thetas, sse_from_estimate, n=n), use.names=FALSE)
df <- data.frame(theta=thetas, sse=sses)
plot(sse ~ theta,
  data = df,
  type = "l",
  ylab = "Sum of Squared Errors",
  main = "SSE vs Theta (n=5000)")
```



In [8]:
```
ns <- seq(10, 5000, 10)
theta <- 0.5
sses <- unlist(lapply(ns, sse_from_estimate, theta=theta), use.names=FALSE)
df <- data.frame(n=ns, sse=sses)
plot(sse ~ n,
  data = df,
  type = "l",
  ylab = "Sum of Squared Errors",
  main = "SSE vs n (theta = 0.5)")
```