Mathematical Statistics & Application in R

Take-Home Quiz-1

$$Issued: Monday\ 16th\ March,\ 2020 \qquad Due: Thursday\ 26th\ March,\ 2020$$

- Members：郭俊麟 (GUO JUNLIN) / 陈梦玄 (CHEN MENGXUAN)
- Student ID：2019270146 / 2019214596
- Date：2020 / 03 / 16

Introduction

Background

Date contains the daily numbers of confirmed COVID-19 cases in Shenzhen from 2020/1/20 to 2020/2/15 (1/22, 1/23, and 2/9 data were missing or removed due to). This gives a total of `24 days of data` with `402 confirmed cases`. Because cases were reported in batches during a day, the actual occurrence time of each case is unknown. For the purposes of this quiz, we have randomly spread out the occurrences of the cases within 24 hours `according to a certain distribution` to mimic the actual occurrence times. We will not release information about this distribution. You can treat the generated occurrence times as the “real” occurrence times. In addition, a random ID was also randomly assigned to each case to denote the “actual” order of this case within all real cases on that day. For example, an ID = 200 means this is the Number 200 “real” infected cases on that day (if the numbers of equal, just add one to distinguish them).

Assumptions

1. Confirmed cases showed up `uniformly` and `independently` over 24 hours while assuming that the distribution of the **actual occurrence time (ID)** is as follows.

   1.1 Poisson Distribution

   1.2 Binomial Distribution
2. Daily data are independent between each other and they possess the same distribution with different parameters.
3. The sampling process is independent, which ensures the random samples are i.i.d from each other.

Data Overview

1. read data
2. observe data
3. split data
4. extract information

Before analyzing data, we should read the excel file using `gdata` package. And we can briefly observe the dimension of the data frame.

In [2]:

```
library(gdata)     # Apply: install.packages('gdata') , to install.
PATH = '/Users/kcl/Documents/TBSI/课程/2020_02_Semester/R Statistics/quiz/take-home-quiz1/'
```

In [3]:

```
df = read.xls(paste(PATH, 'SZ_Data.xlsx', sep=''), sheet=1, header=TRUE)
```

There are originally **402 row data** contained in the excel file with 2 columns under the names called: `Date.and.Time` and `Random.ID`.

In [4]:

```
dim(df)
```

402 · 2

In [5]:

```
names(df)
```

'Date.and.Time' · 'Random.ID'

In [6]:

```
summary(df)
```

```
          Date.and.Time   Random.ID
 2020/01/20 21:10:  2   Min.   :  0.0
 2020/01/26 15:05:  2   1st Qu.: 29.0
 2020/01/26 1:04 :  2   Median : 76.0
 2020/01/31 9:52 :  2   Mean   :101.3
 2020/02/02 0:02 :  2   3rd Qu.:160.0
 2020/02/02 2:03 :  2   Max.   :323.0
 (Other)         :390
```

However, it is inconvenient to analysze data while Data and Time are contained in a same column. We need to separate them and save them back to the data frame under different column.

In [7]:

```
# Determine the days being reported.
Date = c()
Time = c()
for (date in df[['Date.and.Time']]){
    # Split date and time by space.
    day = strsplit(date, ' ')
    # Index the date part into Date array.
    Date = c(Date, day[[1]][1])
    Time = c(Time, day[[1]][2])
}
df['Date'] = Date
df['Time'] = Time
uni_date = unique(Date)
```

Even though there are 402 data totally, they might come from a same day. It would be useful for us to know how these data is distributed into how many days.

In [8]:

```
length(uni_date)
```

```
24
```

Data Analytics

> **Assumption 1**: Confirmed cases showed up `uniformly and independently` over 24 hours while ignoring the distribution of actual occurrence cases.

Parameter discription

- If θ is contained in a function, n would be the total sample numbers, and $\hat{\theta}$ would be the estimator for actual maximum ID.
- If M is contained in a function, k would be the total sample numbers, and N would be the actual maximum ID. As for M, that is a random variable of maximum ID in a random sample.

Method 1: Probability of each sample

The estimator used to predict the maximum value can also be determined by assuming that the probability of getting each sample is uniform where θ represents the actual maximum ID in each day.

$$P(x) = \frac{1}{\theta}$$

Method 2: Probability of maximum sample

According to the `Assumption 1`, we consider the observed maximum ID as a r.v `M`, and take the maximum ID we encountered in one specific day as `m` (i.e. $x_{n:n}$). Assume that the `N` is the actual maximum ID and `k` represents the number of ill sample, the probability mass function (PMF) of getting the maximum ID can be expressed as follows:

$$P(M = m) = \frac{C_{k-1}^{m-1}}{C_k^N}$$

Point Estimate

Estimators intuited from discrete uniform distribution

Estimator1: 2*Mean-1

Consider continuous distribution for this problem, i.e. $UNIF(0, \theta)$

$For\ UNIF(0, \theta), E(X) = \frac{\theta}{2}, Var(X) = \frac{\theta^2}{12}$

We consider the following estimator:

$$\hat{\theta}_1 = \frac{2}{n} \sum_{i=1}^{n} X_i - 1\ for\ discrete\ distrubution$$

$$\hat{\theta}_1 = \frac{2}{n} \sum_{i=1}^{n} X_i\ for\ continuous\ distrubution$$

$$E(\hat{\theta}_1) = E(\frac{2}{n} \sum_{i=1}^{n} X_i - 1) = 2E(\overline{X}) = \theta$$

$$Var(\hat{\theta}_1) = \frac{4}{n^2} \sum_{i=1}^{n} Var(X_i) = \frac{4}{n^2} \sum_{i=1}^{n} \frac{\theta^2}{12} = \frac{\theta^2}{3n}$$

$$\therefore \hat{\theta}_1\ is\ an\ unbiased\ estimator\ with\ Var = \frac{\theta^2}{3n}$$

In [9]:

```
th_1 = function(id_day){
    return(2 * sum(id_day) / length(id_day) - 1)
}

th_1_e = function(id_day){
    return (max(id_day))
}

th_1_var = function(id_day){
    return ((max(id_day)^2) / (3*length(id_day)))
}
```

Estimator2: Max + Avg GAP

Consider other form of improvement from MLE estimator, i.e. using average approach to estimate the GAP between maximum and the upper limit:

$$\widehat{\theta}_2 = X_{n:n} + \frac{1}{n-1}\sum_{i>j}(X_i - X_j - 1) \qquad \ldots for\ discrete\ case$$

$$\widehat{\theta}_2 = X_{n:n} + \frac{1}{n-1}\sum_{i>j}(X_i - X_j) \qquad \ldots for\ continuous\ case$$

Calculate the expected value and variance to determine if this estimator is biased or not.

$$E(\widehat{\theta}_2) = E(X_{n:n}) + \frac{1}{n-1}\sum_{i>j}E(X_i - X_j) = \frac{n\theta}{n+1}$$

$$Var(\hat{\theta}_2) = \frac{n\theta^2}{(n+1)(n-1)(n+2)}$$

Therefore, θ_2 is a biased estimator.

In [10]:

```
th_2 = function(id_day){
    delta = c()
    for (i in 1:length(id_day)){
        if (i == length(id_day)) break
        delta = c(delta, sort(id_day)[i+1] - sort(id_day)[i] - 1)
    }
    return (max(id_day) + sum(delta) / (length(id_day) - 1))
}

th_2_e = function(id_day){
    return (length(id_day) * max(id_day) / (length(id_day) + 1))
}

th_2_var = function(id_day){
    n = length(id_day); th = max(id_day)
    return (n*(th^2) / ((n+1)*(n-1)*(n+2)))
}
```

Estimator3: Min+max estimator

We know that maximum sample ID is what's closed to the upper limit, and we could add more information to it. Intuitively, we first consider minimum sample ID + maximum sample ID:

$$\widehat{\theta}_3 = x_{1:n} + x_{n:n}$$

$$F_{X_{n:n}}(x) = [F_X(x)]^n = \frac{x^n}{\theta^n}, f_{X_{n:n}}(x) = n\frac{x^{n-1}}{\theta^n}$$

$$E[X_{n:n}] = \int xn\frac{x^{n-1}}{\theta^n}dx = \frac{n}{n+1}\theta$$

$$E[X_{n:n}^2] = \int x^2 n\frac{x^{n-1}}{\theta^n}dx = \frac{n}{n+2}\theta^2$$

$$F_{X_{1:n}}(x) = 1 - [1 - F_X(x)]^n = 1 - (\frac{\theta - x}{\theta})^n, f_{X_{1:n}}(x) = \frac{n(\theta - x)^{n-1}}{\theta^n}$$

$$E[X_{1:n}] = \int x\frac{n(\theta - x)^{n-1}}{\theta^n}dx = \frac{1}{n+1}\theta$$

$$E[X_{1:n}^2] = \int x^2\frac{n(\theta - x)^{n-1}}{\theta^n}dx = \frac{2}{n(n+1)}\theta^2$$

$$E(\widehat{\theta}_3) = E(x_{1:n}) + E(x_{n:n}) = \theta$$

$$Var(\hat{\theta}_3) = Var(X_{1:n}) + Var(X_{n:n}) + 2Cov(X_{1:n}, X_{n:n})$$

$$= \frac{2}{n(n+1)}\theta^2 - (\frac{1}{n+1}\theta)^2 + \frac{n}{n+2}\theta^2 - (\frac{n}{n+1}\theta)^2 + 2Cov(X_{1:n}, X_{n:n})$$

$$Since\ the\ joint\ distribution\ of\ the\ order\ statistics\ of\ the\ uniform\ distribution\ is$$

$$f_{u_i,v_j}(u, v) = n!\frac{u^{i-1}}{(i-1)!}\frac{(v-u)^{j-i-1}}{(j-i-1)!}\frac{(1-v)^{n-j}}{(n-j)!}$$

$$Cov(u_k, v_j) = \frac{j(n-k-1)}{(n-1)^2(n+2)}$$

$$Var(\hat{\theta}_3) = \frac{2\theta^2}{n(n+2)} + \frac{2n^2\theta^2}{(n+1)^2(n+2)}$$

Therefore, θ_3 is a biased estimator.

In [11]:

```
th_3 = function(id_day){
    return (min(id_day) + max(id_day))
}

th_3_e = function(id_day){
    return (max(id_day))
}

th_3_var = function(id_day){
    th = max(id_day); n = length(id_day)
    return (2*(th^2) / (n*(n+2)) + 2*(n^2)*(th^2) / (((n+1)^2)*(n+2)))
}
```

Estimator4: Mean + 3 Std estimator

$$\widehat{\theta}_4 = \frac{\sum_{i=1}^{n} X_i}{n} + 3\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} = \frac{\sum_{i=1}^{n} X_i}{n} + 3S$$

$$E(\widehat{\theta}_4) = E[\overline{X}] + 3 * E(S) = \frac{1}{2}\theta + 3\frac{\theta}{2\sqrt{3}}$$

$$Var(\widehat{\theta}_4) > Var(\overline{X}) = \frac{\theta^2}{12n}$$

$$\therefore \widehat{\theta}_4 \ is\ a\ biased\ estimator$$

In [12]:

```
th_4 = function(id_day){
    s2 = sum((id_day - mean(id_day))^2) / (length(id_day) - 1)
    return (mean(id_day) + 3*sqrt(s2))
}

th_4_e = function(id_day){
    return (max(id_day) * 0.5 + 3 * max(id_day) / (2*sqrt(3)))
}

th_4_var = function(id_day){
    return ((max(id_day)^2) / (12*length(id_day)))
}
```

MLE estimator

$$
\begin{aligned}
f_{x_1,x_2,\ldots,x_n}(x_1,x_2,\ldots,x_n) &= \prod_{i=1}^{n} f_{x_i}(x_i) \\
&= \prod_{i=1}^{n} \frac{1}{\theta} \\
&= \frac{1}{\theta^n} \\
ln(f_{x_1,x_2,\ldots,x_n}(x_1,x_2,\ldots,x_n)) &= -nln\theta \\
\therefore \widehat{\theta}_{MLE} &= X_{n:n} \\
F_{X_{n:n}}(x) &= [F_X(x)]^n = \frac{x^n}{\theta^n}, f_{X_{n:n}}(x) = n\frac{x^{n-1}}{\theta^n} \\
E[X_{n:n}] &= \int xn\frac{x^{n-1}}{\theta^n}dx = \frac{n}{n+1}\theta \\
E[X_{n:n}^2] &= \int x^2 n\frac{x^{n-1}}{\theta^n}dx = \frac{n}{n+2}\theta^2 \\
E(\widehat{\theta}_{MLE}) &= E(X_{n:n}) = \frac{n}{n+1}\theta \\
Var(\widehat{\theta}_{MLE}) &= \frac{n}{n+2}\theta^2 - (\frac{n}{n+1}\theta)^2 = \frac{n}{(n+1)^2(n+2)}\theta^2 \\
&\therefore the\ MLE\ estimator\ is\ an\ biased\ estimator. \\
f(x) &= \begin{cases} \frac{1}{\theta}, & 0 < x < \theta \\ 0, & 0.w. \end{cases} \\
f(x) &= \frac{1}{\theta} I_{0,\theta}(x) \\
f(x_1,\ldots,x_n) &= \frac{1}{\theta^n}\prod_{i=1}^{n} I_{0,\theta}(x_i) \\
&= \frac{1}{\theta^n} I_{0,\theta}(x_{n:n}) \\
&= g(x_{n:n},\theta) * h(x_1,\ldots,x_n) \\
\therefore\ \ S &= x_{n:n}\ is\ sufficient\ for\ \theta \\
&\therefore the\ MLE\ estimator\ is\ sufficient.
\end{aligned}
$$

In [13]:

```
th_mle = function(id_day){
    return (max(id_day))
}

th_mle_e = function(id_day){
    return (length(id_day) * max(id_day) / (length(id_day) + 1))
}

th_mle_var = function(id_day){
    n = length(id_day); th = max(id_day)
    return (n*(th^2) / (((n+1)^2) * ((n+2)^2)))
}
```

Estimator5: An improvement from MLE: UMVUE

UMVUE under method1

Consider an improved estimator from $\hat{\theta}$, we have:

$$\hat{\theta}_5 = \frac{n+1}{n} x_{n:n} - 1 \qquad \ldots for\ discrete\ case.$$

$$\hat{\theta}_5 = \frac{n+1}{n} x_{n:n} \qquad \ldots for\ continuous\ case.$$

Get the variance to determine if the estimator has been improved to an unbiased estimator where $Var(x) = \mathbb{E}(x^2) - [\mathbb{E}(x)]^2$.

$$E(\hat{\theta}_5) = \theta$$

$$E(X_{n:n}^2) = \int x^2 \cdot \frac{n \cdot x^{n-1}}{\theta^n} dx = \frac{n\theta^2}{n+2}$$

$$Var(\hat{\theta}_5) = \left(\frac{n+1}{n}\right)^2 \cdot Var(X_{n:n})$$

$$= \left(\frac{n+1}{n}\right)^2 \left(\frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2\right) = \frac{\theta^2}{n(n+2)}$$

Therefore, $\hat{\theta}_5$ is an unbiased estimator.

To obtain the maximum value of θ, the best option is to get the maximum value of each random sample `m` (i.e. $\hat{\theta} = x_{n:n}$). According to the `Ex 3.4`, we know that

$$F_{x_{n:n}}(x) = [F_x(x)]^n = \frac{x^n}{\theta^n} \quad \rightarrow \quad f_{x_{n:n}}(x) = \frac{n \cdot x^{n-1}}{\theta^n}$$

where `n` is the total number of sample in each day. To obtain the expected value of `m` (i.e. $x_{n:n}$), we can determine if the estimator is an biased estimator such that:

$$\mathbb{E}(x_{n:n}) = \int x \cdot \frac{n \cdot x^{n-1}}{\theta^n} dx = \frac{n\theta}{n+1} = \mathbb{E}(\hat{\theta}) \neq \theta$$

Therefore, the estimator is an biased estimator.

$$f(x) = \begin{cases} \frac{1}{\theta} & 0 < x < \theta \\ 0 & o.w. \end{cases}$$

$$f(x) = \frac{1}{\theta} I_{0,\theta}(x)$$

$$\begin{aligned} f(x_1, \ldots, x_n) &= \frac{1}{\theta^n} \prod_{i=1}^{n} I_{0,\theta}(x_i) = \frac{1}{\theta^n} I_{0,\theta}(x_{n:n}) \\ &= g(x_{n:n}, \theta) \cdot h(x_1, \ldots, x_n) \end{aligned}$$

Therefore, $S = x_{n:n}$ is sufficient for θ. According to `Lehmann-Scheffe` Theorem, we have:

$$T = \hat{\theta}_2 = \frac{n+1}{n} X_{n:n}$$

which is unbiased for $\tau(\theta) = \theta$. Thus, T is `UMVUE`.

UMVUE under method2

The expected value of $M = m$ can be calculated as follows where $C_k^N = \frac{N!}{k!(N-k)!}$, $\sum_{m=k}^N C_k^m = C_{k+1}^{N+1}$:

$$\begin{aligned}
\mathbb{E}(M = m) &= \sum_{m=k}^{N} m \cdot P(m) \\
&= \sum_{m=k}^{N} m \cdot \frac{\frac{(m-1)!}{(k-1)!(m-k)!}}{\frac{N!}{k!(N-k)!}} \\
&= \sum_{m=k}^{N} \frac{m!k(N-k)!k!}{k!(m-k)!N!} \\
&= \frac{k(N-k)!k!}{N!} \cdot \sum_{m=k}^{N} C_k^m \\
&= \frac{k(N-k)!k!}{N!} \cdot \frac{(N+1)!}{(k+1)!(N-k)!} = \frac{k(N+1)}{k+1}
\end{aligned}$$

Since we are looking for the maximum ID from our observation, the best guess of `M` should be the maximum ID of THAT particular day `m`. Therefore, we get:

$$m = \frac{k(N+1)}{k+1} \quad \rightarrow \quad k\hat{N} = mk + m - k \quad \rightarrow \quad \hat{N} = m + \frac{m}{k} - 1$$

By Finding the expected value of $\hat{N}$, we have:

$$\begin{aligned}
\mathbb{E}(\hat{N}) = \mathbb{E}\left[\mathbb{E}(M) + \frac{\mathbb{E}(M)}{k} - 1\right] &= \mathbb{E}(M) + \frac{\mathbb{E}(M)}{k} - 1 \\
= \frac{k(N+1)}{k+1} + \frac{N+1}{k+1} - \frac{k+1}{k+1} &= \frac{N(k+1)}{k+1} = N
\end{aligned}$$

Therefore, $\hat{N}$ is proved to be unbiased.

In [14]:

```
th_5 = function(id_day){
    return (max(id_day) + max(id_day) / length(id_day) - 1)
}

th_5_e = function(id_day){
    return (max(id_day))
}

th_5_var = function(id_day){
    return ((max(id_day)^2) / (length(id_day) + 2))
}
```

Estimator6: Bayes Estimator

The Bayesian approach is to consider the credibility $P(N = n|M = m, K = k)$ that the maximum random ID N is equal to the number n, and the maximum observed serial number M is equal to the number m. Consider Conditional probability rule instead of using a proper prior distribution.

$$P(n|m, k)P(m|k) = P(m|n, k)P(n|k) = P(m, n|k)$$

$P(m|n, k)$ answers the question: "What is the probability of a specific serial number m being the highest number observed in a sample of k patients, given there are n in total?" The probability of this occurring is:

$$P(m|n, k) = k\frac{(n-k)!}{n!}\frac{(m-1)!}{(m-k)!} = \frac{C_{k-1}^{m-1}}{C_k^n} I_{k\le m} I_{m\le n}$$

$P(m|k)$ is the probability that the maximum serial number is equal to m once k tanks have been observed but before the serial numbers have actually been observed.

$$\begin{aligned} P(m|k) =& P(m|k) * \sum_{n=0}^{\infty} P(n|m, k) \\ =& P(m|k) * \sum_{n=0}^{\infty} \frac{P(m|n, k)P(n|k)}{P(m|k)} \\ =& \sum_{n=0}^{\infty} P(m|n, k)P(n|k) \end{aligned}$$

$P(n|k)$ is the credibility that the total number of tanks, N, is equal to n when the number K patients observed is known to be k, but before the serial numbers have been observed. Assume that it is some discrete uniform distribution:

$$\begin{aligned} P(n|k) =& \frac{1}{\Omega - k} I_{k\le\Omega} I_{n\le\Omega} \\ & where\ \ \Omega\ is\ the\ upper\ limit\ and\ it's\ finite \end{aligned}$$

$$
\begin{aligned}
P(n|m,k) =& \frac{P(m|n,k)P(n|k)}{P(m|k)} \\
=& \frac{P(m|n,k)P(n|k)}{\sum_{n=0}^{\infty} P(m|n,k)P(n|k)}, k \leq m, m \leq n, k \leq \Omega, n \leq \Omega \\
=& \frac{P(m|n,k)}{\sum_{n=m}^{\Omega-1} P(m|n,k)} I_{m \leq n} I_{n \leq \Omega} \\
for\ k \geq 2,\ P(n|m,k) =& \frac{P(m|n,k)}{\sum_{n=m}^{\infty} P(m|n,k)} I_{m \leq n} \\
=& \frac{\frac{C_{k-1}^{m-1}}{C_k^n}}{\sum_{n=m}^{\infty} \frac{C_{k-1}^{m-1}}{C_k^n}} I_{m \leq n} \\
=& \frac{k-1}{k} \frac{C_{k-1}^{m-1}}{C_k^n} I_{m \leq n} \\
P(N > x|M=m, K=k) =& \sum_{n=x+1}^{\infty} P(n|m,k) I_{m \leq x} \\
=& I_{m<x} + I_{m \geq x} \sum_{n=x+1}^{\infty} \frac{k-1}{k} \frac{C_{k-1}^{m-1}}{C_k^n} \\
=& I_{m<x} + I_{m \geq x} \frac{k-1}{k} \frac{C_{k-1}^{m-1}}{1} \sum_{n=x+1}^{\infty} \frac{1}{C_k^n} \\
=& I_{m<x} + I_{m \geq x} \frac{k-1}{k} \frac{C_{k-1}^{m-1}}{1} \frac{k}{k-1} \frac{1}{C_{k-1}^x} \\
=& I_{m<x} + I_{m \geq x} \frac{C_{k-1}^{m-1}}{C_{k-1}^x} \\
P(N \leq x|M=m, K=k) =& 1 - P(N > x|M=m, K=k) \\
=& I_{m \geq x}(1 - \frac{C_{k-1}^{m-1}}{C_{k-1}^x}) \\
\mu_{Bayes} =& \sum_n nP(n|m,k) \\
=& \sum_n \frac{k-1}{k} \frac{C_{k-1}^{m-1}}{C_k^n} I_{m \leq n} \\
=& \sum_n \frac{m-1}{n} \frac{C_{k-2}^{m-2}}{C_{k-1}^{n-1}} I_{m \leq n} \\
=& (m-1) C_{k-2}^{m-2} \sum_{n \geq n} \frac{1}{C_{k-1}^{n-1}} \\
=& (m-1) C_{k-2}^{m-2} \frac{k-1}{k-2} \frac{1}{C_{k-2}^{m-2}} \\
=& \frac{(m-1)(k-1)}{k-2}
\end{aligned}
$$

Hence, $\mu_{Bayes} = \frac{(X_{n:n}-1)(n-1)}{n-2}$ using the standard of writing in other chapter.

$E(\mu_{Bayes}) = \frac{n}{n+2}\theta$ The Bayes estimator is a biased one. To measure its uncertainty, we calculate its variance:

$$\begin{aligned}
\mu^2 + \sigma^2 - \mu &= \sum_n n(n-1)P(n|m,k) \\
&= \sum_n n(n-1)\frac{m-1}{n}\frac{m-2}{n-1}\frac{k-1}{k-2}\frac{C_{k-3}^{m-3}}{C_{k-2}^{n-2}} I_{m \le n} \\
&=(m-1)(m-2)\frac{k-1}{k-2}C_{k-3}^{m-3}\sum_{n \ge m}\frac{1}{C_{k-2}^{n-2}} \\
&=(m-1)(m-2)\frac{k-1}{k-2}C_{k-3}^{m-3}\frac{k-2}{k-3}\frac{1}{C_{k-3}^{m-3}} \\
&=\frac{(m-1)(m-2)(k-1)}{k-3} \\
\sigma_{Bayes}^2 &=\frac{(m-1)(m-2)(k-1)}{k-3} - (\frac{(m-1)(k-1)}{k-2})^2 + \frac{(m-1)(k-1)}{k-2} \\
&=\frac{(m-1)(k-1)(m+1-k)}{(k-3)(k-2)^2} \\
Var(\widehat{\theta}_{Bayes}) &=\frac{(x_{n:n}-1)(n-1)(x_{n:n}+1-n)}{(n-3)(n-2)^2}
\end{aligned}$$

In [15]:

```
th_bayes = function(id_day){
    upper = (max(id_day) - 1) * (length(id_day) - 1)
    return (upper / (length(id_day) - 2))
}

th_bayes_e = function(id_day){
    return (length(id_day)*max(id_day) / (length(id_day)+2))
}

th_bayes_var = function(id_day){
    upper = (max(id_day) - 1)*(length(id_day) - 1)*(max(id_day) + 1 - length(id_
day))
    return (upper / ((length(id_day) - 3)*((length(id_day) - 2)^2)))
}
```

Point Estimation Conclusion

According to the distribution of the question, we find six possible estimators, in which four of them are intuitive from the distribution and the background of the question, one is maximum likelihood estimator (MLE) and one the Bayes estimator and the improved estimator from MLE. We proved that the improved estimator from MLE is exactly uniformly minimum-variance unbiased estimator (UMVUE) which is unbiased estimator with the smallest variance.

Also, what is most important in our findings is the $X_{n:n}$ plays an important role in estimating the upper limit of th discrete uniform distribution since the maximum sample give the closest information of the upper limit intuitively and we also proved that it' s the sufficient statistics to estimate N.

To compare the unbiasedness, effectiveness of the estimators we find, we summarize the results and give the following table:

No.	$Function$	$E(\widehat{\theta})$	$Var(\widehat{\theta})$
$\widehat{\theta}_1$	$\frac{2}{n}\sum_{i=1}^{n} X_i - 1$	θ	$\frac{\theta^2}{3n}$
$\widehat{\theta}_2$	$X_{n:n} + \frac{1}{n-1}\sum_{i>j}(X_i - X_j - 1)$	$\frac{n\theta}{n+1}$	$\frac{n\theta^2}{(n+1)(n-1)(n+2)}$
$\widehat{\theta}_3$	$x_{1:n} + x_{n:n}$	θ	$\frac{2\theta^2}{n(n+2)} + \frac{2n^2\theta^2}{(n+1)^2(n+2)}$
$\widehat{\theta}_4$	$\frac{\sum_{i=1}^{n} X_i}{n} + 3\sqrt{\frac{\sum E(X_i-\overline{X})^2}{n-1}}$	$\frac{1}{2}\theta + \frac{2\sqrt{3}}{3}\theta$	$Var(\widehat{\theta}_4) > \frac{\theta^2}{12n}$
$\hat{\theta}_5$	$\frac{n+1}{n}x_{n:n} - 1$	θ	$\frac{\theta^2}{n(n+2)}$
$\widehat{\theta}_{MLE}$	$X_{n:n}$	$\frac{n}{n+1}\theta$	$\frac{n}{(n+1)^2(n+2)}\theta^2$
$\hat{\theta}_{Bayes}$	$\frac{(X_{n:n}-1)(n-1)}{n-2}$	$\frac{n}{n+2}\theta$	$\frac{(x_{n:n}-1)(n-1)(x_{n:n}+1-n)}{(n-3)(n-2)^2}$

Intervel Estimation

In addition to point estimation, interval estimation can be carried out. Based on the observation that the probability that *k* observations in the sample will fall in an interval covering *p* of the range (0 ≤ *p* ≤ 1) is p^k (assuming in this section that draws are with replacement, to simplify computations; if draws are without replacement, this overstates the likelihood, and intervals will be overly conservative).

Thus the sampling distribution of the quantile of the sample maximum is the graph $x^{1/k}$ from 0 to 1: the *p*-th to *q*-th quantile of the sample maximum *m* are the interval $[p^{1/k}N, q^{1/k}N]$. Inverting this yields the corresponding confidence interval for the population maximum of $[m/q^{1/k}, m/p^{1/k}]$.

In [16]:

```
interval = function(id_day){
    itv_1 = c(max(id_day) / (0.025^(1/length(id_day))),
              max(id_day) / (0.975^(1/length(id_day))))
    itv_2 = c(max(id_day) / (0.050^(1/length(id_day))),
              max(id_day) / (0.950^(1/length(id_day))))
    itv_3 = c(max(id_day) / (0.075^(1/length(id_day))),
              max(id_day) / (0.925^(1/length(id_day))))
    itv_4 = c(max(id_day) / (0.010^(1/length(id_day))),
              max(id_day) / (0.900^(1/length(id_day))))
    return (matrix((c(itv_1, itv_2, itv_3, itv_4)), nrow=4, byrow=T))
}
```

Generate More Data

> **Assumption 2**: Confirmed cases showed up `uniformly` and `independently` over 24 hours while assuming that the actual occurrence cases follows a certain distribution.

Here are the steps to iterate through data:

1. The random ID would be generated under certain distribution.
2. The new sample ID would update the sample mean $\bar{x}$.
3. The new sample ID might also update the maximum ID.
4. the new $\bar{x}$ and maximum ID can further update the parameter of a assumed distribution.
5. The random ID would be generated under an adjusted distribution. (Go back to the 2^{nd} step.

Our code will iterate the above 5 steps for several times until the parameters are not updated.

1. Poisson Distribution

Let us assum that the unknown distribution is a `Poisson` distribution: $POI(\lambda)$ where:

$$P(X = x_{id}) = \frac{e^{-\lambda}\lambda^{x_{id}}}{x_{id}!}$$

We can apply MLE method under log space to determine λ as follows:

$$\mathcal{L}(\lambda) = \ln \prod_{i=1}^{n} P(x_{id_i}|\lambda) = \sum_{i=1}^{n} \ln\left(\frac{e^{-\lambda}\lambda^{x_{id_i}}}{x_{id_i}}\right) = -n\lambda + \ln\lambda \sum_{i=1}^{n} x_{id_i} - \sum_{i=1}^{n} \ln x_{id_i}!$$

By differentiating this log function, we can get the optimum estimator.

$$\frac{\partial\mathcal{L}(\lambda)}{\partial\lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_{id_i} = 0 \quad \rightarrow \quad \hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_{id_i}$$

In [17]:

```r
lambda_poi = function(id_day){
    return (sum(id_day) / length(id_day))
}
```

2. Binomial Distribution

If we assum that the unknown distribution is a `Binomial` distribution: $BIN(n, p)$ where:

$$P(X = x_{id}) = f(x_{id}, N, p) = C_{x_{id}}^{N} p^{x_{id}}(1-p)^{N-x_{id}}$$

Apply MLE method under log space, we get:

$$\begin{aligned}\mathcal{L}(p) &= \ln \prod_{i=1}^{n} f(x_{id_i}, N, p) \\ &= \sum_{i=1}^{n}\left[\ln\left(\frac{N!}{x_{id_i}!(N - x_{id_i})!}\right) + x_{id_i}\ln p + (N - x_{id_i})\ln(1-p)\right]\end{aligned}$$

By differentiating this equation, we can get the optimal value.

$$\frac{\partial\mathcal{L}(p)}{\partial p} = 0 + \frac{1}{p}\sum_{i=1}^{n} x_{id_i} + \frac{-1}{1-p}\left(nN - \sum_{i=1}^{n} x_{id_i}\right) = 0$$

$$\sum_{i=1}^{n} x_{id_i} - p\sum_{i=1}^{n} x_{id_i} - pnN + p\sum_{i=1}^{n} x_{id_i} = 0 \quad \rightarrow \quad \hat{p} = \frac{\sum_{i=1}^{n} x_{id_i}}{nN} = \frac{\bar{x}_{id}}{N}$$

The result finally shows that $\bar{x}_{id}$ equals to $\hat{\lambda}_{MLE}$ from Poisson distribution. Therefore, here is how we get $\hat{p}$ with code:

In [18]:

```r
p_hat = function(id_day){
    return (mean(id_day) / max(id_day))
}
```

Exprimental Results

First of all, set up a series of empty array for containing the computed maximum ID from different estimators and their corresponding variance.

In [19]:

```
TH_1 = c(); TH_2 = c(); TH_3 = c(); TH_4 = c()
TH_5 = c(); TH_mle = c(); TH_bayes = c()
```

In [20]:

```
TH_1_var = c(); TH_2_var = c(); TH_3_var = c(); TH_4_var = c()
TH_5_var = c(); TH_mle_var = c(); TH_bayes_var = c()
```

1. Point Estimation

In [21]:

```
for (date in uni_date){
    id_day = as.integer(matrix(df[df['Date'] == date], ncol=4)[, 2])

    TH_1 = c(TH_1, round(th_1(id_day), digit=0))
    TH_2 = c(TH_2, round(th_2(id_day), digit=0))
    TH_3 = c(TH_3, round(th_3(id_day), digit=0))
    TH_4 = c(TH_4, round(th_4(id_day), digit=0))
    TH_5 = c(TH_5, round(th_5(id_day), digit=0))
    TH_mle = c(TH_mle, round(th_mle(id_day), digit=0))
    TH_bayes = c(TH_bayes, round(th_bayes(id_day), digit=0))

    TH_1_var = c(TH_1_var, th_1_var(id_day))
    TH_2_var = c(TH_2_var, th_2_var(id_day))
    TH_3_var = c(TH_3_var, th_3_var(id_day))
    TH_4_var = c(TH_4_var, th_4_var(id_day))
    TH_5_var = c(TH_5_var, th_5_var(id_day))
    TH_mle_var = c(TH_mle_var, th_mle_var(id_day))
    TH_bayes_var = c(TH_bayes_var, th_bayes_var(id_day))
}
```

The following is our experimental result based on the point estimation:

In [22]:

```
matrix(c(TH_1, TH_2, TH_3, TH_4, TH_5, TH_mle, TH_bayes), ncol=7, byrow=F)
```

A matrix: 24 × 7 of type dbl

```
242  293  291  341  293  261  303
417  262  414  300  292  244  324
443  334  412  427  347  298  371
279  275  292  350  278  239  298
281  264  257  364  263  238  271
392  345  370  433  347  323  351
322  323  323  435  323  302  326
379  328  328  475  328  315  329
201  291  268  321  289  268  294
325  299  295  456  299  292  300
264  271  271  369  271  265  272
218  237  234  333  237  232  238
171  192  188  263  192  188  192
159  156  151  219  156  151  156
108  110  113  151  110  106  111
119  114  111  155  113  110  114
 95   90   88  126   90   87   90
 56   66   69   83   66   63   67
 59   52   51   76   52   48   53
 29   41   40   48   41   39   42
  7   18   15   23   17   15   19
 27   24   22   35   23   22   24
 15   11   15   16   12   11   12
  8    8    9   12    8    8    8
```

In [23]:

```
c(sum(TH_1), sum(TH_2), sum(TH_3), sum(TH_4),
  sum(TH_5), sum(TH_mle), sum(TH_bayes))
```

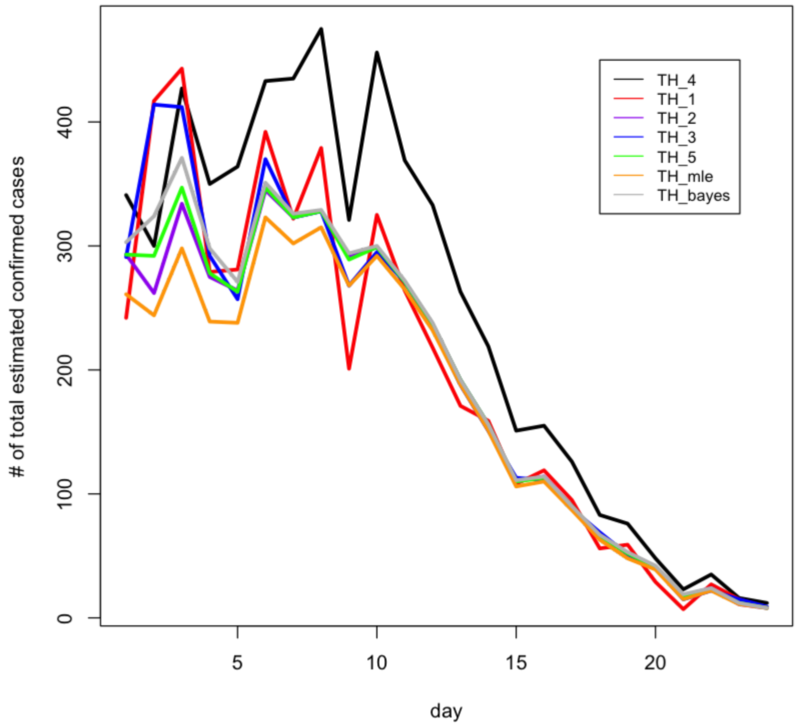
```
4616 ·  4404 ·  4627 ·  5811 ·  4447 ·  4125 ·  4565
```

Here, we could conclude TN in the following table:

No.	*Function*	TN
$\hat{\theta}_1$	$\frac{2}{n}\sum_{i=1}^{n} X_i - 1$	4616
$\hat{\theta}_2$	$X_{n:n} + \frac{1}{n-1}\sum_{i>j}(X_i - X_j - 1)$	4404
$\hat{\theta}_3$	$x_{1:n} + x_{n:n}$	4627
$\hat{\theta}_4$	$\frac{\sum_{i=1}^{n} X_i}{n} + 3\sqrt{\frac{\sum E(X_i - \overline{X})^2}{n-1}}$	5811
$\hat{\theta}_5$	$\frac{n+1}{n} x_{n:n} - 1$	4447
$\hat{\theta}_{MLE}$	$X_{n:n}$	4125
$\hat{\theta}_{Bayes}$	$\frac{(X_{n:n}-1)(n-1)}{n-2}$	4565

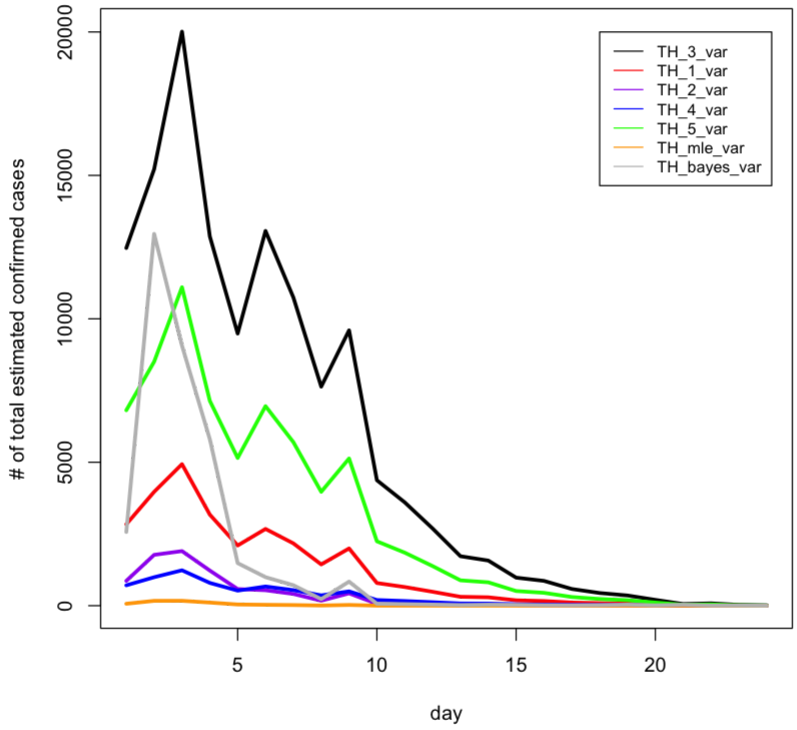
In [24]:

```r
plot(TH_4, type='l', lwd=3, xlab='day',
     ylab='# of total estimated confirmed cases')
lines(TH_1, col="red", lwd=3)
lines(TH_2, col='purple', lwd=3)
lines(TH_3, col='blue', lwd=3)
lines(TH_5, col='green', lwd=3)
lines(TH_mle, col='orange', lwd=3)
lines(TH_bayes, col='gray', lwd=3)
legend(18, 450, legend=c("TH_4", "TH_1", "TH_2", "TH_3",
                         "TH_5", "TH_mle", "TH_bayes"),
       col=c("black", "red", "purple", "blue", "green", "orange", "gray"),
       lty=1, cex=0.8)
```



In [25]:

```
plot(TH_3_var, type='l', lwd=3, xlab='day',
     ylab='# of total estimated confirmed cases')
lines(TH_1_var, col="red", lwd=3)
lines(TH_2_var, col='purple', lwd=3)
lines(TH_4_var, col='blue', lwd=3)
lines(TH_5_var, col='green', lwd=3)
lines(TH_mle_var, col='orange', lwd=3)
lines(TH_bayes_var, col='gray', lwd=3)
legend(18, 20000, legend=c("TH_3_var", "TH_1_var", "TH_2_var",
                           "TH_4_var", "TH_5_var", "TH_mle_var",
                           "TH_bayes_var"),
       col=c("black", "red", "purple", "blue", "green", "orange", "gray"),
       lty=1, cex=0.8)
```



2. Intervel Estimation

In [26]:

```
id_day = as.integer(matrix(df[df['Date'] == uni_date[1]], ncol=4)[, 2])
```

In [27]:

```
interval(id_day)
```

```
A matrix: 4 × 2 of type
dbl

 413.9025   261.8273
 379.5502   262.6788
 360.7928   263.5559
 464.1309   264.4601
```

In [28]:

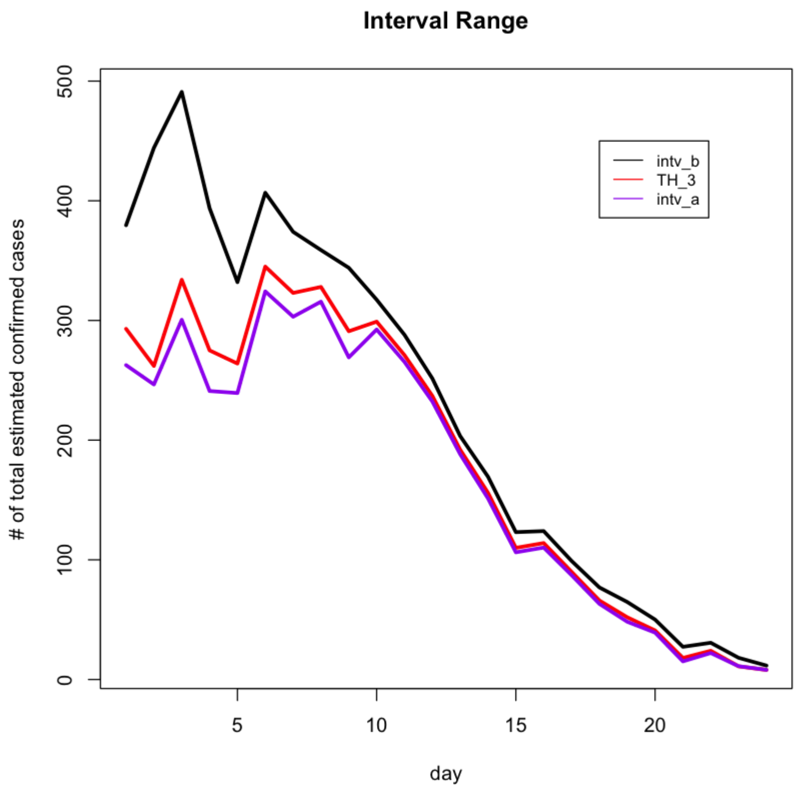
```
intv_b = c(); intv_a = c()
for (date in uni_date){
    id_day = as.integer(matrix(df[df['Date'] == date], ncol=4)[, 2])

    intv_b = c(intv_b, interval(id_day)[2, 1])
    intv_a = c(intv_a, interval(id_day)[2, 2])
}
```

In [29]:

```
plot(intv_b, type='l', lwd=3, xlab='day',
     ylab='# of total estimated confirmed cases',
     main='Interval Range')
lines(TH_2, col="red", lwd=3)
lines(intv_a, col='purple', lwd=3)

legend(18, 450, legend=c('intv_b', 'TH_3', 'intv_a'),
       col=c("black", "red", "purple"),
       lty=1, cex=0.8)
```



Generate more Data

According to the assumption we defined above, we pick up both `Poisson` and `Binomial` distribution to generate more data. There are `data_gen` number of data being produced in one day as part of our random sample.

In [30]:

```
data_gen = 10     # Generate "num" data at one iteration.
iteration = 50    # Generate data for "num" times.
```

Because the number of random sample from each day are different, we have to save the original r.s. to a list so that it would be more convenient to index the data.

In [31]:

```
ID_days = list()
for (date in uni_date){
    id = as.integer(matrix(df[df['Date'] == date], ncol=4)[, 2])
    ID_days = append(ID_days, list(id))
}
```

1. Possion Distribution

In [34]:

```
# Create an empty list to record the change of estimator in each iter.
rec_poi = list()
for (i in 1:iteration){
    # Record the transitions.
    rec_poi = append(rec_poi, list(TH_5))

    for (j in 1:length(uni_date)){
        # Update the parameter by current existing samples.
        lambda = lambda_poi(ID_days[[j]])

        # Generate more data according to poisson distribution.
        data_get = c(rpois(data_gen, lambda))
        ID_days[[j]] = append(ID_days[[j]], data_get)

        # Update the estimate by an estimator.
        TH_5[j] = round(th_5(ID_days[[j]]), digit=0)
    }
}
```

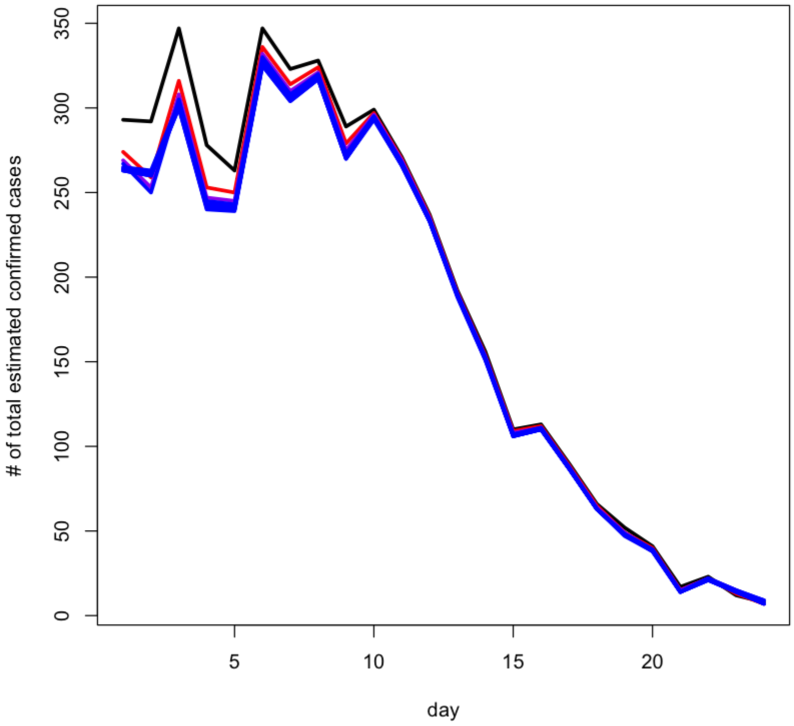
In [35]:

```
length(rec_poi)
```

```
50
```

In [33]:

```
plot(rec_poi[[1]], type='l', lwd=3, xlab='day',
     ylab='# of total estimated confirmed cases')
lines(rec_poi[[2]], col="red", lwd=3)
lines(rec_poi[[3]], col='purple', lwd=3)
lines(rec_poi[[4]], col='blue', lwd=3)
lines(rec_poi[[5]], col='blue', lwd=3)
lines(rec_poi[[6]], col='blue', lwd=3)
lines(rec_poi[[7]], col='blue', lwd=3)
lines(rec_poi[[8]], col='blue', lwd=3)
lines(rec_poi[[9]], col='blue', lwd=3)
lines(rec_poi[[10]], col='blue', lwd=3)
```



2. Binonmial Distribution

In [32]:

```
# Create an empty list to record the change of estimator in each iter.
rec_bin = list()
for (i in 1:iteration){
    # Record the transitions.
    rec_bin = append(rec_bin, list(TH_5))

    for (j in 1:length(uni_date)){
        # Update the parameter by current existing samples.
        P_hat = p_hat(ID_days[[j]])

        # Generate more data according to poisson distribution.
        data_get = c(rbinom(data_gen, TH_5[j], P_hat))
        ID_days[[j]] = append(ID_days[[j]], data_get)

        # Update the estimate by an estimator.
        TH_5[j] = round(th_5(ID_days[[j]]), digit=0)
    }
}
```

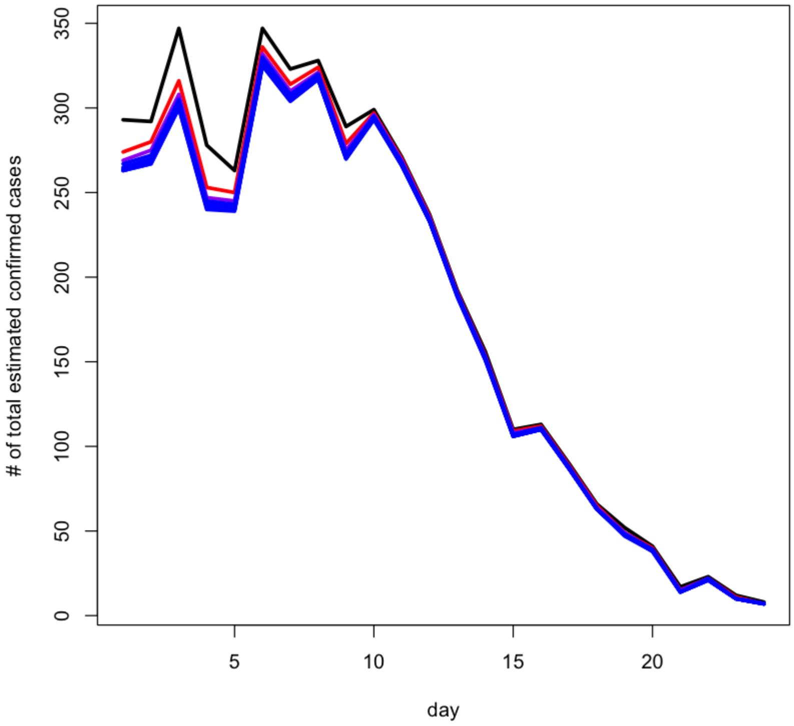
In [33]:

```
length(rec_bin)
```

```
50
```

In [34]:

```
plot(rec_bin[[1]], type='l', lwd=3, xlab='day',
     ylab='# of total estimated confirmed cases')
lines(rec_bin[[2]], col="red", lwd=3)
lines(rec_bin[[3]], col='purple', lwd=3)
lines(rec_bin[[4]], col='blue', lwd=3)
lines(rec_bin[[5]], col='blue', lwd=3)
lines(rec_bin[[6]], col='blue', lwd=3)
lines(rec_bin[[7]], col='blue', lwd=3)
lines(rec_bin[[8]], col='blue', lwd=3)
lines(rec_bin[[9]], col='blue', lwd=3)
lines(rec_bin[[10]], col='blue', lwd=3)
```



In [35]:

```
result = c()
for (tn in rec_bin){
    result = c(result, sum(tn))
}
```

In [36]:

```
result
```

```
4447 · 4293 · 4238 · 4212 · 4193 · 4182 · 4176 · 4169 · 4164 · 4159 · 4157 ·
4154 · 4150 · 4150 · 4150 · 4148 · 4145 · 4143 · 4142 · 4140 · 4137 · 4136 ·
4136 · 4136 · 4136 · 4136 · 4136 · 4136 · 4135 · 4135 · 4135 · 4135 · 4135 ·
4135 · 4134 · 4134 · 4134 · 4134 · 4134 · 4134 · 4134 · 4134 · 4134 · 4133 ·
4133 · 4133 · 4133 · 4132 · 4131 · 4131
```

The results show us the trend that the estimator would give us during the iterations. As there are more samples being generated, the result will converge to a certain number.

In [41]:

```
sum_max = c()
for (date in uni_date){
    id_day = as.integer(matrix(df[df['Date'] == date], ncol=4)[, 2])
    sum_max = c(sum_max, max(id_day))
}
```

In [42]:

```
sum(sum_max)
```

```
4125
```

Coincidentally, the converged number will increasingly equals to the sum of all maximum ID from 24 days.

Reference

- wiki/German_tank_problem (https://en.wikipedia.org/wiki/German_tank_problem)
- wiki/Order_statistic (https://en.wikipedia.org/wiki/Order_statisticn)
- wiki/Discrete_uniform_distribution (https://en.wikipedia.org/wiki/Discrete_uniform_distribution)
- wikieducator/Point*estimation*-_German_tank_problem#Maximum_value_estimator (https://wikieducator.org/Point_estimation_-_German_tank_problem#Maximum_value_estimator)