

# Applied Statistic HW2

2020270026 王姿文、2020211316 周斯萤、2020211314 徐颢轩

2020/10/15

**Question 1** 试基于以下几种思路对元件寿命参数  $\frac{1}{\lambda}$  进行估计，尽可能给出点估计的理论表达式。

## 1.1 利用取值小于T的样本值占样本容量的比例

$$P(X < T) = \int_0^T \lambda e^{-\lambda x} dx = 1 - e^{-\lambda T}$$

$$\Rightarrow \frac{1}{\hat{\lambda}} = \frac{-T}{\log(1-p_{(x < y)})}$$

## 1.2 利用样本中所有取值小于T的部分的均值

$$E(X|X < T) = \frac{1}{1-e^{-\lambda T}} \int_0^T \lambda x e^{-\lambda x} dx = \frac{1}{1-e^{-\lambda T}} \frac{1}{\lambda} \int_0^{\lambda T} \lambda x e^{-x} dx = \frac{1}{\lambda(1-e^{-\lambda T})} (-\lambda T e^{-\lambda T} + 1 - e^{-\lambda T}) = \frac{1-(\lambda T+1)e^{-\lambda T}}{\lambda(1-e^{-\lambda T})}$$

$$\Rightarrow \frac{1-(\hat{\lambda} T+1)e^{-\hat{\lambda} T}}{\hat{\lambda}(1-e^{-\hat{\lambda} T})} - m(x)_{(x < t)} = 0$$

## 1.3 利用样本均值

$$E[Y] = \frac{1-e^{-\lambda T}}{\lambda} \text{ Let } \frac{1}{\lambda} = \theta$$

$$\Rightarrow \hat{\theta}^{-1} \bar{y} + e^{-\frac{t}{\theta}} - 1 = 0$$

## 1.4 利用极大似然估计

$$\text{Likelihood function} = \frac{n!}{(n-r)!} \prod_{i=1}^k (\lambda e^{-\lambda y_{(i)}} (e^{-\lambda T})^{n-k})$$

Ignore constant,

$$\ln L = k \ln \lambda - \lambda \sum_{i=1}^k y_{(i)} - \lambda T(n-k)$$

$$\frac{\partial \ln L}{\partial \lambda} = \frac{k}{\lambda} - \sum_{i=1}^k y_{(i)} - T(n-k) = 0$$

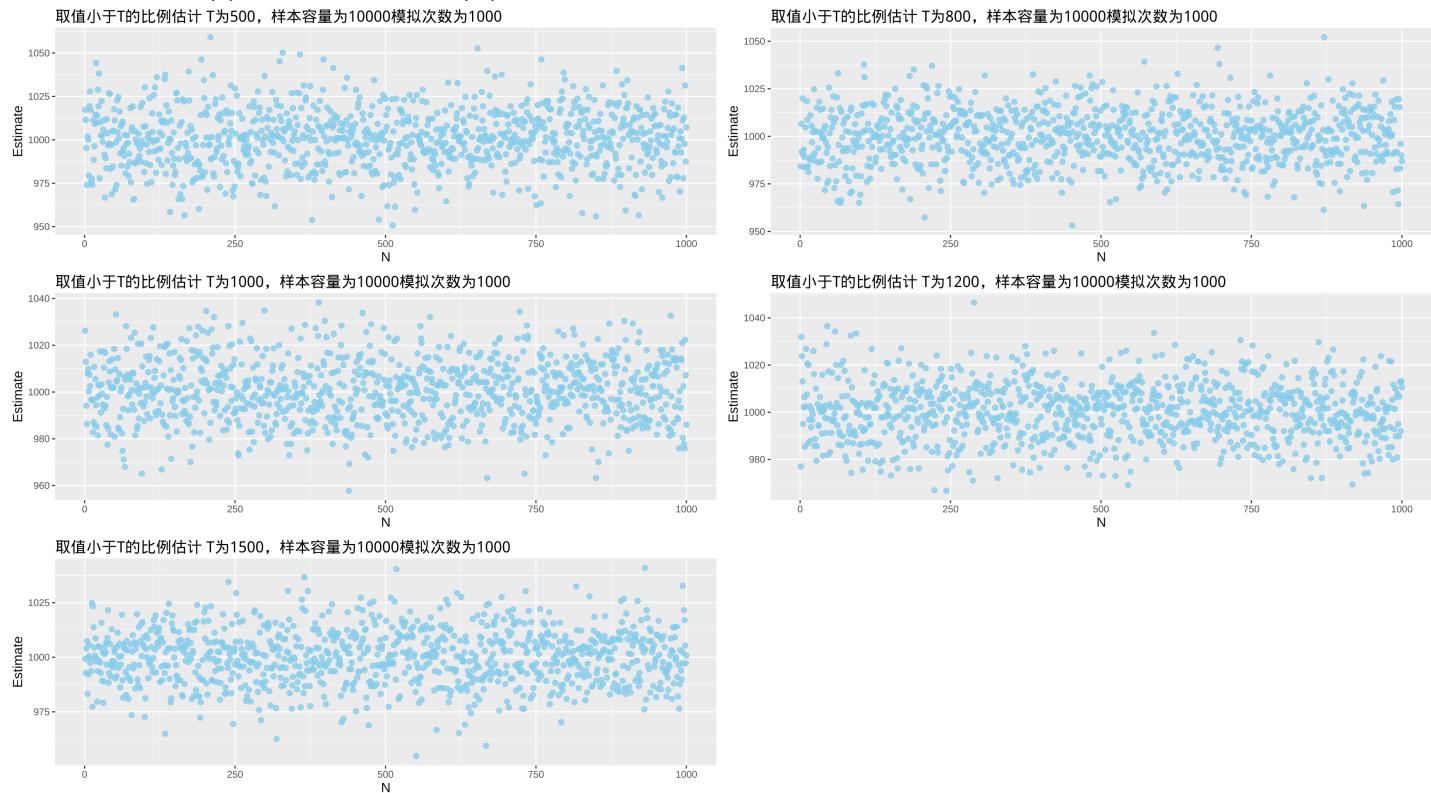
$$\Rightarrow \frac{1}{\hat{\lambda}} = \frac{\sum_{i=1}^k y_{(i)} + T(n-k)}{k} = \frac{\sum_{i=1}^k y_{(i)}}{k} \quad k \text{ 为 } n \text{ 个样本中小于 } T \text{ 的样本个数}$$

**Question 2** 设  $\frac{1}{\lambda} = \frac{1}{1000}$ , 分别取  $T=500, 800, 1000, 1200, 1500$ ，按照上述抽样规则模拟生成样本观测值，样本容量自己确定，可多尝试几种不同的样本容量，计算出问题1中四种估计方法的估计值；将每种估计方法重复模拟多次（可选取不同的次数），利用模拟结果比较各种估计方法的优劣，并且比较的取值对估计效果的影响。

## 2.1 利用取值小于T的样本值占样本容量的比例估计

i 比较T

下图为模拟次数( $n$ )=1000、样本容量( $m$ )=10000时， $T$ 分别为500,800,1000,1200,1500的图。

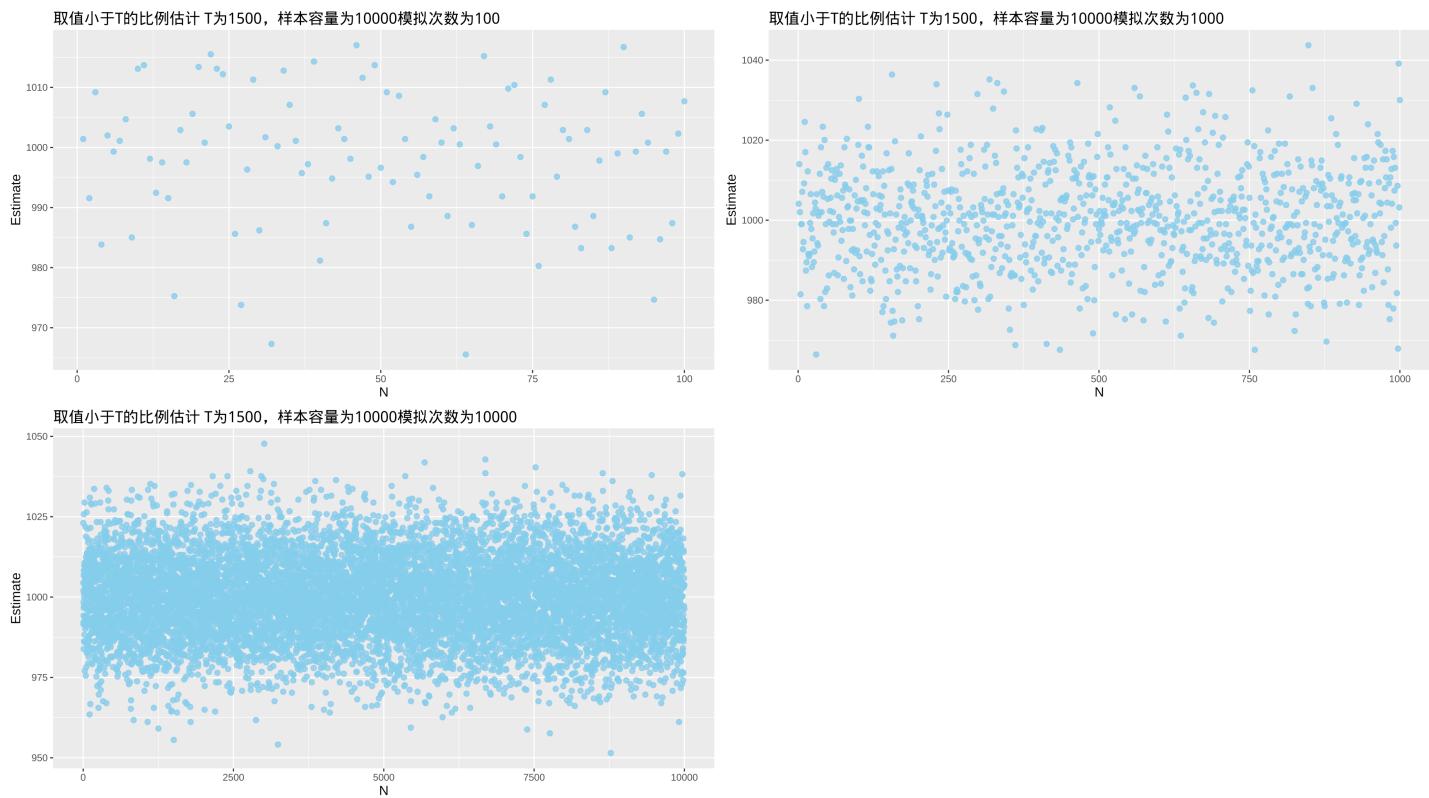


可以结合问题三，从无偏性、有效性和均方误差，以 $T$ 不同的值來比較优劣，下表为无偏性、有效性和均方误差的结果。在期望值与实际值（1000）差不多的条件下，可以看出 $T$ 越大，VAR及MSE越小，因此**T越大估计效果越好**。

模拟次数	样本容量	$T$	Expectation	Variation	MSE
1000	10000	500	999.8	270.9	270.9
1000	10000	800	999.8	194.7	194.7
1000	10000	1000	999.8	165.4	165.4
1000	10000	1200	1001.0	164.4	165.4
1000	10000	1500	1000.0	160.0	160.0

## ii 比較模拟次数( $n$ )

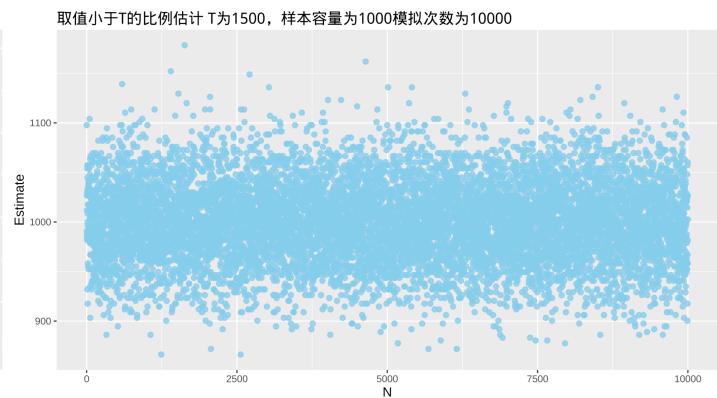
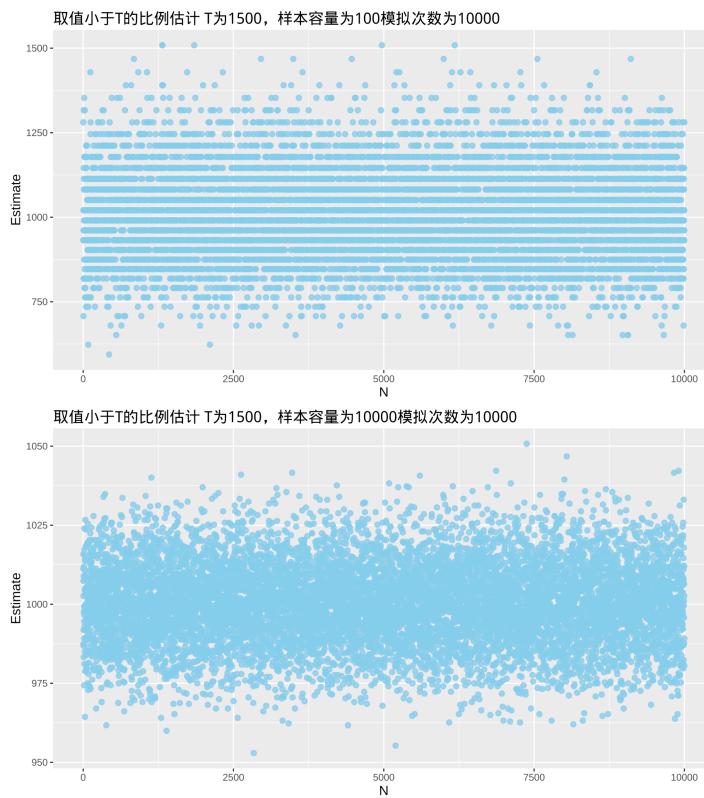
以下我们则实验在固定 $T=1500$ 的条件下，样本容量( $m$ )=10000不变，分别取模拟次数( $n$ )=100,1000,10000。从下表可以看出在期望值与实际值（1000）差不多的条件下，模拟次数( $n$ )越大，VAR及MSE不一定越小，因此**模拟次数(n)越大估计效果不一定越好**。



模拟次数	样本容量	T	Expectation	Variation	MSE
100	10000	1500	1002	123.3	125.8
1000	10000	1500	1000	164.2	164.3
10000	10000	1500	1000	155.1	155.1

### iii 比較样本容量(m)

以下我们则实验在固定T=1500的条件下，模拟次数(n)=10000不变，分别取样本容量(m)=100,1000,10000。从下表可以看出在期望值与实际值(1000)差不多的条件下，样本容量(m)越大，VAR及MSE越小，因此样本容量(m)越大估计效果越好。



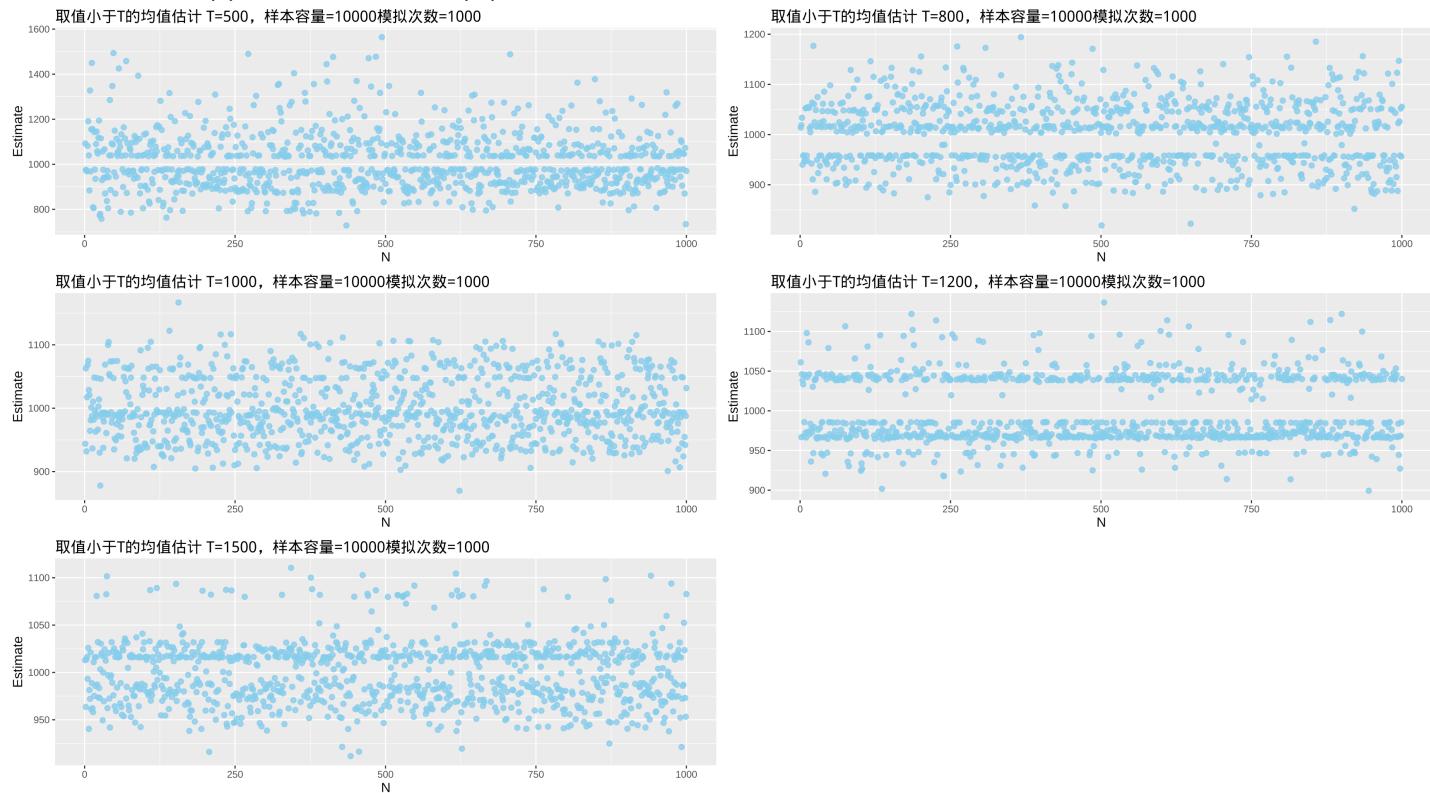
模拟次数	样本容量	T	Expectation	Variation	MSE
10000	100	1500	1004	15633.4	15646.6
10000	1000	1500	1000	1521.1	1521.2
10000	10000	1500	1000	155.1	155.2

综上所述，T、样本容量(m)都是越大越好，模拟次数(n)则不一定。

## 2.2 利用样本中所有取值小于T的部分的均值

i 比較T

下图为模拟次数( $n$ )=1000、样本容量( $m$ )=10000时， $T$ 分别为500,800,1000,1200,1500的图。

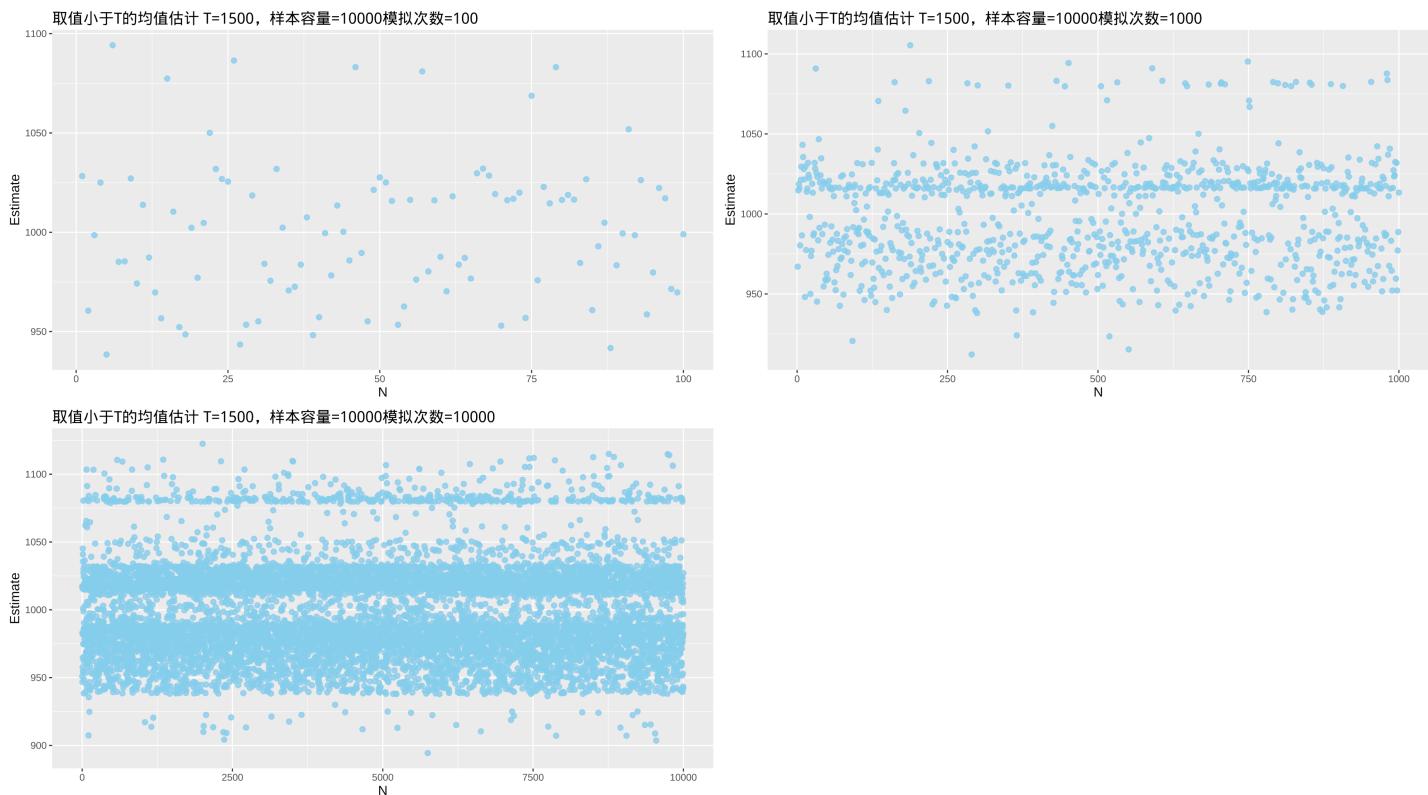


可以结合问题三，从无偏性、有效性和均方误差，以 $T$ 不同的值來比較優劣，下表為無偏性、有效性和均方誤差的結果。在期望值與實際值（1000）差不多的條件下，可以看出 $T$ 越大，VAR及MSE越小，因此**T**越大估計效果越好。

模拟次数	样本容量	T	Expectation	Variation	MSE
1000	10000	500	1016.7	14896.6	15175.8
1000	10000	800	1002.9	4036.8	4045.1
1000	10000	1000	999.3	2329.9	2330.3
1000	10000	1200	997.1	1816.4	1824.7
1000	10000	1500	999.8	990.5	990.6

## ii 比較模擬次數( $n$ )

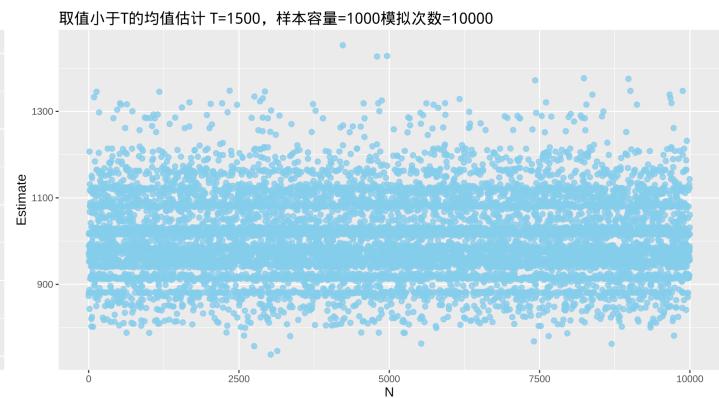
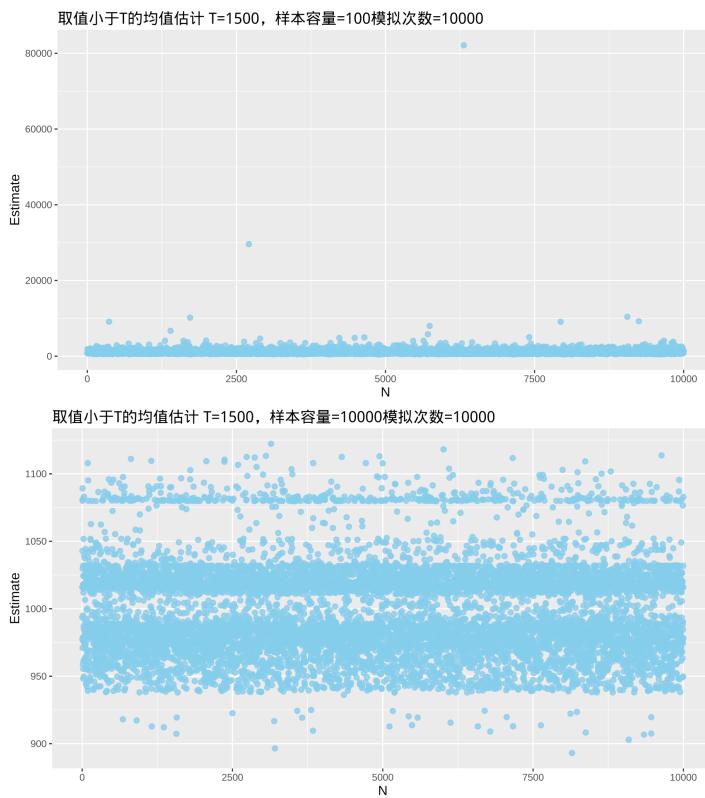
以下我們則實驗在固定 $T=1500$ 的條件下，樣本容量( $m$ )=10000不變，分別取模擬次數( $n$ )=100,1000,10000。從下表可以看出在期望值與實際值（1000）差不多的條件下，模擬次數( $n$ )越大，VAR及MSE不會越小，因此模擬次數( $n$ )越大估計效果不一定越好。



模拟次数	样本容量	T	Expectation	Variation	MSE
100	10000	1500	1002.9	1058	1067
1000	10000	1500	999.5	1091	1092
10000	10000	1500	1000.1	1113	1113

### iii 比較样本容量(m)

以下我们则实验在固定T=1500的条件下，模拟次数(n)=10000不变，分别取样本容量(m)=100,1000,10000。从下表可以看出在期望值与实际值(1000)差不多的条件下，样本容量(m)越大，VAR及MSE越小，因此样本容量(m)越大估计效果越好。



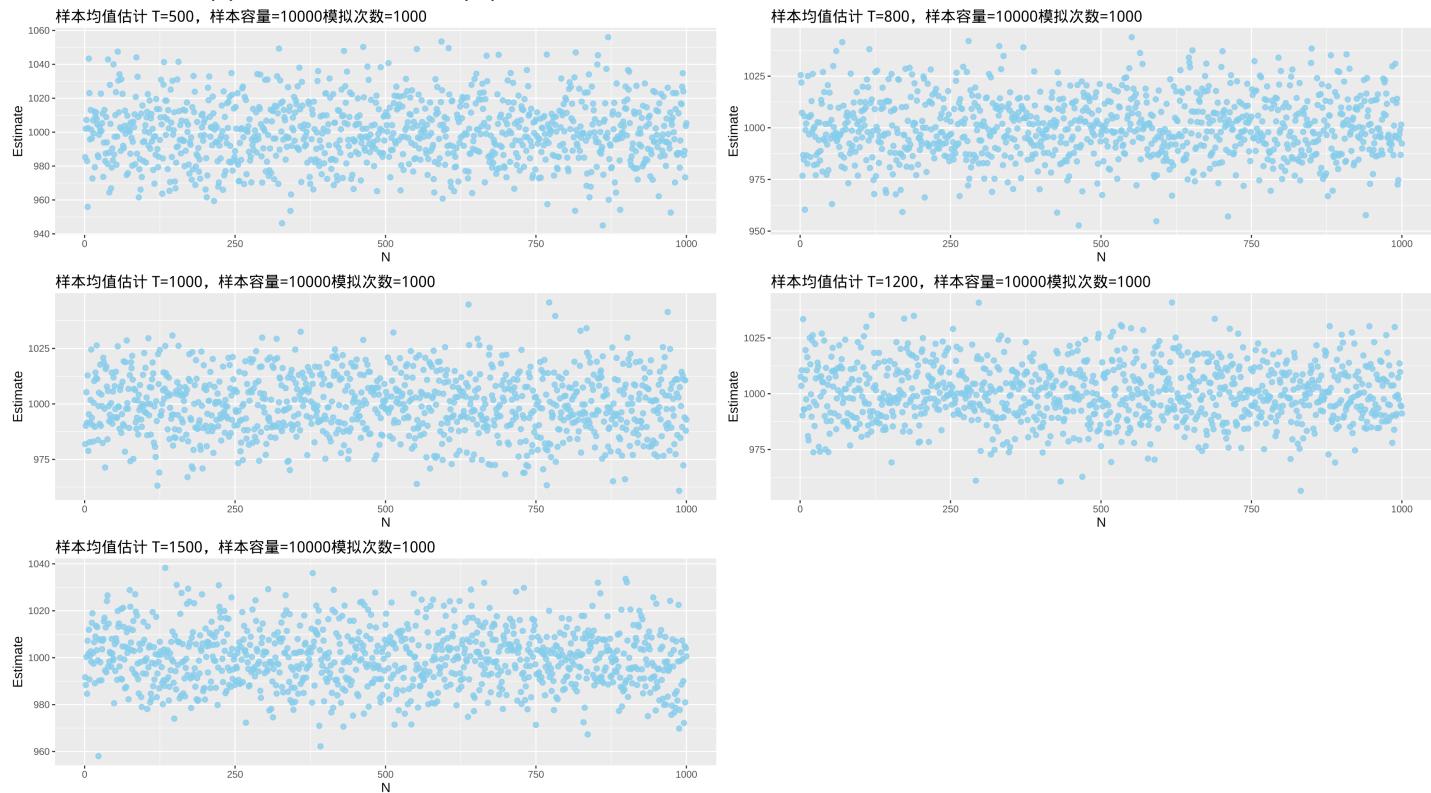
模拟次数	样本容量	T	Expectation	Variation	MSE
10000	100	1500	1090.3	705723	713883
10000	1000	1500	1007.8	8609	8669
10000	10000	1500	999.8	1076	1076

综上所述，T、样本容量(m)都是越大越好，模拟次数(n)则不一定。

## 2.3 利用样本均值

### i 比較T

下图为模拟次数( $n$ )=1000、样本容量( $m$ )=10000时， $T$ 分别为500,800,1000,1200,1500的图。

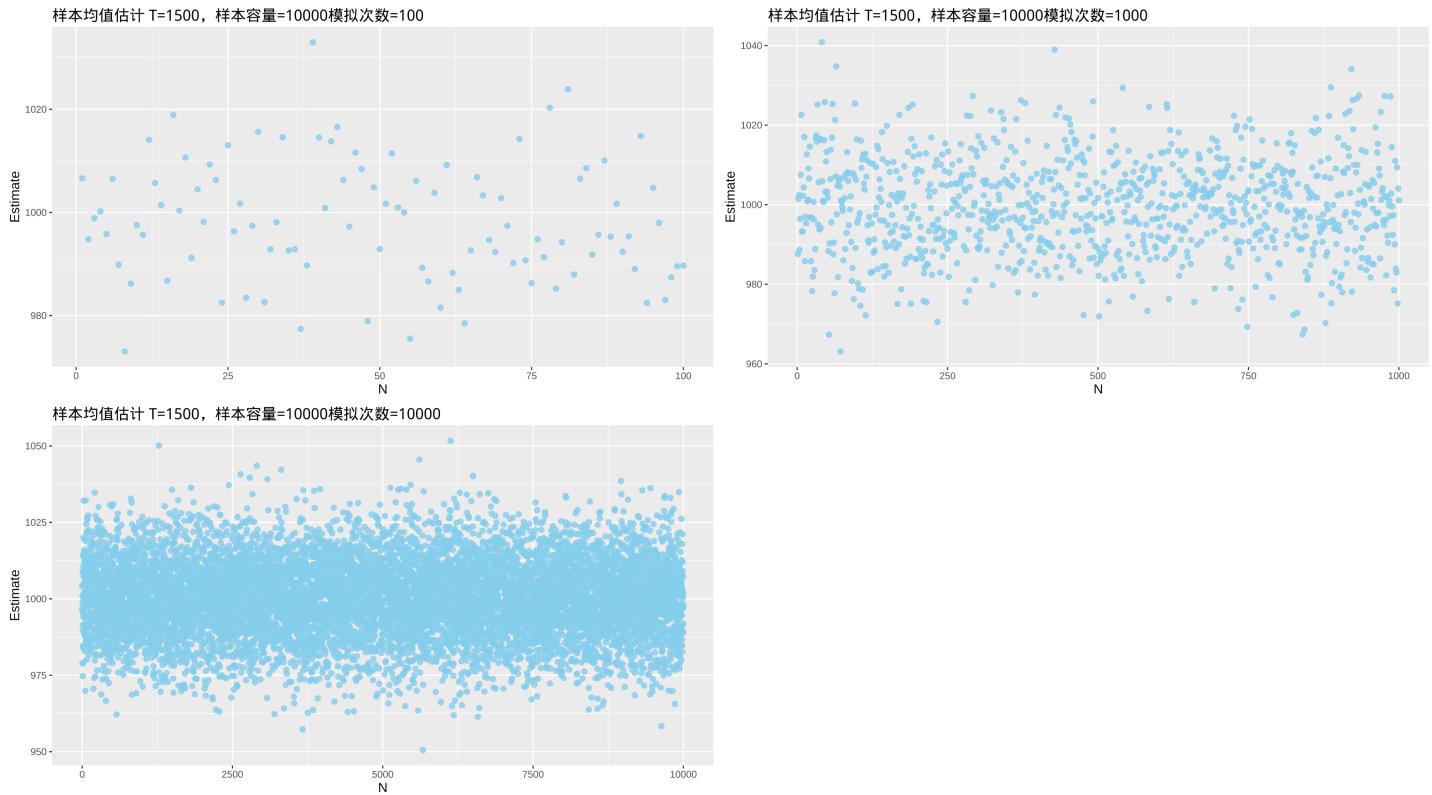


可以结合问题三，从无偏性、有效性和均方误差，以 $T$ 不同的值來比較優劣，下表為無偏性、有效性和均方误差的結果。在期望值与实际值（1000）差不多的条件下，可以看出 $T$ 越大，VAR及MSE越小，因此**T越大估计效果越好**。

模拟次数	样本容量	$T$	Expectation	Variation	MSE
1000	10000	500	999.5	311.8	312.1
1000	10000	800	1000.5	214.1	214.3
1000	10000	1000	1000.4	173.1	173.2
1000	10000	1200	1000.1	170.4	170.4
1000	10000	1500	1000.9	140.7	141.5

## ii 比較模拟次数( $n$ )

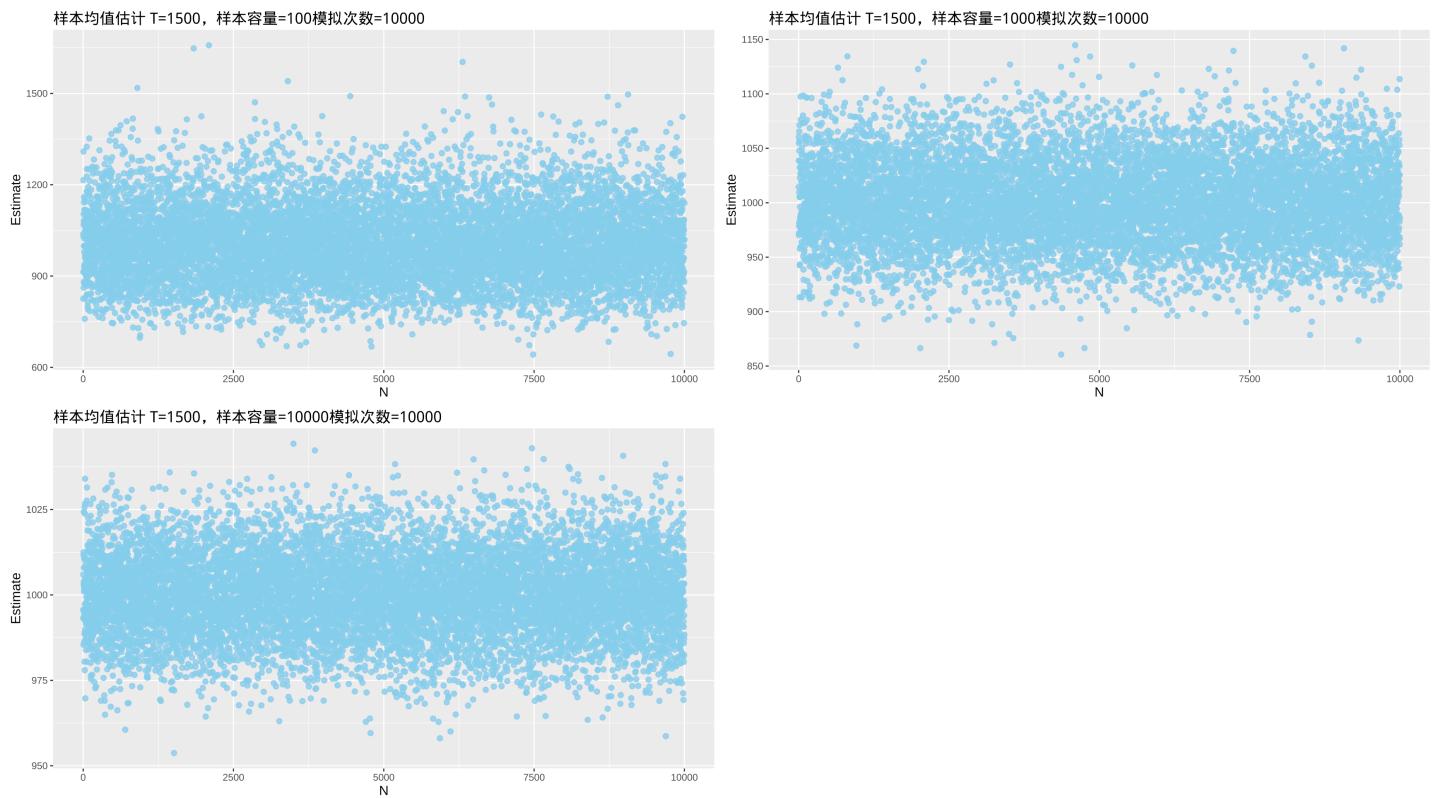
以下我們則實驗在固定 $T=1500$ 的條件下，樣本容量( $m$ )=10000不變，分別取模擬次數( $n$ )=100,1000,10000。從下表可以看出在期望值與實際值（1000）差不多的條件下，模擬次數( $n$ )越大，VAR及MSE不一定越小，因此**模擬次數( $n$ )越大估計效果不一定越好**。



模拟次数	样本容量	T	Expectation	Variation	MSE
100	10000	1500	1001.3	140.3	141.9
1000	10000	1500	999.7	139.8	139.9
10000	10000	1500	1000.0	147.2	147.2

### iii 比較样本容量(m)

以下我们则实验在固定T=1500的条件下，模拟次数(n)=10000不变，分别取样本容量(m)=100,1000,10000。从下表可以看出在期望值与实际值(1000)差不多的条件下，样本容量(m)越大，VAR及MSE越小，因此**样本容量(m)越大估计效果越好。**



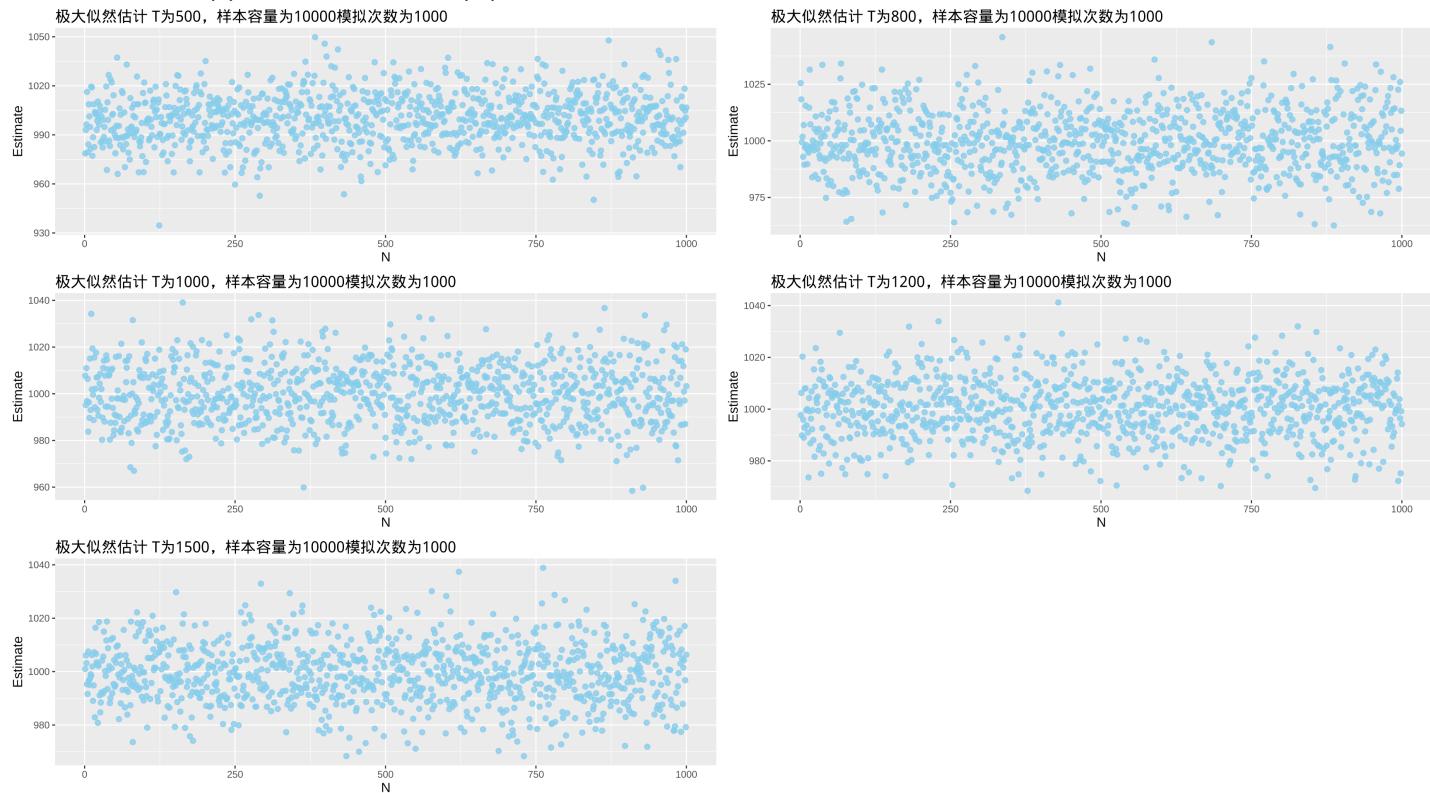
模拟次数	样本容量	T	Expectation	Variation	MSE
10000	100	1500	1010.1	14804.9	14907.6
10000	1000	1500	1001.5	1430.9	1433.1
10000	10000	1500	999.9	147.4	147.4

综上所述，T、样本容量(m)都是越大越好，模拟次数(n)则不一定。

## 2.4 利用极大似然估计

### i 比較T

下图为模拟次数( $n$ )=1000、样本容量( $m$ )=10000时， $T$ 分别为500,800,1000,1200,1500的图。

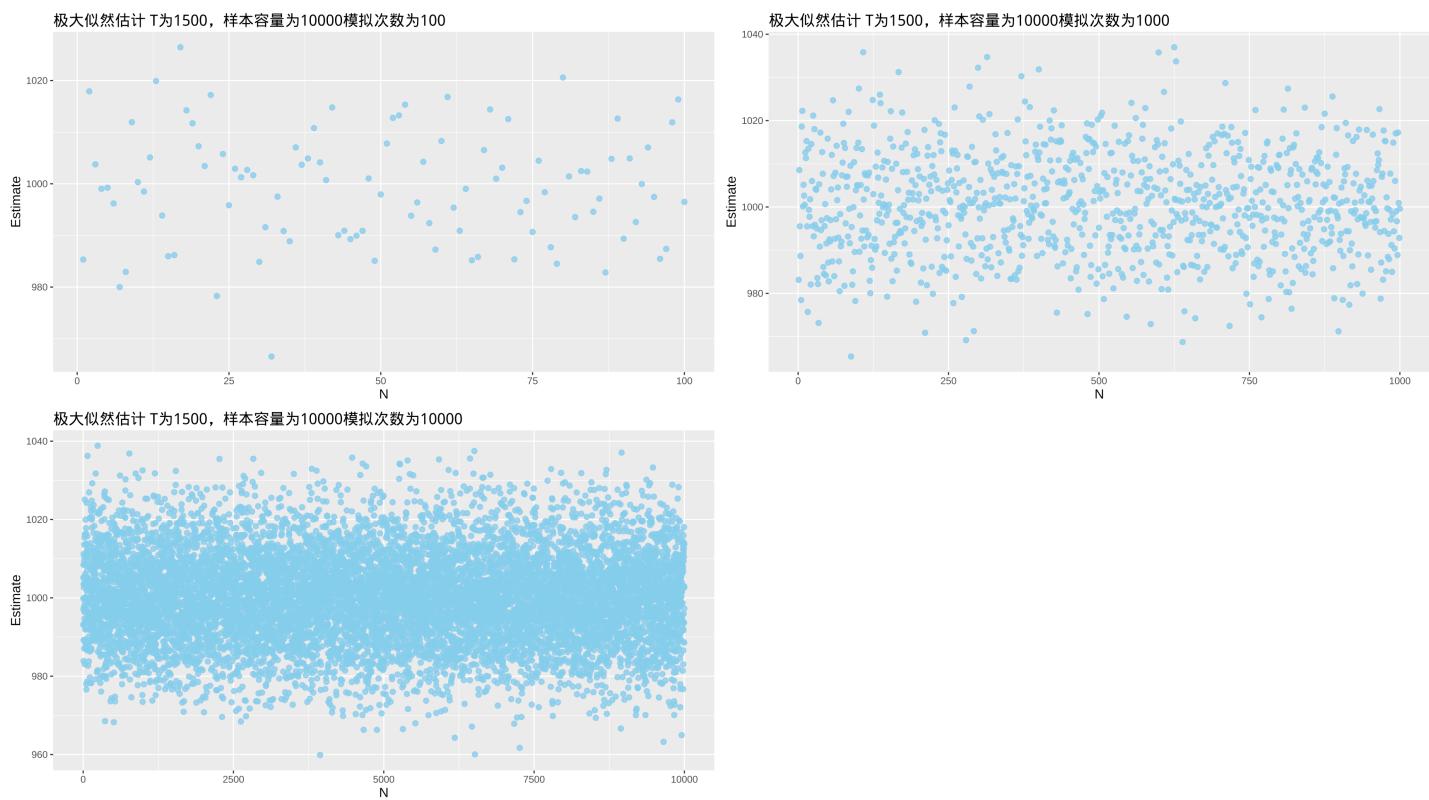


可以结合问题三，从无偏性、有效性和均方误差，以 $T$ 不同的值來比較优劣，下表为无偏性、有效性和均方误差的结果。在期望值与实际值（1000）差不多的条件下，可以看出 $T$ 越大，VAR及MSE越小，因此**T越大估计效果越好**。

模拟次数	样本容量	$T$	Expectation	Variation	MSE
1000	10000	500	999.6	260.2	260.4
1000	10000	800	1000.6	187.8	188.1
1000	10000	1000	999.9	152.5	152.5
1000	10000	1200	999.4	150.9	151.3
1000	10000	1500	999.7	127.4	127.5

## ii 比較模拟次数( $n$ )

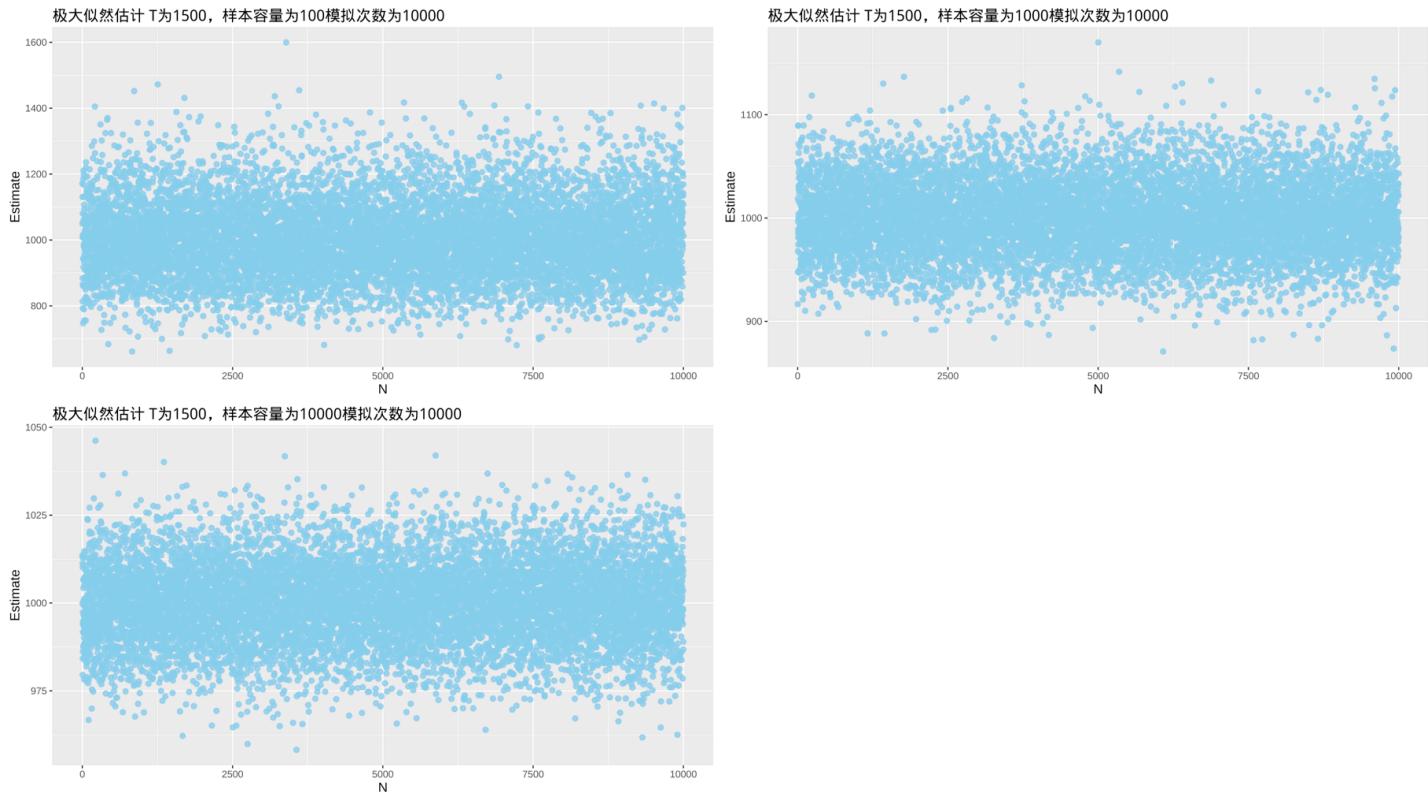
以下我们则实验在固定 $T=1500$ 的条件下，样本容量( $m$ )=10000不变，分别取模拟次数( $n$ )=100,1000,10000。从下表可以看出在期望值与实际值（1000）差不多的条件下，模拟次数( $n$ )越大，VAR及MSE不一定越小，因此**模拟次数(n)越大估计效果不一定越好**。



模拟次数	样本容量	T	Expectation	Variation	MSE
100	10000	1500	999.1	157.1	158.0
1000	10000	1500	999.7	123.4	123.5
10000	10000	1500	999.9	128.1	128.1

### iii 比較样本容量(m)

以下我们则实验在固定T=1500的条件下，模拟次数(n)=10000不变，分别取样本容量(m)=100,1000,10000。从下表可以看出在期望值与实际值 (1000) 差不多的条件下，样本容量(m)越大，VAR及MSE越小，因此样本容量(m)越大估计效果越好。



模拟次数	样本容量	T	Expectation	Variation	MSE
10000	100	1500	1005.5	13042.5	13073.2
10000	1000	1500	1000.6	1269.0	1269.3
10000	10000	1500	999.8	129.3	129.3

综上所述，T、样本容量(m)都是越大越好，模拟次数(n)则不一定。

### Question 3 问题1中四种估计方法中哪些能够进行无偏性、有效性和均方误差的理论分析，并给出相应的分析；用数值模拟的方法给出四种估计量的无偏性、有效性和均方误差的分析和判断。

固定模拟次数(n)=1000、样本容量(m)=10000、T=1000，分别将1.1命为方法一；1.2命危方法二；1.3命为方法三；1.4命为方法四。下表比较四种估计方法的评估比较。下表可以根据无偏性、有效性和均方误差来评估，Expectation越接近1000（真实值）越好；VAR越小越好；MSE越小越好，无偏性代表是否准确，有效性代表预测是否精确，而均方误差则可同时表示偏差及有效的特质，因为

$MSE = Var(\hat{\theta}) + Bias(\hat{\theta}, \theta)^2$ ;  $Bias(\hat{\theta}, \theta) = E(\hat{\theta} - \theta)$ 。根据下表和理论，方法二最不好而其他方法的差异没到很大。

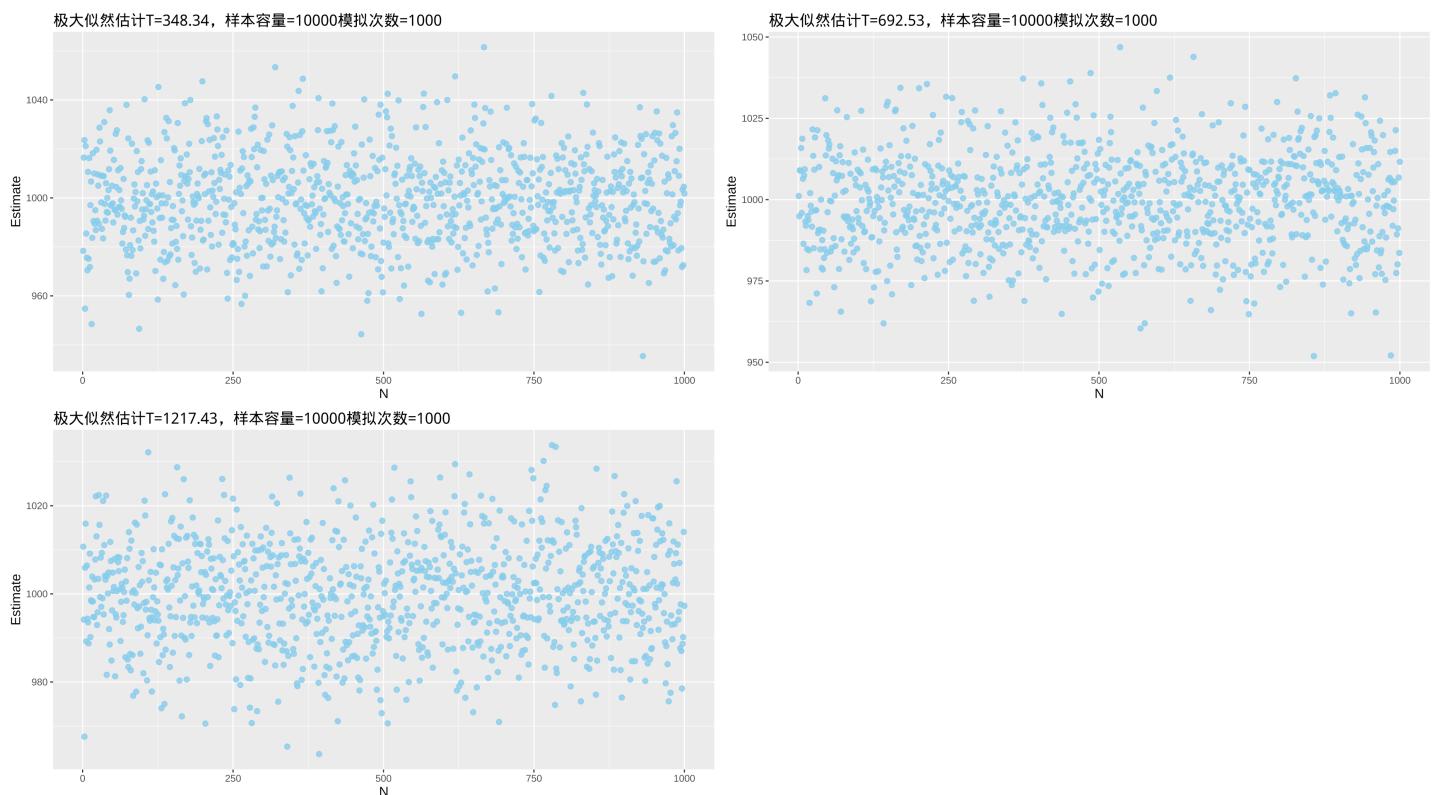
估计方法	Expectation	Variation	MSE
方法一	999.3	165.0	165.5
方法二	999.5	2644.7	2644.9

方法三	999.8	181.1	181.2
方法四	999.5	147.3	147.5

**Question 4** 仍然取  $\frac{1}{\lambda} = \frac{1}{1000}$  模拟生成样本值，因为实际抽样过程中可能并不知道寿命平均值的大致范围，所以难以给出合适的T的取值，可以分别考虑在30%、50%、70%（或其他比例）的样品失效时，停止试验。利用这样的抽样原则构造估计参数  $\frac{1}{\lambda}$  的方法，停止时间的规则如何确定比较合理，给出你的理由。

我们分别将取每个样本的百分之三十、百分之五十、百分之七十的样本分位数作为T的取值。在題三，我们得到在四种方法中，极大似然估计所得到的结果较好，因此用极大似然估计用于估计。

根据结果发现，当所取分位数比例从百分之三十增加到百分之五十时，所得结果的期望基本维持在1000不变，但是方差和均方误差均有较大幅度下降。分析是因为当T取值增加时，我们得到的样本信息更多，且多得到的样本信息对估计参数有效。因此，在不考虑成本及其它因素的情况下，T应取值越大越好。但是由于当T取值过大时，在实际操作时，会造成较大成本。在同等样本量的情况下，T应在成本允许的范围内越大越好。在样本量较大的情况下，由于样本量的增加会使得方差和均方误差变小，可以考虑适当缩减T的取值。



T	样本容量	Expectation	Variation	MSE
30%	10000	1000.3	332.7	332.9
50%	10000	1000.3	204.8	204.9
70%	10000	999.9	135.2	135.2

30%	1000	999.4	3520.1	3520.5
50%	1000	999.3	2034.1	2034.6
70%	1000	1002.1	1448.4	1453.0

## Question 5 是否能给出更多对参数 $\frac{1}{\lambda}$ 进行估计的办法，并进行模拟和分析。

使用截尾矩估计法来估计参数。

$x_1, x_2, \dots, x_n$  是相互独立且同分布的随机变量，其分布为  $G \frac{(x-\mu)}{\sigma}$ ，在  $(n, T)$  方案下得到的观测值是  $(y_1, \delta_1), \dots, (y_n, \delta_1)$ ，这里的  $\delta_i$  可以理解为被记录的标识变量，如被记录则为1，没有被记录则为0，其中， $y_i = \min(x_i, T)$ ,  $\delta_i = I_{(x_i \leq T)}$ ，在此提出一种矩估计方法：

令  $G^{-1}(u) = \inf(x : G(x) \geq u)$ ,  $(0 < u < 1)$ ，即反函数。

$$h(p) = pG^{-1}(u) - \int_0^p G^{-1}(u)du$$

$$\delta = I_{(x \leq T)}$$

$$a_G = \inf(x : G(x) > 0), b_G = \sup(x : G(x) < 1)$$

$$p = P(x \leq T)$$

定理：设  $0 < p < 1$ ，且  $G(x)$  是  $a_G, b_G$  上的增函数， $Y = \min(x, T)$ ，则有  $E[\delta(T - Y)] = \sigma h(p)$

$$\text{则 } \hat{\sigma} = \frac{1}{h(p)} E[\delta(T - Y)]$$

$$\text{用 } \hat{p} = \frac{1}{n} \sum_{i=1}^n \delta_i (T - y_i)$$

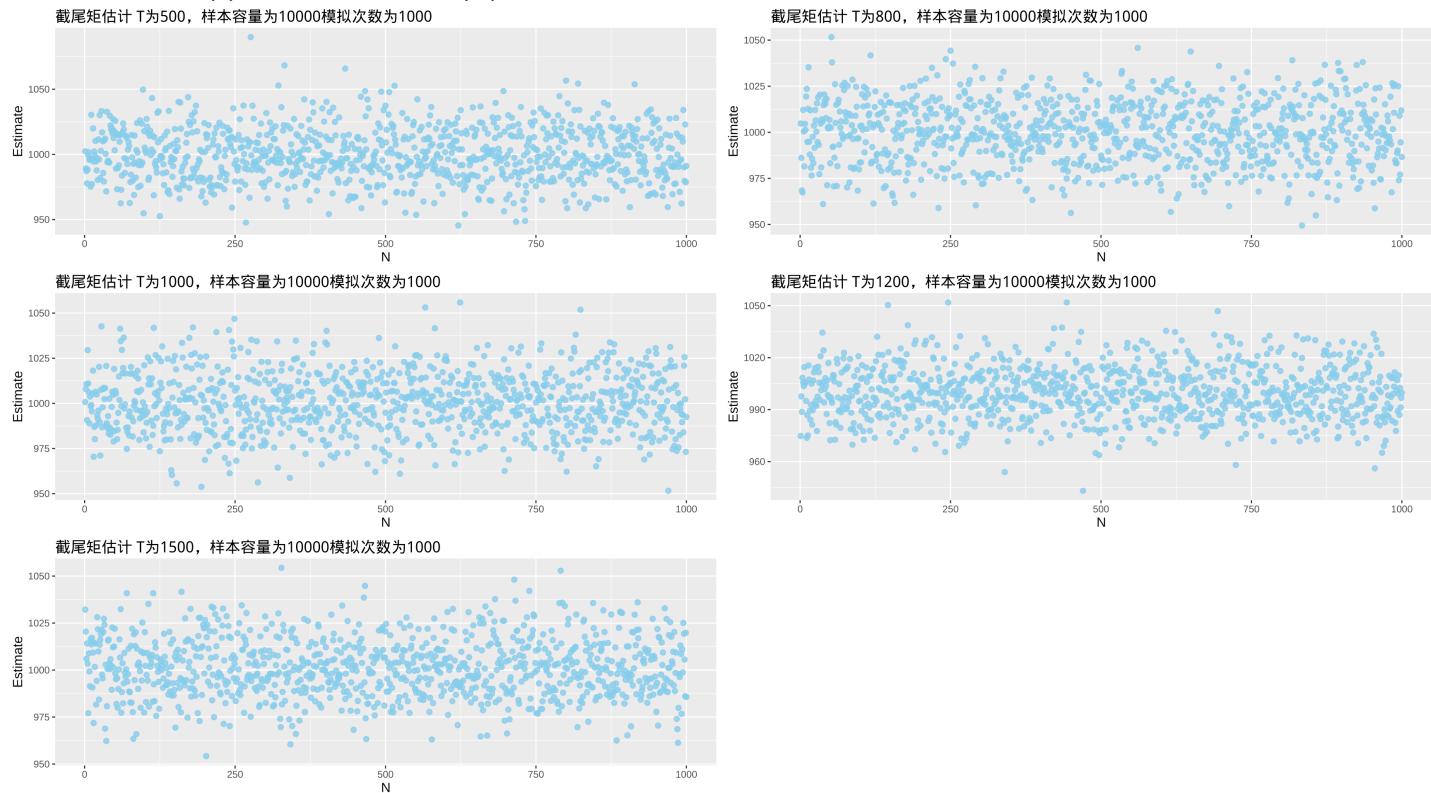
$$\text{又 } \mu = T - \sigma G^{-1}(p) \text{, 用 } \hat{\mu} = T - \hat{\sigma} G^{-1}(\hat{p}) \text{。}$$

在此题中，m 代表截止 T 失效的产品个数，n 为样本量， $x_i$  服从  $\text{Exp}(=1000)$ ，因此

$$\frac{1}{\hat{\lambda}} = \frac{1}{n(\ln \frac{1}{1 - \frac{m}{n}} - \frac{m}{n})} [mT - \sum_{i=1}^m x_{(i)}] \text{。}$$

i 比较 T

下图为模拟次数( $n$ )=1000、样本容量( $m$ )=10000时， $T$ 分别为500,800,1000,1200,1500的图。

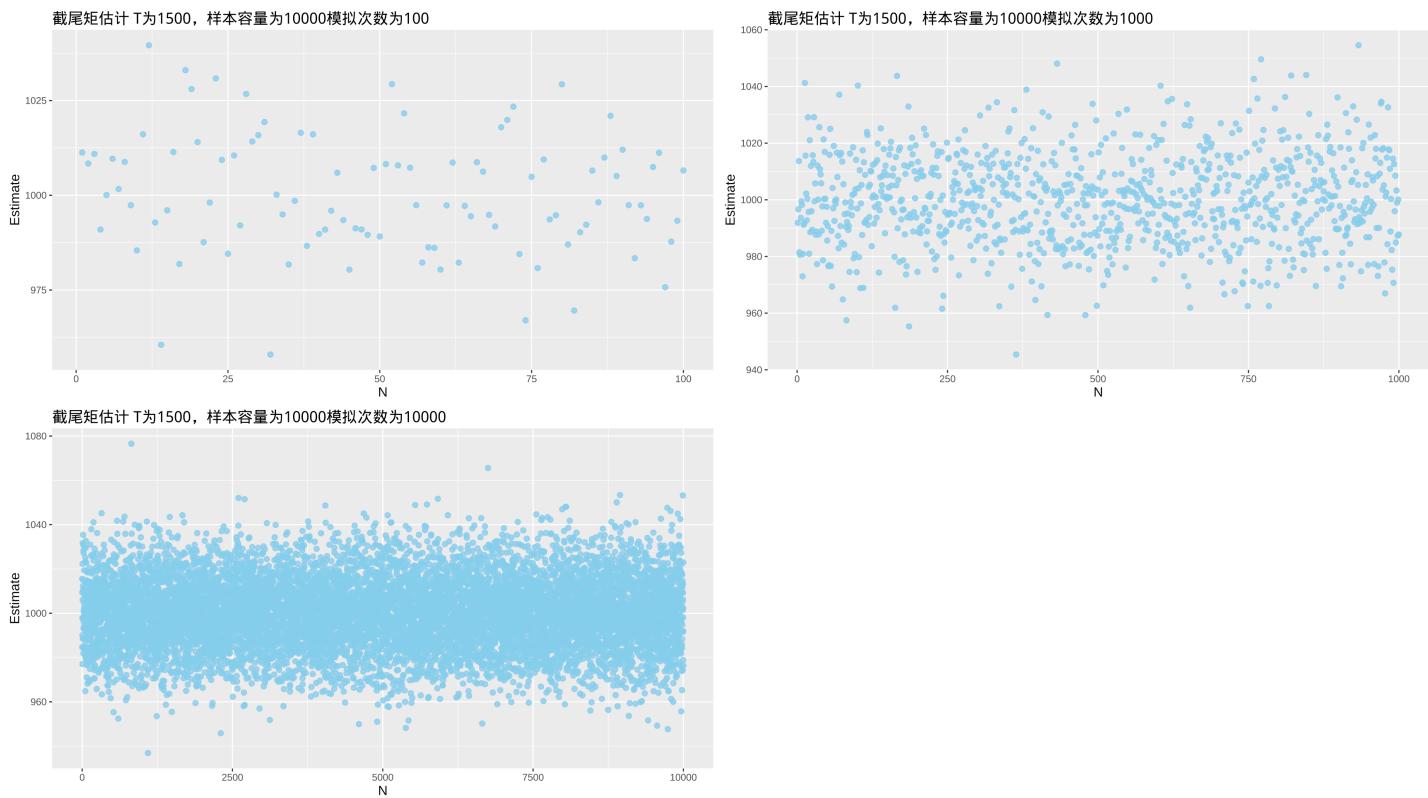


可以结合问题三，从无偏性、有效性和均方误差，以 $T$ 不同的值來比較优劣，下表为无偏性、有效性和均方误差的结果。在期望值与实际值（1000）差不多的条件下，可以看出 $T$ 越大，VAR及MSE不一定越小，因此 $T$ 越大估计效果不一定越好。

模拟次数	样本容量	$T$	Expectation	Variation	MSE
1000	10000	500	1000.2	388.1	388.1
1000	10000	800	999.5	269.4	269.7
1000	10000	1000	1000.3	270.8	270.9
1000	10000	1200	1000.2	247.2	247.3
1000	10000	1500	999.4	223.9	224.3

## ii 比較模拟次数( $n$ )

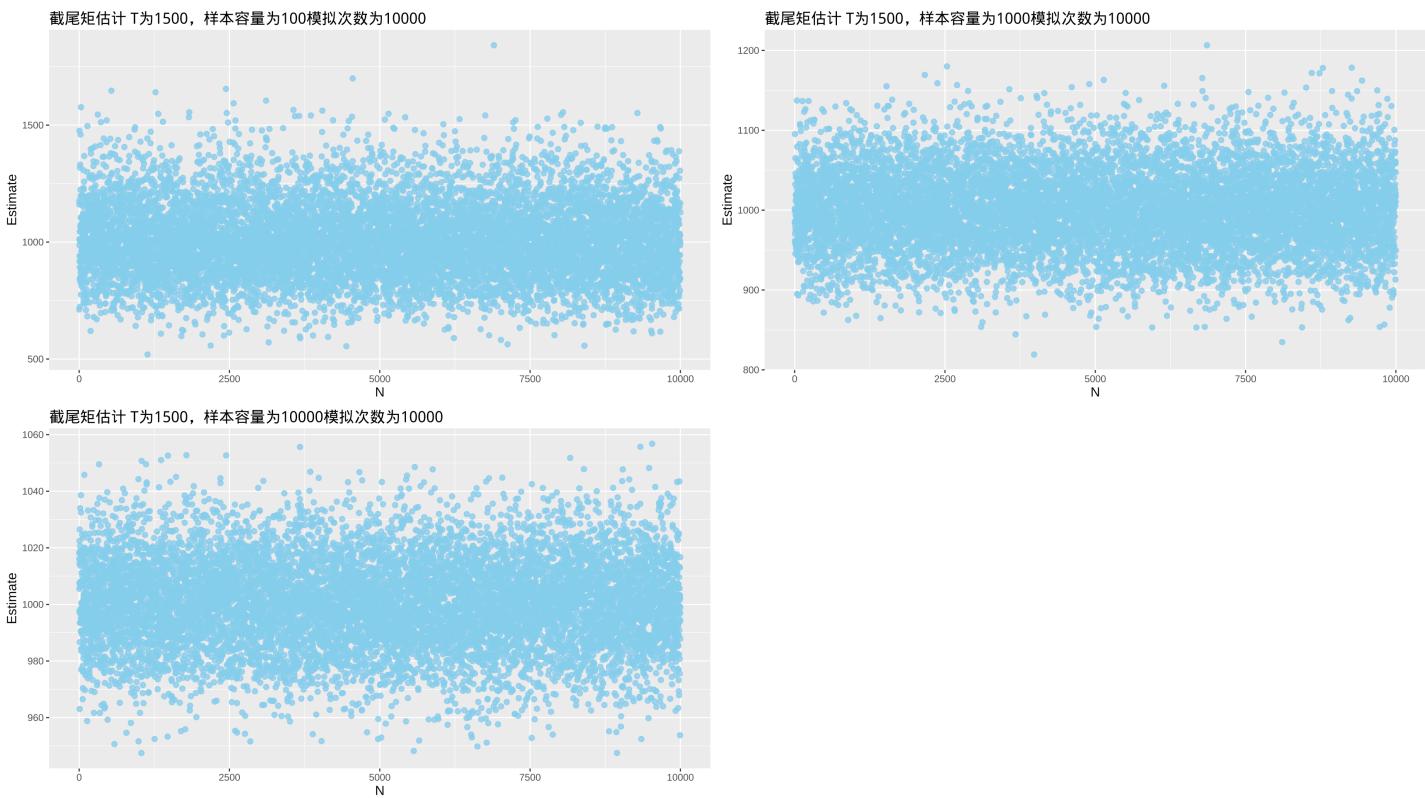
以下我们则实验在固定 $T=1500$ 的条件下，样本容量( $m$ )=10000不变，分别取模拟次数( $n$ )=100,1000,10000。从下表可以看出在期望值与实际值（1000）差不多的条件下，模拟次数( $n$ )越大，VAR及MSE不会越小，因此模拟次数( $n$ )越大估计效果不一定越好。



模拟次数	样本容量	T	Expectation	Variation	MSE
100	10000	1500	998.3	213.8	216.6
1000	10000	1500	999.9	241.9	241.9
10000	10000	1500	1000.1	239.7	239.7

### iii 比較样本容量(m)

以下我们则实验在固定T=1500的条件下，模拟次数(n)=10000不变，分别取样本容量(m)=100,1000,10000。从下表可以看出在期望值与实际值(1000)差不多的条件下，样本容量(m)越大，VAR及MSE越小，因此样本容量(m)越大估计效果越好。



模拟次数	样本容量	T	Expectation	Variation	MSE
10000	100	1500	1005.7	24789.9	24822.8
10000	1000	1500	999.9	2394.8	2394.8
10000	10000	1500	1000.3	241.5	241.7

综上所述，T、样本容量(m)都是越大越好，模拟次数(n)则不一定。