The 3rd Practical Report

Candidate Number: 1023122

Word Count:2166

Number of Pages:12

1.1 Abstract

In this report, our objective is to compare the samples which simulate the same distribution f(x) from three different MCMC methods using non-parametric tests. In particular, the report is consisted of four parts which correspond to the answers to four specific questions. At the beginning, we perform an exploratory analysis to get a brief picture of samples X, Y. As these are not enough to determine the relationship between distributions of two samples, non-parametric tests are introduced to check the hypothesis. In particular, according to different conditions, three tests including signed rank test, Wilcoxon rank sum test and Monte-Carlo permutation tests are adopted and give similar results. We further investigate the effects of parameters K and T, and it is found that null hypothesis is accepted when T passes a threshold and power of test is greater if K increases and T decreases respectively. For the last part, we use the Wilcoxon rank sum test to compare X and Y with the exact samples Z to test the validity of two samplers. A brief summary of conclusions and reflections about this report is also concluded.

1.2 Introduction

The data collects K samples of three different samplers respectively. Among these three samplers, two data vectors X = ($X1$, ..., XK) and Y = ($Y1$, ..., YK) are generated using functions m1 (T,x0) and m2 (T,x0) which converge in distribution to f as T increases and Z = (Z1, ..., ZK) is generated by inversion method. As X, Y both simulate a Markov chain of T steps with starting state x0 and return the Tth state, X[k] and Y[k] are correlated but conditionally independent given x0[k]. To check whether these two samplers give the same distribution, we initially perform an exploratory data analysis.

1.3 Data

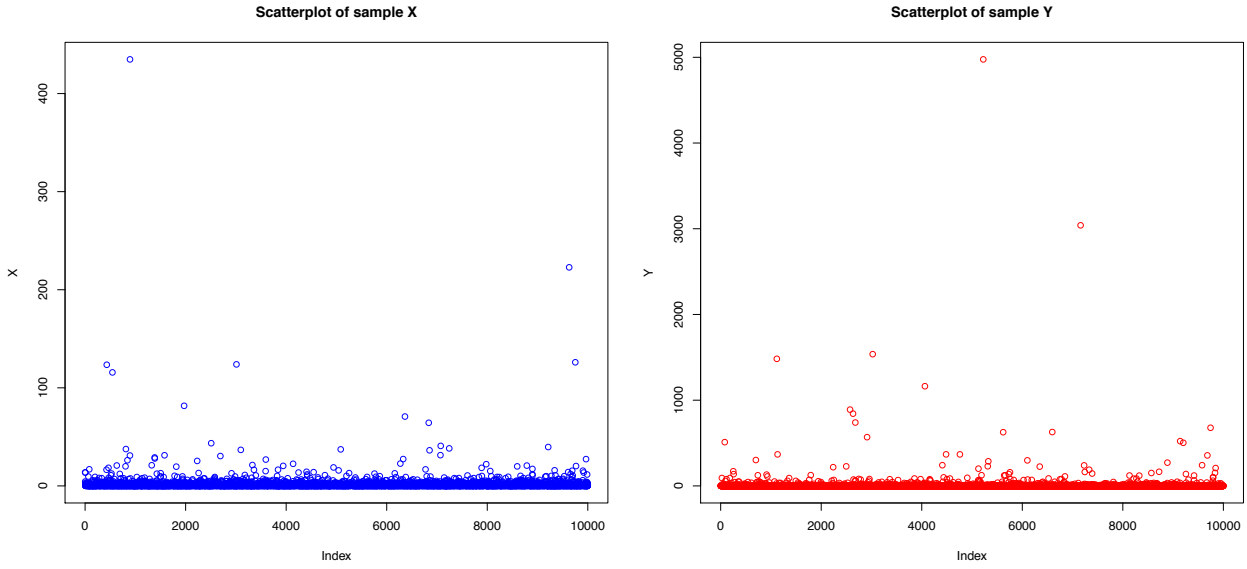

FIGURE 1. Scatterplots of samples X and Y

Figure 1 illustrates that 10000 data points of X and Y are collected respectively. Over 99% of data points in X and Y are below 500 and over 75% of data points gather in the range (0,3). Generally, X and Y show similar distribution properties that they both include large numbers of small-valued data points, while Y have a greater number of outliers and much higher maximum.

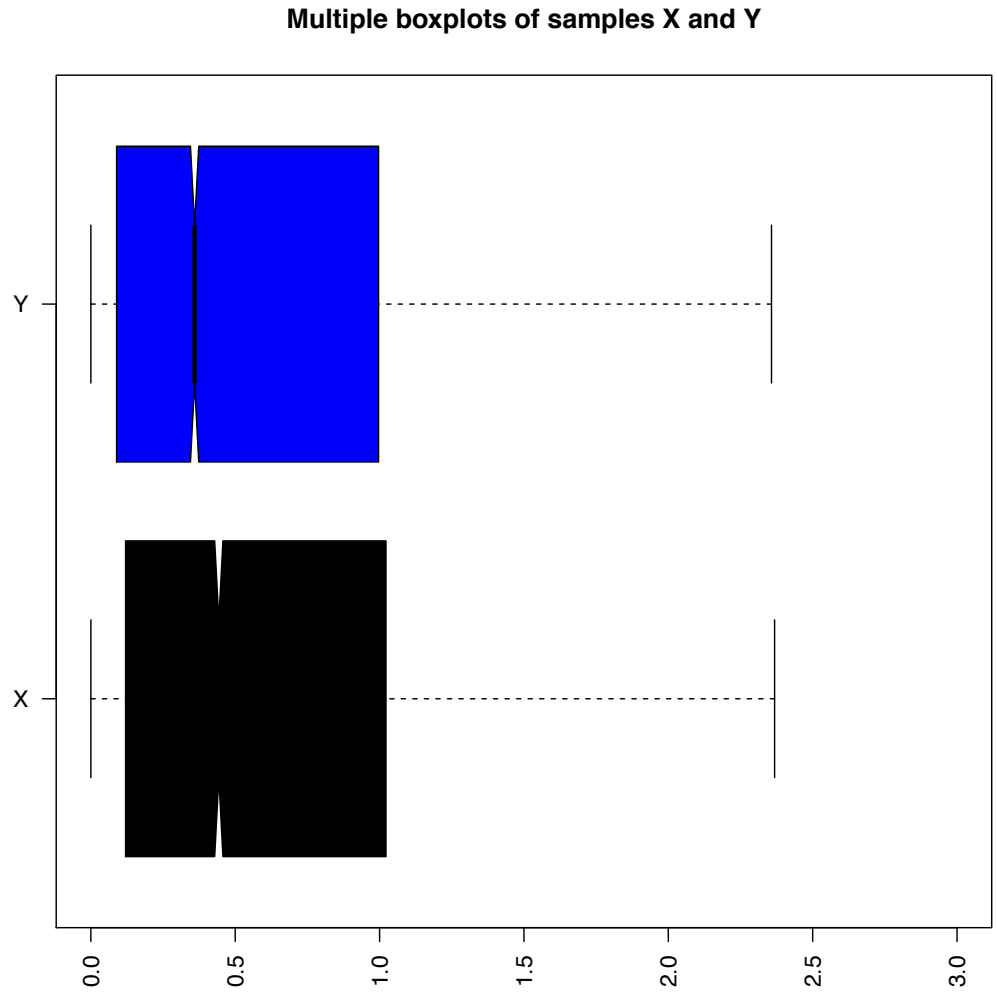

FIGURE 2. Boxplots of samples X and Y

```
> summary(X)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
    0.0000     0.1160     0.4298     1.1307     1.0039 434.8412
> summary(Y)
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
    0.0000     0.0887     0.3707     3.9558     1.0077 1536.8724
```

TABLE 1. Summary of samples X and Y

Boxplots in Figure 2 and Table 1 give a clearer picture of how data points are

distributed: X has the same minimum 0 as Y, much higher 1st quartile, higher median, approximately the same 3rd quartile and lower maximum compared to those of Y. And the lower mean of X is explained by the large number of outliers and higher maximum of Y.

2.1Tests

Q1: Check for X~Y with T=10, K=10000.

As we cannot determine whether X and Y have the same distribution purely from plots, we use two-sided non-parametric tests to test the null hypothesis H0: X~Y. Here as x0 is not fixed, thus X[k] and Y[k] are paired correlated observations. In addition, T=10 is small and correlation is strong. Therefore, signed rank test is a suitable method here.

Signed rank test

- Assumptions

Two samples are dependent observations with paired observations independent of one another. Given X~Y-Δ, X and Y are exchangeable, thus X-Y is symmetric about 0. Two measurements are of ordinal scale and are thus can be compared.

- Calculations

The null hypothesis is $H_0: \Delta = 0$ $(X \sim Y)$ against general alternative $H_1: \Delta \neq 0$.
Define

$$R_i = \sum_{j=1}^{n} I\{|X_j - \mu_0| \leq |X_i - \mu_0|\}$$

$$I_i = I\{X_i > \mu_0\}$$

where μ_0 is the median.
Then the SRT statistic is

$$W = \sum_{j=1}^{n} I_j R_j.$$

Under the null, $W_{obs} = 25720305$. Instead of getting the exact p value, we approximate it using Monte-Carlo method to generate a row vector W_{rand} of length n representing the simulated numbers of W.

$$p = 2\min\left(\frac{|\{i: W_{rand}(i) \geq W_{obs}\}|}{n}, \frac{|\{i: W_{rand}(i) < W_{obs}\}|}{n}\right)$$

This formula gives an approximate p value of a two-sided test p=0.014, thus we reject the null hypothesis that X and Y have the same distribution.

- Advantages & Disadvantages

The advantage of signed rank test is that it not only assigns a sign to the observations but it also shows the magnitude that data points are above or below

the median. However, this test is used specifically to compare related paired samples, it might not be useful when X and Y data vector have different lengths. ($m \neq n$)

- Conclusions

In the process of calculating p value, several K values are used for consideration. It is found that if K increases, p value drops significantly. For example, when K=1000, p=0.89; when K=5000, p=0.034; when K=10000, p=0.014. And one reason for this is that when sample size K increases, the differences between the true distribution of test statistic and the hypothesized distribution are more obvious. Therefore, if H0 is false, p value will drop greatly.

Q2: Check for X~Y with x0[k]=10 for all k, T=10, K=1000. (Here we choose K=1000 to save running time)

With a different initial state, we adopt different tests to check whether X~Y. As x0 is a fixed vector with all entries 10 and each sample in X and Y are simulated using x0, they are conditionally independent given x0. Then we could use powerful tests for location shift such as rank sum test and Monte-Carlo permutation test.

Wilcoxon rank sum test

- Assumptions

$X \sim f$, $Y - \Delta \sim f$, with Δ unknown. This is a strong assumption which implies that X[i] and Y[j] are independent identically distributed and are jointly independent for i=1,…n, j=1,…m and distributions are the same except for a location shift.

- Calculations

The null hypothesis is $H_0: \Delta = 0$ $(X \sim Y)$ against general alternative $H_1: \Delta \neq 0$. Define a vector which combines X and Y: $v = (X, Y)$, a variable $R_i = R(v)_k$ where $R()$ is the rank function and $k = 1, 2 \ldots, n + m$. Here $n = m = K = 1000$. Then we get our test statistic

$$W = \sum_{i=n+1}^{n+m} R_i.$$

To save computations, a normal approximation to distribution of W is adopted. Under the null hypothesis, $X \sim f$ and $Y \sim f$ with

$$E(W) = \frac{m(n+m+1)}{2} = 1000500$$

$$var(W) = \frac{mn(n+m+1)}{12} = 166750000$$

By central limit theorem,

$$\frac{W - E(W)}{\sqrt{var(W)}} \to Z \; as \; n, m \to infinity$$

Where Z has a standard normal distribution.

As $W_{obs} = 1097329 > E(W)$, we use a continuity correction to improve the accuracy of significance level:

$$p = P(W \geq W_{obs}) = \Phi\left(-\left|(W_{obs} - 0.5 - E(W))/\sqrt{var(W)}\right|\right) = 6.46e - 14$$

As the p<<0.05, we reject H0, then X and Y don't have the same distribution.

- Advantages & Disadvantages

The advantage of rank sum test is that it simulates the properties of a normal distribution when m and n are sufficiently large , and this makes analysis much easier. However, it costs many computations if we use the function in r directly instead of using a normal approximation of rank sum test.

Monte-Carlo permutation test

- Assumptions

$X[i]$ and $Y[j]$ are independent identically distributed and are jointly independent. X[i]~Fx and Y[j]~Fy for i = 1,...,n and j = 1,...,m with unknown Fx and Fy.

- Calculations

The null hypothesis is H0: Fx=Fy against general alternative H1: Fx≠Fy, without giving Fx, Fy.
The test statistic is still

$$W = \sum_{i=n+1}^{n+m} R_i.$$

Under H0, $R \sim Unif(P_{m+n})$ as they are all independent identically distributed samples and all permutations are equally likely. Then we approximate the p value as

$$p = 2P(W < -|W_{obs}|) = 2\min\left(P(W < W_{obs}), P(W > W_{obs})\right)$$

From R code, we get $W_{obs} = 1097329$ and $p = 0$. As shown in previous method that p is approximately $6.46e - 14$, and we have 1000 samples generated by Monte-Carlo method, thus that the approximate p value equals 0 is acceptable. Therefore, as p value is small, we reject the null hypothesis. Results are consistent with the previous methods.

- Advantages & Disadvantages

The advantage of permutation test is that it is generally applicable to any test statistic. In addition, it won't distort the distribution when there are ties, especially compared to rank sum test. However, it requires more time especially when the number of permutations is large.

Q3: For at least what value of T do we not reject X~Y at K=10000?

To solve this problem, we expect T to be large as X and Y are distributed according to f when T is large. We continue with the Wilcoxon rank sum test and use a for

loop in R to return the value of T that makes p value exceed 0.05. The result shows that at K=10000, T should be at least 25 to ensure a p value higher than 0.05 and we don't reject X~Y.

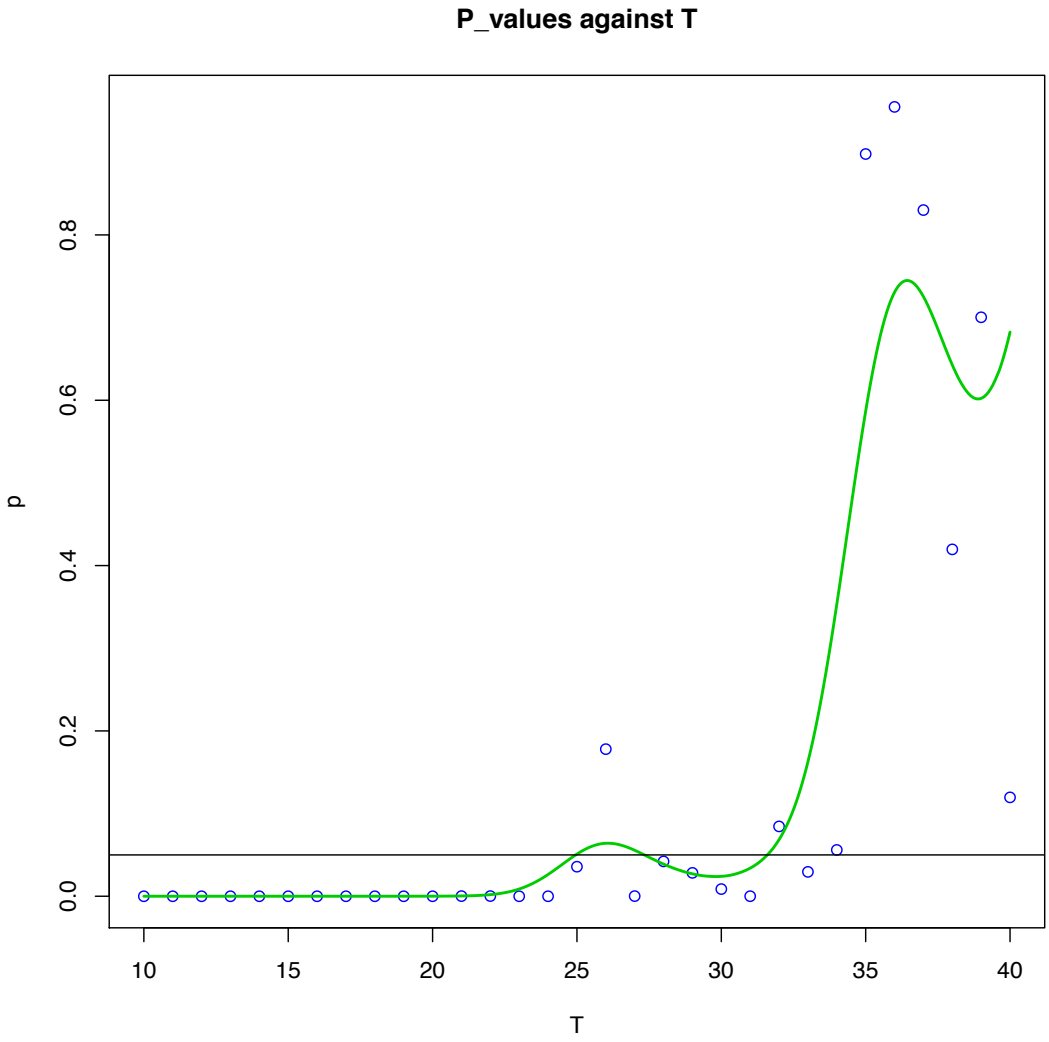

FIGURE 3. Plot of P values against T

It is shown in figure3 that p values are lower than 0.05 for T smaller than 25. And when T is larger than 32, p values increase significantly. The green line is a smoothing curve of datapoints using locpoly() function in R. It is shown that p values follow a general trend of increasing as T increases even though the smoother generates some fluctuations of the curve.

The power of a test is the probability of rejecting H0 when H0 is false. In general, the power of this test is greater when T decreases and when K increases. There are two major factors which affect the power of the test:

- Sample size: As the sample size K increases, we gather more information regarding the true distribution of the test statistic. Hence it is easier to detect whether the observed statistic behaves like the hypothesized distribution or

not. Therefore, the probability of correctly rejecting H0 is higher with greater K.

- Significance level: It is shown that T needs to increase from 10 (problem 1) to 25 (problem 3) in order to make the null hypothesis accepted. This is represented by a higher p value with a higher value of T. And if p value is higher, which means that the probability of accepting H0 is higher, thus a lower probability of rejecting H0. Then there is a smaller critical region which corresponds to a lower power of test.

Therefore, the power of test depends on T and K, with a proportional relationship with K and an inverse proportional relationship with T.

Q4. Do the Z samples offer any further opportunity for testing?

As the Z samples are simulated exactly using inversion method, we could perform nonparametric tests to check whether samples X and Y generally follow the distribution with pdf f(x). Here, we compare the samples of X and Y with Z with fixed T=100 and K=1000. As X,Y converge to f(x) when T and K both increase, we increase T to 100 while decrease K to ensure running time is not too long. We select the Wilcoxon rank sum test as Z doesn't depend on the parameters of X and Y and thus X, Z and Y, Z are clearly both jointly independent.

- Calculations

For H0: X~Z against the general alternative, the observed test statistic $W_{obs} = 997728$ with $E(W) = 1000500 > W_{obs}$ and $var(W) = 166750000$. This gives the p value using normal approximation $p = 0.8300591$. And this p value is much greater than 0.05, thus we strongly agree that X and Z have the same distribution.

For H0: Y~Z against the general alternative, the observed test statistic $W_{obs} = 1001479$ with $E(W) = 1000500 < W_{obs}$ and $var(W) = 166750000$. This gives the p value $p = 0.9395978$. Therefore, we strongly agree that Y and Z have the same distribution.

- Conclusions

It is shown that X and Y are good simulations for fixed T=100, K=1000. However, the resulting p value changes as T increases: when T=10, p values for the first and the second tests above are approximately 0.008956067 and 0.04753479; when T=1000, p values for the first and second tests above are approximately 0.7784788 and 0.9843994. Therefore , we expect the performances of these two samplers to be better as T increases.

3.1Conclusions

In brief, there are three groups of samples generated by three different methods. Two data vectors X,Y are generated by m1 (T,x0), m2(T,x0) and Z is generated by inversion. Initially, we compare the distributions of X and Y with different conditions: when x0 is not fixed with T small, null (X~Y) is rejected using signed

rank test; when x0 is fixed with jointly independent samples, null is also rejected by Wilcoxon rank sum test and Monte-Carlo permutation test. To further investigate how parameters K and N influence the testing results, it is found that at K=10000, T has to be approximately at least 25 to accept the null hypothesis. And the power of test is greater if K increases and T decreases respectively. Z samples are used to test whether X and Y samples follow f distribution. The result shows that although X and Y are both considered good simulators, they are not so accurate when T is small.

However, there are still certain weaknesses in this report: firstly, even though we check whether X~Z and Y~Z for Q4 and reach a conclusion that they both work well, we don' t include a comparison between X and Y using Z to see which is a better simulator. Moreover, we include only nonparametric tests which assume less specific information than parametric tests and thus are less powerful sometimes. Finally, the running time and computational machines limit the values of parameters used which might generate incomprehensive results.

practical3.R

```
## Question 1
# Generate the data with T=10 and K=10000
set.seed(1)
T=10;
K=10000;
x0=X=Y=Z=numeric(K);
for (k in 1:K) {
  x0[k]=abs(rt(1,df=1))
  X[k]=m1(T,x0=x0[k])
  Y[k]=m2(T,x0=x0[k])
}

# Perform an exploratory data analysis
summary(X)
summary(Y)
pdf("practical3_boxplot.pdf", height=8, width=8)
b<-boxplot(X,Y,ylim=c(0,3),main="Multiple boxplots of samples X and Y
",at=c(1,2),names=c("X","Y"),las=2,col=c("black","blue"),horizontal=
TRUE,notch=TRUE,outline=FALSE)
dev.off()

pdf("practical3_scatterplot.pdf", height=8, width=17)
par(mfrow=c(1,2))
plot(X,main="Scatterplot of sample X",xlab="Index",ylab="X",col="blu
e")
# plot(Y,main="Scatterplot of sample Y",xlab="Index",ylab="Y",col="r
ed")
dev.off()

# Signed rank test
n=length(X);
m=length(Y);
srt=data.frame(X,Y);
srt$diff=numeric(n);
for (i in 1:n){srt$diff[i]=X[i]-Y[i]}
# srt$diff is the vector of entries equaling the differences of entri
es of X and Y
srt$adiff=abs(srt$diff)
srt$rank=rank(srt$adiff)
srt$srank=srt$rank*sign(srt$diff)
W=sum(srt$srank[srt$srank>0])
get_pvalue_srt <- function(w_observed, n, nsim=1000){
```

```
  w_observed<-W
  nsim<-1000
  w_rand <- numeric(nsim)
  rnk<-1:n
  for (sim in 1:nsim){
    sign <- sample(c(-1,1), replace=TRUE, size=n)
    srank <- rnk*sign
    w_rand[sim] <- sum(srank[srank>0])
  }
  prob_larger <- mean(w_observed <= w_rand)
  prob_smaller <- mean(w_observed >= w_rand)
  p_value <- 2*min(prob_smaller, prob_larger)
  return(list(p_value = p_value, w = w_rand))
}
sim=get_pvalue_srt(W,n=10000);
sim$p_value

## Question 2
# Initialise with x0[k]=10 and regenerate data with t=10 and K=1000
set.seed(1)
T=10;
K=1000;
x0=X=Y=Z=numeric(K);
for (k in 1:K) {
  x0=rep(10,K)
  X[k]=m1(T,x0=x0[k])
  Y[k]=m2(T,x0=x0[k])
}

# Wilcoxon rank sum test (Normal Approximation)
n=length(X);
m=length(Y);
rst=c(X,Y);
rst.rank=rank(rst);
W=sum(rst.rank[1001:2000])
e=m*(n+m+1)/2
v=n*m*(n+m+1)/12
(W-e)/sqrt(v)
p=2*pnorm(-abs((W-0.5-e)/sqrt(v))) # Continuity Correction for the up
per tail

# Monte-Carlo permutation test (with ties)
get_pvalue <- function(w_observed, n, m, nsim = 10000){
  ranks <- 1:(n+m)
```

```
  w_rand <- numeric(nsim)
  for (sim in 1:nsim){
    w_rand[sim] <- sum(sample(ranks, m, replace=FALSE))
  }
  prob_larger <- mean(w_observed <= w_rand)
  prob_smaller <- mean(w_observed >= w_rand)
  p_value <- 2* min(prob_smaller, prob_larger)
  return(p_value)
}
get_pvalue(W,1000,1000)

## Question 3
# Find threshold of T with K=10000, x0[k]=0
K=10000;
p=numeric(25);
for(T in 10:40){
  x0=X=Y=Z=numeric(K)
  for (k in 1:K) {
    x0=rep(10,K)
    X[k]=m1(T,x0=x0[k])   #K independent approximate simulations X~f
using m1()
    Y[k]=m2(T,x0=x0[k])   #K independent approximate simulations X~f
using m2()
  }
test_res<-wilcox.test(X,Y,paired=FALSE)
p[T]=test_res$p.value
}

# Plot p values against T
library(KernSmooth)
pdf("practical3_pvalues.pdf", height=8, width=8)
plot(10:40,p[10:40],main="P_values against T",xlab="T",ylab="p",col=
"blue")
k <- locpoly(10:40,p[10:40],bandwidth=dpill(10:40,p[10:40]),degree=
1)
lines(k,col=3,lwd=2)
abline(h=0.05)
dev.off()

## Question 4
# Regenerate data with T=100, K=1000
set.seed(1)
T=100
K=1000
```

```
x0=X=Y=Z=numeric(K);
for (k in 1:K) {
  x0[k]=abs(rt(1,df=1))
  X[k]=m1(T,x0=x0[k])
  Y[k]=m2(T,x0=x0[k])
  Z[k]=Finv(runif(1))
}

# Wilcoxon rank sum test
n=length(X);
t=length(Y);
m=length(Z);
rst1=c(X,Z);
rst1.rank=rank(rst1);
W1=sum(rst1.rank[1001:2000])
e1=m*(n+m+1)/2
v1=n*m*(n+m+1)/12
(W1-e1)/sqrt(v1)
p=12*pnorm(-abs((W1+0.5-e1)/sqrt(v1)))

rst2=c(Y,Z);
rst2.rank=rank(rst2);
W2=sum(rst2.rank[1001:2000])
e2=m*(t+m+1)/2
v2=t*m*(t+m+1)/12
(W2-e2)/sqrt(v2)
p2=2*pnorm(-abs((W2-0.5-e2)/sqrt(v2)))
```