ZHANG Fan

EE511 Pro.4 USC ID:1417-68-5115

This is the updated version. Please ignore the another version I handed out. Thank you!!!

Q1:

Approximate the following integrals using a Monte Carlo simulation. Compare your estimates with the exact values (if known):

a. $\int_{-2}^{2} e^{x+x^2}\,dx.$

b. $\int_{-\infty}^{\infty} e^{-x^2}\,dx.$

c. $\int_0^1 \int_0^1 e^{-(x+y)^2}\,dy\,dx.$

Code:

```
%Q1_ab
clear
N=10000;
bound(1:2)=[-inf,inf]
bound(3)=0
syms x
%y=exp(x+x^2);
y=exp(-x^2)
handout=vpa(int(y,bound(1),bound(2)))
fx=inline(y);
bound(4)=max(fx(bound(1):0.01:bound(2)));
B = bound;
     R = rand(2, N);
     %Set the random samplings to the correct intervals
     R(1, :) = (B(2)-B(1))*R(1, :)+ B(1);
     R(2, :) = R(2, :)*(B(4) - B(3)) + B(3);
     area = (B(2)-B(1))*(B(4)-B(3));
     s = fx(R(1,:))>=R(2,:);
     total = sum(s);
     avgF = total/N;
     Approx = avgF*area
     plot(R(1,:),fx(R(1,:)),'*')
     hold on
     plot(R(1,:),R(2,:),'b*')
```

Results:

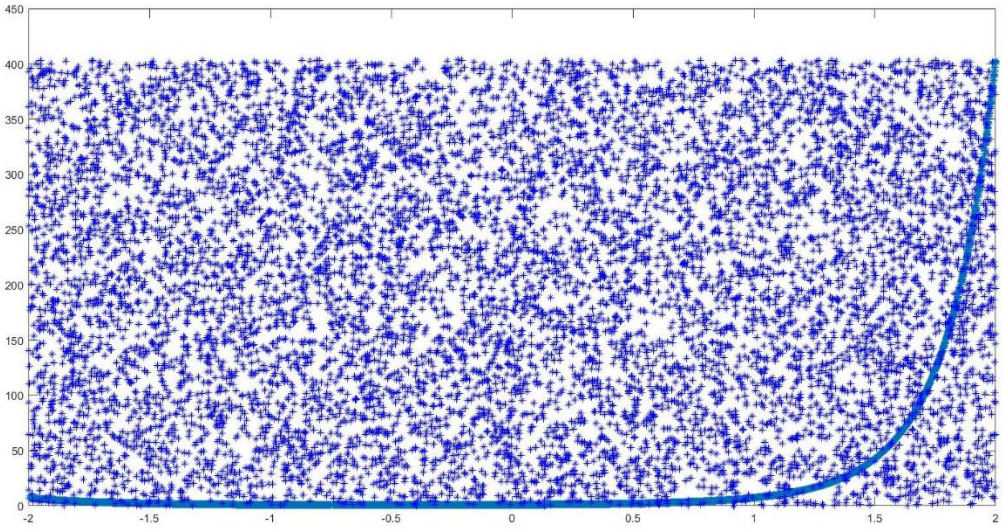
a.

```
handout =

93.162753292441975538173131907087


Approx =

   92.3045
```



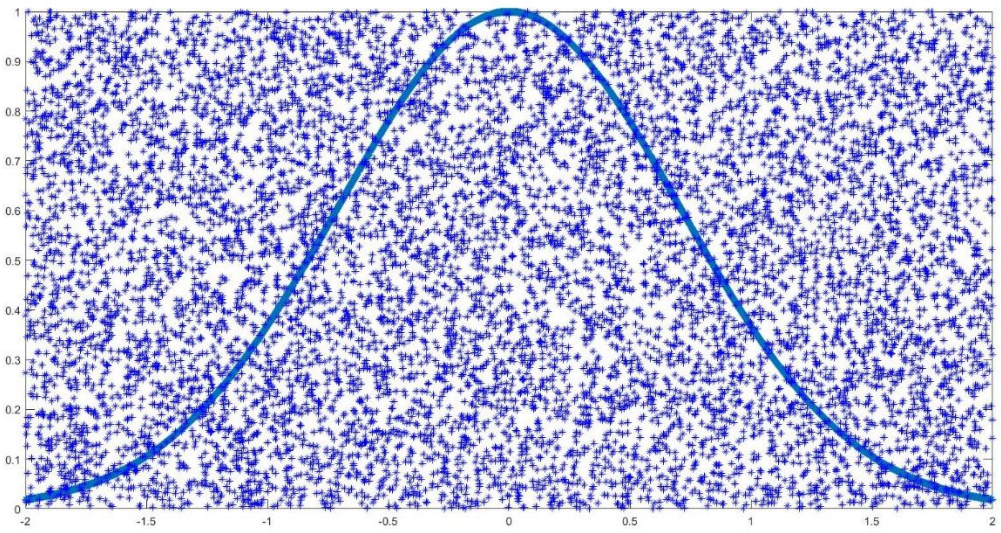
b.

```
handout =

1.764162781524843359934962071828 1


Approx =

    1.7600
```




```
 %code for Q1_c
clear
N=10000;
bound(1:4)=[0,1,0,1]
bound(5)=0
syms x y
%y=exp(x+x^2);
%y=exp(-x^2)
z=exp(-(x+y).^2);
%handout=vpa(int(y,bound(1),bound(2)))
%fx=inline(y);
fx=inline(z);
handout=vpa(int(int(z, x, 0, 1), y, 0, 1))
bound(6)=max(fx(bound(1):0.01:bound(2),bound(3):0.01:bound(4)));
B = bound;
     R = rand(3, N);
     %Set the random samplings to the correct intervals
     R(1, :) = (B(2)-B(1))*R(1, :)+ B(1);
     R(2, :) = R(2, :)*(B(4) - B(3)) + B(3);
     R(3, :) = R(3, :)*(B(6) - B(5)) + B(5);
     V = (B(2)-B(1))*(B(4)-B(3))*(B(6)-B(5));
     s = fx(R(1,:),R(2,:))>=R(3,:);
     total = sum(s);
     avgF = total/N;
     Approx = avgF*V
```

Results for c:

```
handout =

0.411792894172914077687362440039 56


Approx =

    0.4127
```

Q2

Define the random variable $X = Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2$ where $Z_k \sim N(0,1)$. Then $X \sim \chi^2(4)$. Generate 10 samples from X by first sampling Z_i for $i = 1, 2, 3, 4$ and then computing X. Plot the empirical distribution $F_{10}^*(x)$ for your samples and overlay the theoretical distribution $F(x)$. Estimate a lower bound for $\|F_{10}^*(x) - F(x)\|_\infty$ by computing the maximum difference at each of your samples: $\max_{x_i} |F_{10}^*(x_i) - F(x_i)|$. Then find the 25th, 50th, and 90th percentiles using your empirical distribution and compare the value to the theoretical percentile values for $\chi^2(4)$. Repeat the above using 100 and 1000 samples from X.

Code:

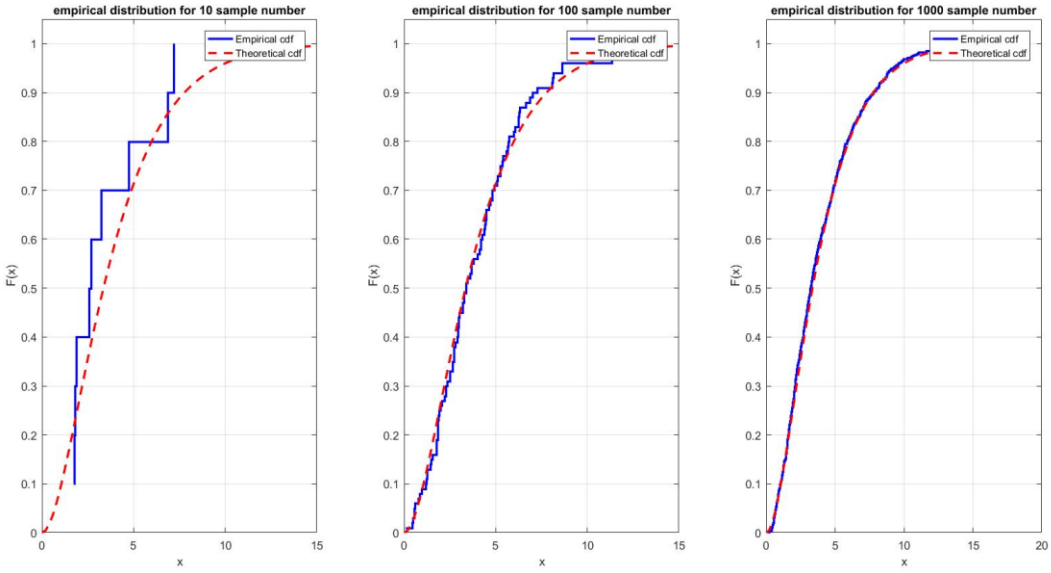
```
%Q2
clear
n=[10,100,1000];
for q=1:3
for i=1:n(q);
       for j=1:4;
       z(j)=randn;
       end
    X(i)=z(1).^2+z(2).^2+z(3).^2+z(4).^2;%randn: Normally distributed random
numbers
 end
X=sort(X);
subplot(1,3,q);

figure(1);
stairs(X,1/n(q):1/n(q):1,'b','linewidth',2);%X is the sample.
%The probablity of X1,X2,X3 may be 0.2, 0.5,0.3(pdf) or 0.2,0.7,1(cdf),
%in this case, it is cdf. The theory of cdf is:
%sort all of the samples,
%give each sample a number, 1, 2, 3, ...n in a accending order
%when the sequence(or the sample number) is standared, it become the cdf of
the samples scpace.
hold on
grid on
x=0:0.2:15;
y=chi2cdf(x,4);
plot(x,y,'r--','linewidth',2);
```

```
hold off
legend('Empirical cdf','Theoretical cdf');
ylim([0 1.05]);
xlabel('x');
ylabel('F(x)');
title(['empirical distribution for ',num2str(n(q)) , ' sample number']);

sprintf('Sample number: %d, Theoretical 25th, 50th, 90th
percentiles:%f,%f,%f',...
n(q),chi2inv(0.25,4),chi2inv(0.5,4),chi2inv(0.90,4))
sprintf('Smaple number: %d, Empirical 25th, 50th, 90th
percentiles:%f,%f,%f',...
n(q),X(ceil(0.25*n(q))),X(ceil(0.5*n(q))),X(ceil(0.9*n(q))))
sprintf('sample number: %d difference of 25th, 50th, 90th percentiles between
Empirical and theoretical confidence interval:\n%f,%f,%f',...
  n(q),abs(chi2inv(0.25,4)-X(ceil(0.25*n(q)))),abs(X(ceil(0.5*n(q)))-
chi2inv(0.5,4)),abs(X(ceil(0.9*n(q)))-chi2inv(0.90,4)))
end
```

Results:



When the sample is 10:

```
ans =

Sample number: 10, Theoretical 25th, 50th, 90th percentiles:1.922558,3.356694,7.779440


ans =

Smaple number: 10, Empirical 25th, 50th, 90th percentiles:1.186583,2.463753,7.278311


ans =

sample number: 10 difference of 25th, 50th, 90th percentiles between Empirical and theoretical confidence interval:
0.735975,0.892941,0.501129
```

When the sample is 100:

```
ans =

Sample number: 100, Theoretical 25th, 50th, 90th percentiles:1.922558,3.356694,7.779440


ans =

Smaple number: 100, Empirical 25th, 50th, 90th percentiles:1.863339,3.465558,8.152232


ans =

sample number: 100 difference of 25th, 50th, 90th percentiles between Empirical and theoretical confidence interval:
0.059219,0.108864,0.372792
```

When the sample number is 1000:

```
ans =

Sample number: 1000, Theoretical 25th, 50th, 90th percentiles:1.922558,3.356694,7.779440


ans =

Smaple number: 1000, Empirical 25th, 50th, 90th percentiles:1.925638,3.246557,8.004171


ans =

sample number: 1000 difference of 25th, 50th, 90th percentiles between Empirical and theoretical confidence interval:
0.003080,0.110137,0.224730
```

Comment:

the more sample number, the lower value $\left|\left|F_{10}^{*}(x) - F(x)\right|\right|_{\infty}$ is, which means the closer that the theretical results to the emprical results.

Q3:

A geyser is a hot spring characterized by an intermittent discharge of water and steam. Old Faithful is a famous cone geyser in Yellowstone National Park, Wyoming. It has a predictable geothermal discharge and since 2000 it has erupted every 44 to 125 minutes. Refer to the addendum data file that contains waiting times and the durations for 272 eruptions. Compute a 95% statistical confidence interval for the waiting time using data from only the first 15 eruptions. Compare this to a 95% bootstrap confidence interval using the same 15 data samples. Repeat these calculation using all the data samples. Comment on the relative width of the confidence intervals when using only 15 samples vs using all samples.

Analysis:

when 15 samples are used, t-distribution is used. Command "tinv" command in MATLAB can be used to determine $\alpha/2$; similarly, when all the 272 samples are used, it is better to apply z-distribution. And MATLAB command "norminv" $z_{\alpha/2.}$

After that, to obtain the confidence interval, we apply the equation $\overline{X} = \pm \frac{s}{\sqrt{n}} z_{\alpha/2}$.

Finally, we apply bootstr method to get a new confidence interval and compare the statistical results with it.

Code:

```
y15=x(1:15,3)'
yall=x(:,3)'


StdAll=std(yall);
Std15=std(y15);
MeanAll=mean(yall);
Mean15=mean(y15);

%t - distrubition, sample number less than 30
t_afa = tinv(0.95,14)
Bound_15(1)=Mean15-Std15/sqrt(15)*t_afa;
Bound_15(2)=Mean15+Std15/sqrt(15)*t_afa;

%z distribution
z_afa = norminv([0.025 0.975],0,1)
BoundAll(1)=Mean15-z_afa(2)*StdAll/sqrt(272);
BoundAll(2)=Mean15+z_afa(2)*StdAll/sqrt(272);

stats = bootstrp(1000,@mean,yall);
SortAll=sort(stats);
BoundBootAll(1)=SortAll(25);
BoundBootAll(2)=SortAll(975);

stats = bootstrp(1000,@mean,y15);
Sort15=sort(stats);
BoundBoot15(1)=Sort15(25);
BoundBoot15(2)=Sort15(975);
BoundAll
Bound_15
BoundBootAll
BoundBoot15
```

Results：

```
t_afa =

    1.7613


z_afa =

   -1.9600    1.9600
```

The condidence interval for the 15 samples is [69.3177,72.5490] and for all samples is [64.0547,77.8119]. And the bootstr method to get the confidence interval for all samples are [69.1360,72.4779] and for the 15 samples is [63.0667,78.000].

```
BoundAll =

   69.3177   72.5490


Bound_15 =

   64.0547   77.8119


BoundBootAll =

   69.2868   72.4559


BoundBoot15 =

   63.2667   77.8667
```