Weining Bao weiningb@usc.edu 8287723653

Assignment 3

1. Estimate π by the area method (including confidence intervals). Simulate this approach. Draw a graph of the successive values of the estimator as the number of samples increases. How many points do you need to use for your estimate to be within 0.01 of the true value of π (with probability 0.99)?

2. Evaluate an Integral as suggested in class to find π .

Matlab Simulation

with code at last

1.

Estimate π by the area method (including confidence intervals). Simulate this approach. Draw a graph of the successive values of the estimator as the number of samples increases.
How many points do you need to use for your estimate to be within 0.01 of the true value of π (with probability 0.99)?

Draw a quadrant with center at the origin and radius r=1. So the quadrant is in a rectangle with side 1 and one vertex at origin. The points generated randomly inside the unit square should have a probability of $\frac{area \quad of \quad quadrant}{area \quad of \quad square} = \frac{\pi}{4}$ to be inside the quadrant. The points can be generated with x, y coordinates all uniform RV in the range of (0,1). So the result whether the point is in the unit square or not can be represented by the RV $P = \{ \begin{matrix} 1 & X^2 + Y^2 \le 1 \\ 0 & otherwise \end{matrix}$. The mean of P E(P) = probability that one point falls into the quadrant = $p = \frac{\pi}{4}$, and the variance of P $VAR(P) = \sigma_P^2 = p(1-p)$.

The sample mean $\hat{P} = \frac{\sum_1^n P_i}{n}$ can be used to estimate the mean of P E(P). However $\hat{P}$ itself is a RV with mean $E(\hat{P}) = E(\frac{\sum_1^n P_i}{n}) = \frac{n}{n}E(P_i) = E(P) = p$ and variance $\sigma_{\hat{P}}^2 = VAR(\hat{P}) = VAR(\frac{\sum_1^n P_i}{n}) = \frac{n}{n^2}VAR(P_i) = \frac{VAR(P)}{n} = \frac{\sigma_P^2}{n}$. If n is large enough, $\hat{P}$ also follows the normal distribution, according to the central limit theorem. So the confidence interval is $\Pr\{p - \beta\sigma_{\hat{P}} \le \hat{p} \le p + \beta\sigma_{\hat{P}}\} = 1 - \alpha$. If $\alpha = 0.01$, then the estimate sample mean is of probability

0.99 to be within $\beta\sigma_{\hat{P}}$ of the true value while $\beta = 2.56$ according to the standard normal distribution table. If $4\beta\sigma_{\hat{P}} \le 0.01$ is desirable, then $\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}} \le \frac{0.01}{4*2.56}$. If $\pi = 3.1415927$ is used, then $n \ge 176736$. That means at least 176736 points are needed for estimate to be within 0.01 of the true value of π with probability 0.99.

Here the value of π is used to calculate the number of points needed. However the value of π is not available in real estimate. Instead, it is exactly the value we want to estimate. This is also a problem when we estimate π with sample mean $\hat{P}$ and certain confidence interval $\{p - \beta\sigma_{\hat{P}} \le \hat{p} \le p + \beta\sigma_{\hat{P}}\}$. $\hat{P}$ can be computed with all the samples. If $\sigma^2_{\hat{P}} = \frac{\sigma^2_P}{n} = \frac{p(1-p)}{n}$ is used here, the computation will involve the solution of a set of two one-variable quadratic inequalities. The calculation is quite involved. Any other way to get around is using another estimate for $\sigma^2_{\hat{P}}$.

Two estimates can be used as $\sigma^2_{\hat{P}}$. One is using sample mean as mean in the calculation of $\sigma^2_{\hat{P}} = \frac{p(1-p)}{n} \approx \frac{\hat{p}(1-\hat{p})}{n} = s^2_{\hat{P}}$. Another way is using the sample variance $s^2_{P_i} = \frac{\sum_1^n (P_i - \hat{p})^2}{n-1}$ as the variance of P and then calculate the variance of $\hat{P}$, $s^2_{\hat{P}} = \frac{1}{n} s^2_{P_i}$. In this way the confidence interval becomes $\Pr\{p - \beta s_{\hat{p}} \le \hat{p} \le p + \beta s_{\hat{p}}\} = 1 - \alpha$. So the equivalent confidence interval $\Pr\{4(\hat{p} - \beta s_{\hat{p}}) \le \pi \le 4(\hat{p} + \beta s_{\hat{p}})\} = 1 - \alpha$ gives a estimate of π with certain confidence.

Two estimates are both used in the simulation to see which way gives better interval. Besides 1000, 10000, 100000 and 1000000 points are used sequentially to compute the corresponding points needed for the estimate to be within 0.01 of the true value of π with probability 0.99 with two estimates for $s^2_{\hat{P}}$. Here are the results:

Prob 1:

Generating 1000 random points inside unit square.
Generating 1000 corresponding RVs.
Mean: 0.804000. Variance: 0.000158.
With confidence of 0.99 from 1000 points estimate,
With first estimate, 3.087455 <= pi <= 3.344545
With second estimate, 3.087390 <= pi <= 3.344610
With first estimate, 165238.800384 points are needed to estimate within 0.01 of the true value of pi
With second estimate, 165404.204589 points are needed to estimate within 0.01 of the true value of pi

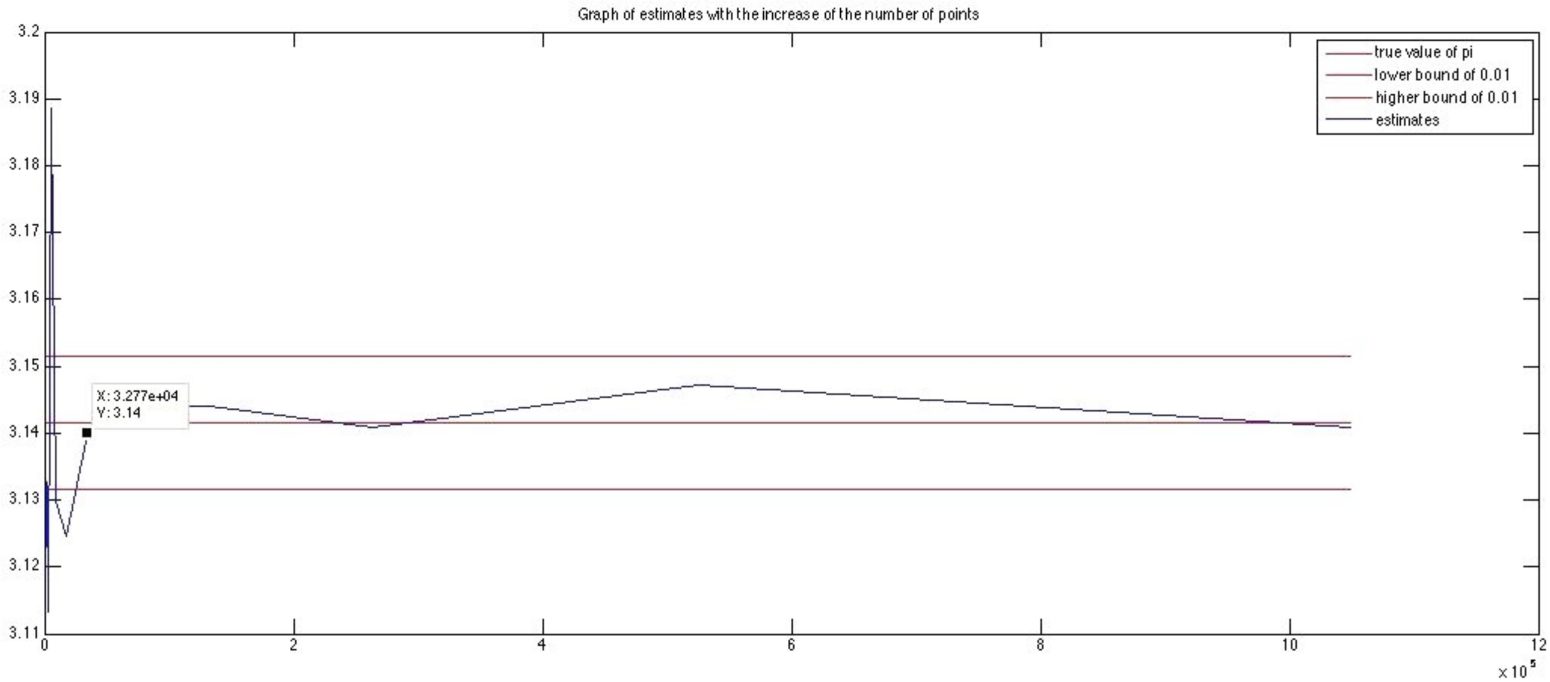
Generating 10000 random points inside unit square.
Generating 10000 corresponding RVs.
Mean: 0.782200. Variance: 0.000017.
With confidence of 0.99 from 10000 points estimate,
With first estimate, 3.086534 <= pi <= 3.171066
With second estimate, 3.086532 <= pi <= 3.171068
With first estimate, 178638.720860 points are needed to estimate within 0.01 of the true value of pi
With second estimate, 178656.586519 points are needed to estimate within 0.01 of the true value of pi

Generating 100000 random points inside unit square.
Generating 100000 corresponding RVs.
Mean: 0.785590. Variance: 0.000002.
With confidence of 0.99 from 100000 points estimate,
With first estimate, 3.129070 <= pi <= 3.155650
With second estimate, 3.129070 <= pi <= 3.155650
With first estimate, 176620.413282 points are needed to estimate within 0.01 of the true value of pi
With second estimate, 176622.179504 points are needed to estimate within 0.01 of the true value of pi

Generating 1000000 random points inside unit square.
Generating 1000000 corresponding RVs.
Mean: 0.785569. Variance: 0.000000.
With confidence of 0.99 from 1000000 points estimate,
With first estimate, 3.138073 <= pi <= 3.146479
With second estimate, 3.138073 <= pi <= 3.146479
With first estimate, 176632.990258 points are needed to estimate within 0.01 of the true value of pi
With second estimate, 176633.166890 points are needed to estimate within 0.01 of the true value of pi

As we can see, two estimates give similar interval. All the cases with different number of points generated give ~ 176600 number of points needed for certain confidence and error, the requirement is actually met when the number of points increase till range between 100000 and 1000000.

To further verify this, graph of the successive values of the estimator as the number of samples increases is also plotted. 11 sample sizes of 2^(10:20) are estimated. So each step doubles the sample size. Since only 11 sample sizes are estimated, it is just coarse grain estimate of the number of points needed to be within 0.01 of the true value of π . Here is the resulting plot. It can be seen that the estimates are definitely within the 0.01 bound after the number of points reach 32770. Calculation of ~176600 points needed gives a more strict requirement to meet the required confidence interval.



Chi-square goodness of fit test is also made on P_i to see if it follows a Bernoulli distributed DRV. There are only 2 possible outcomes with probability $p = \frac{\pi}{4}$ and $1 - p = 1 - \frac{\pi}{4}$. If the number of trials/points in this estimate is n, then the expected numbers of outcomes are np and n(1-p) respectively. The corresponding RV χ^2 has only degree of freedom 1 and is defined as $q_1 = \sum_{all \quad outcomes} \frac{(observed - \exp ected)^2}{\exp ected}$. The q1 should be smaller than critical value α based on certain desired level of significance. According to the chi-square distribution table, for the degree of freedom 1, $\alpha = 6.635$ for 0.99 confidence. Here are the results:

799 points in quadrant and 201 not
The chi-square value is 1.097670.

7922 points in quadrant and 2078 not
The chi-square value is 2.744916.

78433 points in quadrant and 21567 not
The chi-square value is 0.676943.

785834 points in quadrant and 214166 not
The chi-square value is 1.127000.

So the RV P_i fits Bernoulli distributed DRV of $p = \frac{\pi}{4}$ well.

2.

Evaluate an Integral as suggested in class to find π .

$I = \int_0^1 \frac{1}{1+x^2} dx = \frac{\pi}{4} = \int_0^1 g(x) dx = E(g(x))$ if x follows uniform distribution in the range of (0, 1). So to estimate π , sequence of uniform RVs x in the range of (0, 1) are generated and the value $\frac{1}{1+x^2}$ is calculated. Mean of these values can be used as an estimate of integral $\frac{\pi}{4}$.

Similarly, 1000, 10000, 100000 and 1000000 RVs x are generated and mean of values $\frac{1}{1+x^2}$ are calculated. Here are the results:
Generating 1000 uniform RVs x from (0, 1) and corresponding values 1/(1+x^2).
Mean: 0.784028. Variance: 0.026636.
Estimate for pi: 3.136113
Generating 10000 uniform RVs x from (0, 1) and corresponding values 1/(1+x^2).
Mean: 0.785615. Variance: 0.025971.
Estimate for pi: 3.142458
Generating 100000 uniform RVs x from (0, 1) and corresponding values 1/(1+x^2).
Mean: 0.785435. Variance: 0.025838.
Estimate for pi: 3.141741
Generating 1000000 uniform RVs x from (0, 1) and corresponding values 1/(1+x^2).
Mean: 0.785482. Variance: 0.025862.
Estimate for pi: 3.141927

As we can see, the sample mean is a good estimate of the value of π . Besides, the estimate gets close to the true value with the increase of sample size.

Code:

```
% prob 1
fprintf('\nProb 1: \n\n');
for count = 3:6 % test 1000, 10000, 100000, 1000000 points
    num = 10^count;
    fprintf('Generating %d random points inside unit square.\n', num);
    x = rand(num, 1); % generate x coordinate
    y = rand(num, 1); % generate y coordinate
    fprintf('Generating %d corresponding RVs.\n', num);
    p = ones(num, 1); % generate random variable
    for i = 1:num
        if x(i)^2 + y(i)^2 > 1
            p(i) = 0;
        end
    end
    chi = (sum(p)-num*pi/4)^2/(num*pi/4) + (sum(p)-num*pi/4)^2/(num-num*pi/4);
    m = mean(p);
    sv = var(p); % var() function normalize with 1/(n-1)
    v = m * (1 - m) / num; % sample mean as mean to calculate variance
```

```
    fprintf('Mean: %f. Variance: %f.\n', m, v);
    fprintf('%d points in quadrant and %d not\n', sum(p), num-sum(p));
    fprintf('The chi-square value is %f.\n', chi);
    fprintf('With confidence of 0.99 from %d points estimate,\n', num);
    fprintf('With first estimate, %f <= pi <= %f\n', ...
        4*(m-2.56*sqrt(v)), 4*(m+2.56*sqrt(v)));
    fprintf('With second estimate, %f <= pi <= %f\n', ...
        4*(m-2.56*sqrt(sv/num)), 4*(m+2.56*sqrt(sv/num)));
    fprintf('With first estimate, %f points are needed ', 16*2.56^2*m*(1-m)/(0.01)^2);
    fprintf('to estimate within 0.01 of the true value of pi\n');
    fprintf('With second estimate, %f points are needed ', 16*2.56^2*sv/(0.01)^2);
    fprintf('to estimate within 0.01 of the true value of pi\n\n');
end
e = zeros(11,1); % corresponding estimate value
for count = 10:20 % test 2^(10:20) points of estimates
    num = 2^count;
    x = rand(num, 1); % generate x coordinate
    y = rand(num, 1); % generate y coordinate
    p = ones(num, 1); % generate random variable
    for i = 1:num
        if x(i)^2 + y(i)^2 > 1
            p(i) = 0;
        end
    end
    e(count-9) = 4 * mean(p); % generate estimate
end
t = ones(11,1) .* pi; % true value of pi
l = ones(11,1) .* (pi - 0.01); % lower bound of confidence interval
h = ones(11,1) .* (pi + 0.01); % higher bound of confidence interval
n = 2.^(10:20);
figure;
plot(n, t, 'r'); hold on;
plot(n, l, 'r');
plot(n, h, 'r');
plot(n, e); hold off;
title('Graph of estimates with the increase of the number of points');
legend('true value of pi', 'lower bound of 0.01', 'higher bound of 0.01',...
    'estimates');

% prob 2
fprintf('\nProb 2: \n\n');
for count = 3:6 % test 1000, 10000, 100000, 1000000 points
    num = 10^count;
    fprintf('Generating %d uniform RVs x from (0, 1) ', num);
    fprintf('and corresponding values 1/(1+x^2).\n');
    v = 1 ./ (1 + rand(num, 1).^2);
    fprintf('Mean: %f. Variance: %f.\n', mean(v), var(v));
    fprintf('Estimate for pi: %f\n\n', 4*mean(v));
end
```