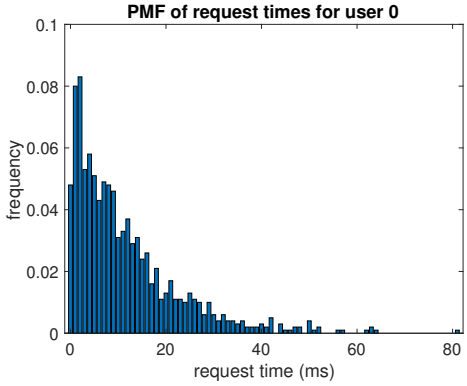
Model Solution

1(a) The PMF of a random variable Y taking values y_i, $i = 1, \ldots, n$ is the probabilities $P(Y = y_i)$, $i = 1, \ldots, n$. We estimate this from the data by listing the set of request times observed and calculating the fraction that each request time occurs in the data, see plot below:



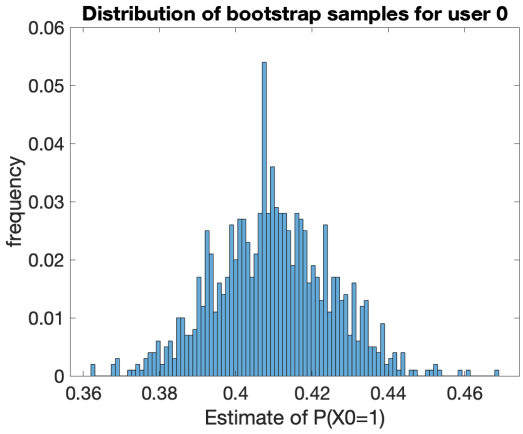
1(b) Since X_0 is an indicator random variance $E[X_0] = Prob(X_0 = 1)$. We estimate $Prob(X_0 = 1)$ by the empirical mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ where $n = 1000$ is the number of request times observed and $Y_i = 1$ is the i'th request exceeds 10ms and $Y_i = 1$ otherwise. From the data we calculate $\bar{X} = 0.41$. Visually, this seems consistent with the plotted PMF, where it looks like roughly half the area of the PMF lies to the left of 10ms and half to the right.

1(c) To use the CLT and Chebyshev we need the mean and variance of $\bar{X}$. We estimate mean $E[\bar{X}] = 0.41 = \mu$ from (b). Assuming the Y_i are independent $var(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} var(Y_i) = \frac{var(Y_i)}{n} = \sigma^2$. From the data using the empirical mean $\frac{1}{n}\sum_{i=1}^{n}(Y_i - 0.41)^2$ we estimate $var(Y_i) = 0.242142$ and so $\sigma^2 = 0.000242$. The 95% confidence interval using the CLT is then $[\mu - 2\sigma, \mu + 2\sigma] = [0.378878, 0.441122]$. The 95% confidence interval using Chebyshev is $[\mu - \sigma/\sqrt{0.05}, \mu + \sigma/\sqrt{0.05}] = [0.340409, 0.479591]$. To generate a confidence interval using bootstrapping we sample 1000 values with replacement from the set of Y_i, $i = 1, \ldots, n$ values, then calculate the empirical mean of these values to obtain an estimate for $\bar{X}$. Note that we need to use 1000 samples since we use $n = 1000$ when calculating $\bar{X}$. We repeat this sampling process to obtain a set of N estimates for $\bar{X}$. An example distribution of values for $N = 1000$ is shown in the figure below. Using this distribution we can then either visually estimate a 95% confidence interval or using the matlab prctile function obtain interval [0.378500,0.442000]. We have some freedom in the choice of N to use in bootstrapping, the value of $N = 1000$ was selected since using larger values did not change the estimated confidence interval by much.

Discussion: Observe that the confidence intervals obtained by the CLT and bootstrapping are v similar. Also that the confidence interval obtained using Chebyshev is larger - we expect this to always be true since Chebyshev gives an upper bound on the interval. In general, the CLT assumes that n is large enough that the data is Gaussian distributed and when this is true it gives an accurate estimate of the confidence interval, Chebyshev gives an upper bound on the confidence interval that is valid for all values of n but may be larger than necessary, bootstrapping gives an accurate empirical estimate provided the data is representative, i.e. is distributed similarly to when longer experiments are run, but otherwise may be inaccurate. For both the CLT and bootstrapping it's hard to quantify the inaccuracy.

2. Using the same approach as in Q1b the probabilities for each of the 6 users that the request times exceed 10ms are calculated to be: 0.4100 0.5220 0.4720 0.2640 0.4420 0.3440.

3. Estimate of $P(Z_n > 10) = \sum_{u=1}^{6} P(Z_n > 10|U_n = i)P(U_n = i) = \sum_{u=1}^{6} P(X_i = 1)P(U_n = i)$. Substituting in the values for $P(X_i = 1)$ from Q2 and the supplied values for $P(U_n = i)$ we calculate estimate $P(Z_n > 10) = 0.400548$



4 . By Bayes Rule $P(U_n = 0|Z_n > 10) = \frac{P(Z_n>10|U_n=0)P(U_n=0)}{P(Z_n>10)}$. Substituting the value for $P(Z_n > 10)$ from Q3, the value for $P(Z_n > 10|U_n = 0) = P(X_0)$ from Q1b and the supplied value of $P(U_n = 0)$ we calculate $P(U_n = 0|Z_n > 10) = 0.051060$.

5. Simulation pseudo-code:

sum=0
for $n = 1$ **to** N **do**
randomly draw user i according to supplied PMF $P(U_n = i)$, $i = 1, \ldots, 6$
draw a Bernoulli RV X_n with prob $P(X_n = 1) = P(Z_n > 10|U_n = i)$
$sum+ = X_n$
end for
output estimate sum/N for $P(Z_n > 10)$

To draw a Bernoulli RV use $rand() < P(X_n = 1$ where $rand()$ is a uniformly distributed RV. To draw a user use: for j=1:length(user_probabilities) {if ($r <$ cumsum(user_probabilities)(j)) break; } where r is a uniformly distributed RV and user_probabilities is a vector with elements $P(U_n = i)$, $i = 1, \ldots, 6$.

The number of iterations N is a key parameter in the simulation. The estimate sum/N is a RV that changes from run to run of the simulation. As we increase N we expect the variability in sum/N to decrease since the law of large numbers tells us that sum/N concentrates onto $P(Z_n > 10)$ with probability one as N goes to infinity. For $N = 100,000$ we observe the variability of sum/N to be fairly low with values of the estimate from three runs being 0.405360, 0.400510, 0.401680.

The estimate sum/N from the simulation for $P(Z_n > 10)$ is a random variable. So we need to be careful when comparing the output from the simulation with the value calculated in Q3 and should ideally estimate confidence intervals for both and compare those. However, when N is large the confidence interval for sum/N is small, and as a rough check with $N = 100,000$ we can see that the estimates from the simulation are similar to the estimate calculated in Q3.

Appendix

```
% user probs taken from first line of file
user_probabilities=[ 0.049882450727516 0.16660145902205 0.15146859569335
    0.24061446046328 0.23941009750605 0.15202293658775];

% load rest of data
data = load('midterm2020.txt');
[m,n]=size(data);

%Q1a - there are many ways to code this, here is just one example (not
    the
%shortest or simplest)
vals = unique(data(:,1)); % get the set of observed request times
pmf=[];
for val=vals',
    pmf=[pmf;val,length(find(data(:,1)==val))/m];
end
figure(1), bar(pmf(:,1),pmf(:,2));
xlabel('request time (ms)'), ylabel('frequency'), title('PMF of request
    times for user 0')
set(gca,'fontsize',18)
%Q1b
m0=sum(data(:,1)>10)/m; % estimate of mean
fprintf("Estimate of P(X0=1)=%f\n",m0);

%Q1c
v0=var(data(:,1)>10)/m; % estimate of variance of mean
fprintf("Estimate of var(Y_i)=%f and var(Xbar)%f\n",v0*m,v0);

% CLT 95% confidence interval
fprintf("CLF confidence interval [%f,%f]\n",m0-2*sqrt(v0),m0+2*sqrt(v0));
% Chebyshev 95% confidence interval
fprintf("Chebyshev confidence interval [%f,%f]\n",m0-sqrt(v0/0.05),m0+
    sqrt(v0/0.05));
% Boostrap confidence interval
N=1000; % number of runs of bootstrap sampling
X=[];
for i=1:N,
    X=[X,mean(data(randi(m,1,m),1)>10)];
end
figure(2),histogram(X,100,'Normalization','probability')
xlabel('Estimate of P(X0=1)'), ylabel('frequency'), title('Distribution
    of bootstrap samples for user 0')
set(gca,'fontsize',18)
fprintf("Bootstrap confidence interval [%f,%f]\n",prctile(X,2.5),prctile(
    X,97.5));

%Q2
p=mean(data>10)

%Q3
PZ10=sum(user_probabilities.*p);
```

```
fprintf("Estimate of P(Zn>10)=%f\n",PZ10)

%Q4
fprintf("Estimate of P(Un=0|Zn>10)=%f\n",p(1)*user_probabilities(1)/PZ10)

%Q5
N=100000; % number of times to run simulation
UP=cumsum(user_probabilities);
count=0;
for i=1:N,
    % draw user randomly
    r=rand();
    for j=1:length(user_probabilities),
        if (r<UP(j)) break; end
    end
    user = j;
    % draw whether request exceeds 10ms randomly
    count = count + (rand()<p(user));
end
% estimate using empirical mean
fprintf("Estimate of P(Zn>10) from sim =%f\n",count/N)
```