Tutorial- 7 Submission

MA 201 Probability and Statistics (2021-22) (3-1-0-4)

B. Tech. II year CSE & IT

Name - Archit Agrawal

Roll No - 202051213

Qus - 1 The numbers of blocked intrusion attempts on each day during the first two weeks of the month were 56, 47, 49, 37, 38, 60, 50, 43, 43, 59, 50, 56, 54, 58.

After the change of firewall settings, the numbers of blocked intrusions during the next 20 days were 53, 21, 32, 49, 45, 38, 44, 33, 32, 43, 53, 46, 36, 48, 39, 35, 37, 36, 39, 45. Comparing the number of blocked intrusions before and after the change,

(a) construct side-by-side stem-and-leaf plots;

(b) compute the five-point summaries and construct parallel boxplots;

(c) comment on your findings.

(d) Construct a 95% confidence interval for the difference between the average number of intrusion attempts per day before and after the change of firewall settings (assume equal variances).

(e) Can we claim a significant reduction in the rate of intrusion attempts? The number of intrusion attempts each day has approximately Normal distribution. Compute Pvalues and state your conclusions under the assumption of equal variances and without it. Does this assumption make a difference?

```
%(a) construct side-by-side stem-and-leaf plots;
clear all
close all
before = [56, 47, 49, 37, 38, 60, 50, 43, 43, 59, 50, 56, 54, 58]';
after = [53, 21, 32, 49, 45, 38, 44, 33, 32, 43, 53, 46, 36, 48, 39, 35, 37, 36, 39, 45
stemleafplot(before)
```

```
  0 |
  1 |
  2 |
  3 | 7 8
  4 | 3 3 7 9
  5 | 0 0 4 6 6 8 9
  6 | 0
key: 36|5 = 365
stem unit: 10
leaf unit: 1
```

```
stemleafplot(after)
```

```
  0 |
  1 |
```

```
  2 | 1
  3 | 2 2 3 5 6 6 7 8 9 9
  4 | 3 4 5 5 6 8 9
  5 | 3 3
key: 36|5 = 365
stem unit: 10
leaf unit: 1
```

(b) compute the five-point summaries and construct parallel boxplots;

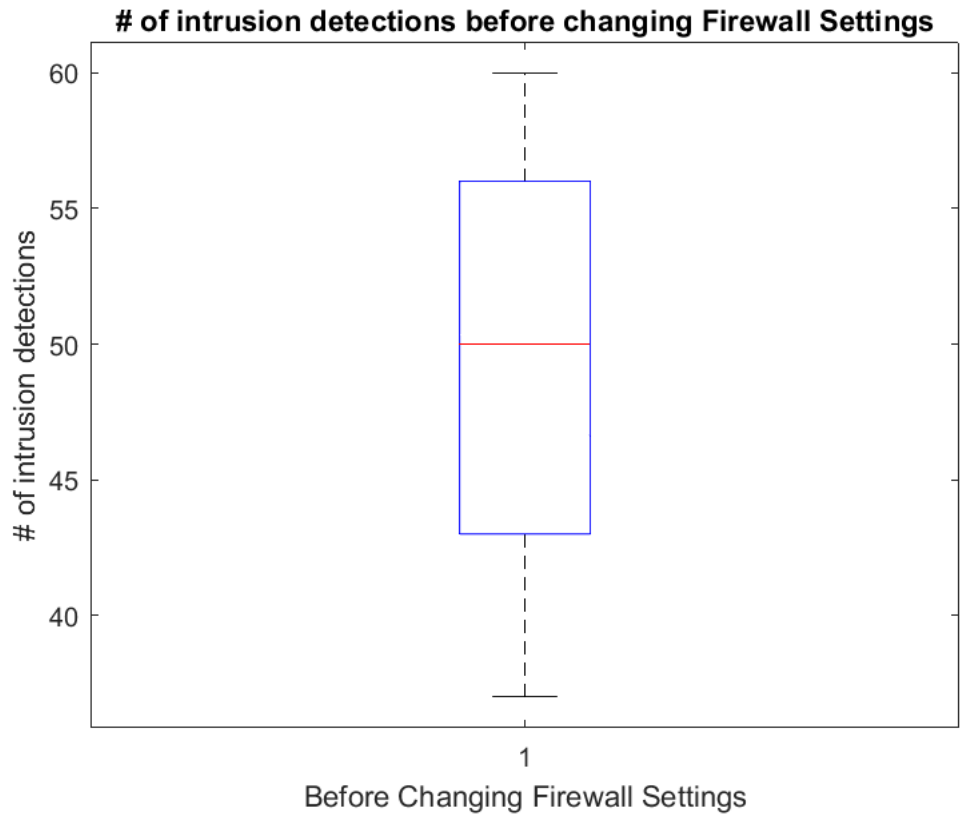
```
%(b) compute the five-point summaries and construct parallel boxplots;
summary = @(sample)([min(sample) quantile(sample,0.25) quantile(sample,0.50) quantile(s
boxplotdatasummary = @(sample)([sort([min(sample) quantile(sample,0.25)-1.5*iqr(sample)
beforesummary = summary(before)
```

```
beforesummary = 1×5
    37    43    50    56    60
```

```
beforeboxplotsummary = boxplotdatasummary(before)
```

```
beforeboxplotsummary = 1×7
   23.5000   37.0000   43.0000   50.0000   56.0000   60.0000   62.5000
```

```
boxplot(before);
```




```
xlabel('Before Changing Firewall Settings'); ylabel('# of intrusion detections');
title('# of intrusion detections before changing Firewall Settings')
aftersummary = summary(after)
```

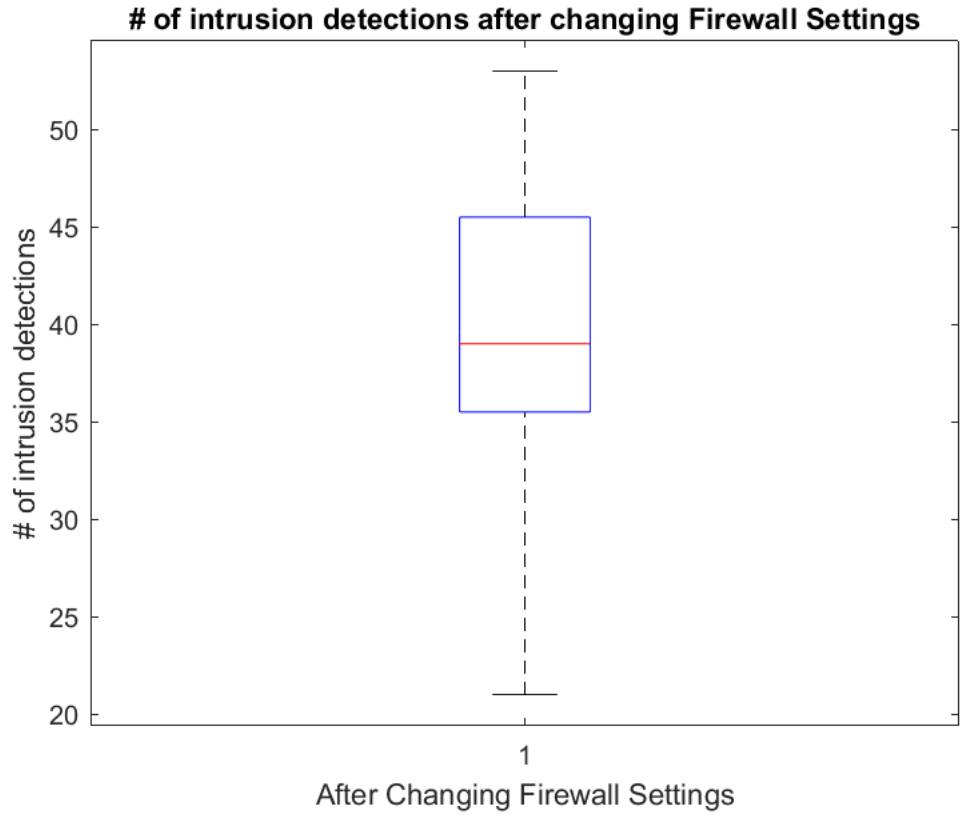
```
aftersummary = 1×5
```

```
    21.0000   35.5000   39.0000   45.5000   53.0000
```

```
afterboxplotsummary = boxplotdatasummary(after)
```

```
afterboxplotsummary = 1×7
    20.5000   21.0000   35.5000   39.0000   45.5000   50.5000   53.0000
```

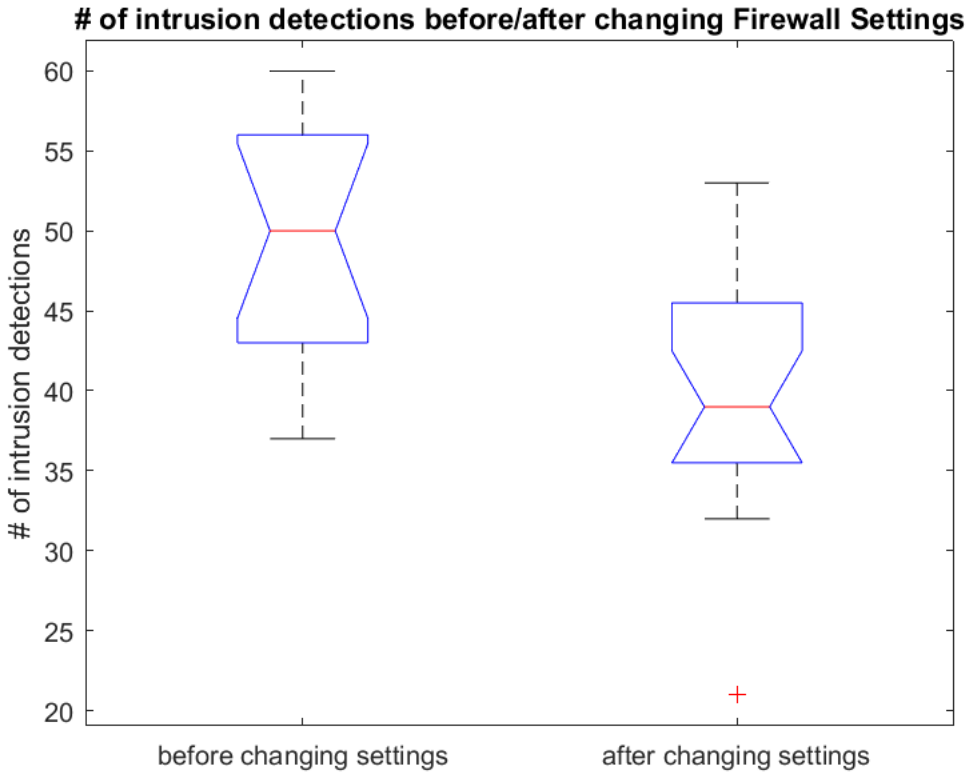
```
boxplot(after);
```




```
xlabel('After Changing Firewall Settings'); ylabel('# of intrusion detections');
title('# of intrusion detections after changing Firewall Settings')
```

(c) comment on your findings.

```
%(c) comment on your findings.
boxplot([before; after],[repmat({'before'},length(before),1); repmat({'after'},length(a
ylabel('# of intrusion detections'); title('# of intrusion detections before/after char
```




```
n = length(before); m = length(after);
xhat = mean(before); yhat = mean(after);
sp = sqrt((n-1)*var(before)+(m-1)*var(after)/(n+m-2))
```

```
sp = 28.1354
```

```
ci =.95; alpha = 1-ci;
ConfInterval = xhat - yhat + tinv([alpha/2 1- alpha/2],n+m-2)*sp*sqrt(1/n+1/m)
```

```
ConfInterval = 1×2
  -10.1706   29.7706
```

P_Value

```
Tobs = (0-yhat + xhat)/(sp*sqrt(1/n+1/m)) % H0: p2=p1 HA: p2>p1
```

```
Tobs = 0.9996
```

```
P_value = tcdf(Tobs,n+m-2,"Upper") % Accepted null hypothesis.
```

```
P_value = 0.1625
```

```
P_value = tcdf(Tobs,(var(before)/n+var(after)/m)^2/((var(before)/n)^2/(n-1)+(var(after)
```

```
P_value = 0.1629
```

Qus-2) A network provider investigates the load of its network. The number of concurrent users is recorded at fifty locations (thousands of people),

17.2 22.1 18.5 17.2 18.6 14.8 21.7 15.8 16.3 22.8 24.1 13.3 16.2 17.5 19.0 23.9 14.8 22.2 21.7 20.7 13.5 15.8 13.1 16.1 21.9 23.9 19.3 12.0 19.9 19.4 15.4 16.7 19.5 16.2 16.9 17.1 20.2 13.4 19.8 17.7 19.7 18.7 17.6 15.9 15.2 17.1 15.0 18.8 21.6 11.9

(a) Compute the sample mean, variance, and standard deviation of the number of concurrent users.

(b) Estimate the standard error of the sample mean.

(c) Compute the five-point summary and construct a boxplot.

(d) Compute the interquartile range. Are there any outliers?

(e) It is reported that the number of concurrent users follows approximately Normal distribution. Does the histogram support this claim?

Hint -

```
X = [17.2 22.1 18.5 17.2 18.6 14.8 21.7 15.8 16.3 22.8 24.1 13.3 16.2 17.5 19.0 23.9 14
%(a) Compute the sample mean, variance, and standard deviation of the number of concurr
mu = sum(X)/length(X);
mean_var_sd = [mu sum((X - mu).^2)/(length(X)-1) sqrt(sum((X - mu).^2)/(length(X)-1))]
```

```
mean_var_sd = 1×3
   17.9540    9.9682    3.1573
```

or

```
% (a) Compute the sample mean, variance, and standard deviation of the number of concur
mean_var_sd = @(X)([mean(X) var(X) std(X)]);
mean_var_sd(X)
```

```
ans = 1×3
   17.9540    9.9682    3.1573
```

(b) Estimate the standard error of the sample mean.

We already know that the standard error of this sample estimator is $\sigma(\overline{X}) = \sigma/\sqrt{n}$, and it can be estimated by $s(\overline{X}) = s/\sqrt{n}$

```
%(b) Estimate the standard error of the sample mean.
mu = sum(X)/length(X);
std_err_mean_estimator = sqrt(sum((X - mu).^2)/(length(X)-1))/length(X)
```

```
std_err_mean_estimator = 0.0631
```

or

```
%(b) Estimate the standard error of the sample mean.
std_err_mean_estimator = std(X)/length(X)
```

```
std_err_mean_estimator = 0.0631
```

(c) Compute the five-point summary and construct a boxplot.

```
%(c) Compute the five-point summary and construct a boxplot.
```

```
summary(X)
```

```
ans = 1×5
   11.9000   15.8000   17.5500   19.9000   24.1000
```

```
boxplotdatasummary(X)
```

```
ans = 1×7
    9.6500   11.9000   15.8000   17.5500   19.9000   21.9500   24.1000
```

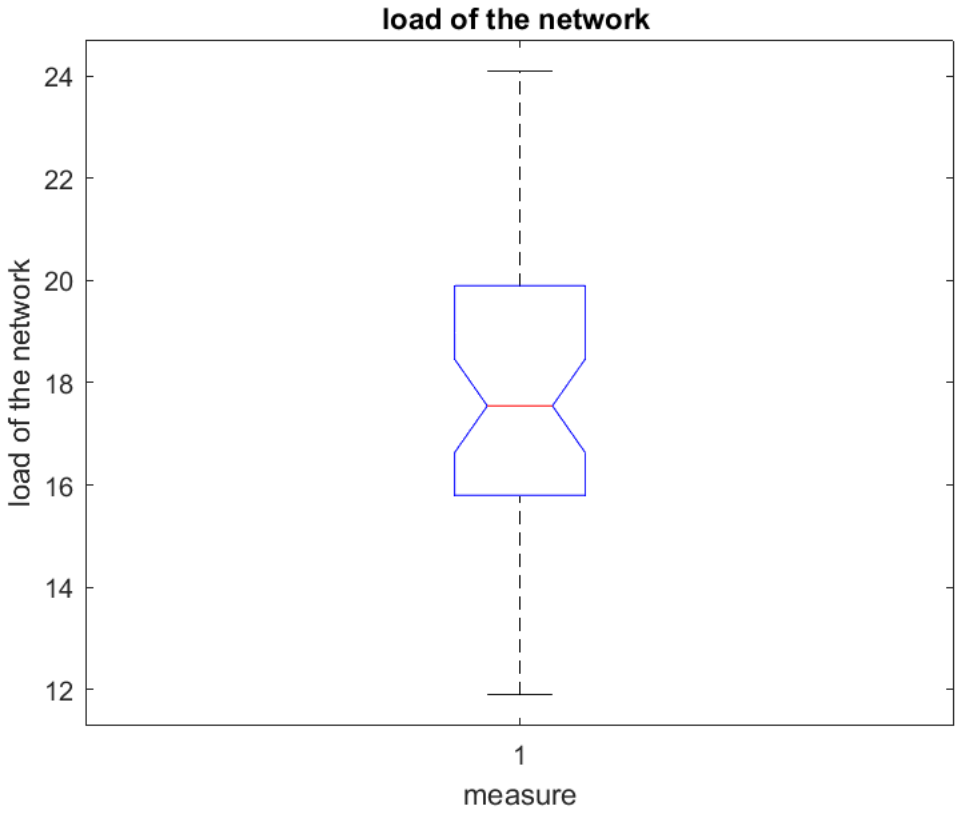
```
boxplot(after,'Notch','on','Labels',{'location'}); ylabel('load of the network'); title
%(d) Compute the interquartile range. Are there any outliers?
iqr(X)
```

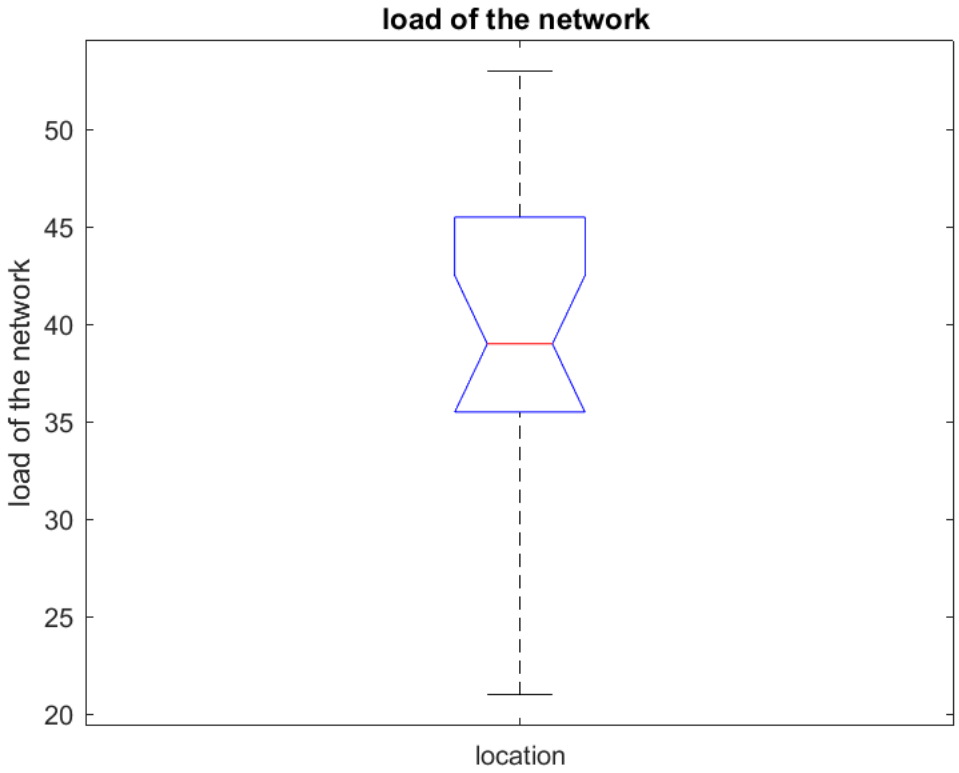
```
ans = 4.1000
```

```
[sum(X < (quantile(X,0.25)-1.5*iqr(X))) sum(X > (quantile(X,0.75)+1.5*iqr(X)))]
```

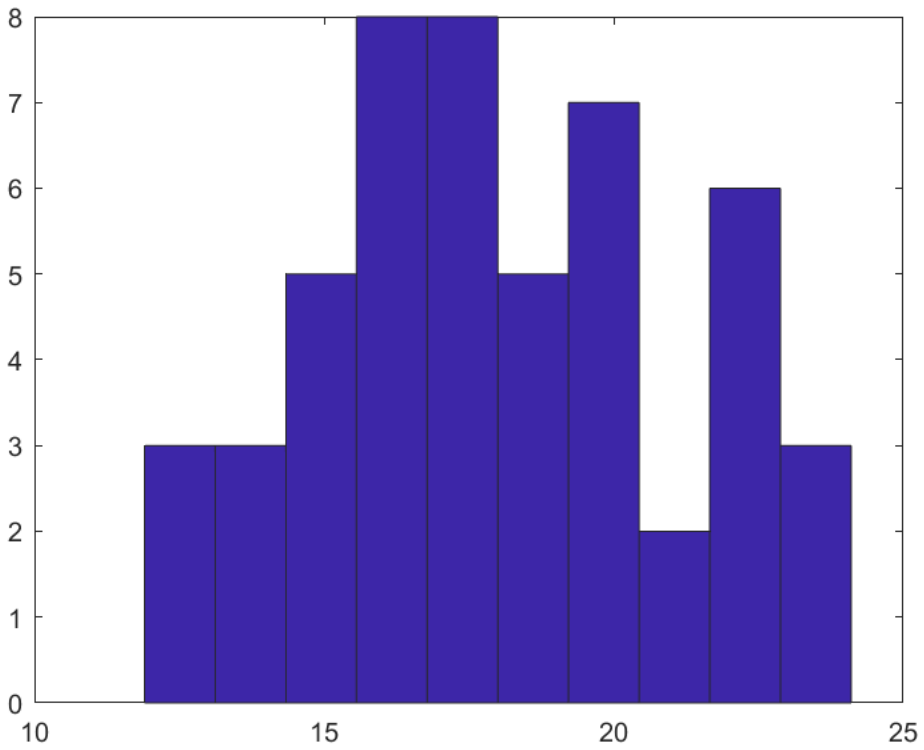
```
ans = 1×2
     0     0
```

```
MentionBoxplotOutlier(X); xlabel('measure'); ylabel('load of the network'); title('load
```






```
%(e) It is reported that the number of concurrent users follows approximately Normal di
hist(X,10)% claim rejected
```



Qus.3 Consider three data sets.

(1) 19, 24, 12, 19, 18, 24, 8, 5, 9, 20, 13, 11, 1, 12, 11, 10, 22, 21, 7, 16, 15, 15, 26, 16, 1, 13, 21, 21, 20, 19

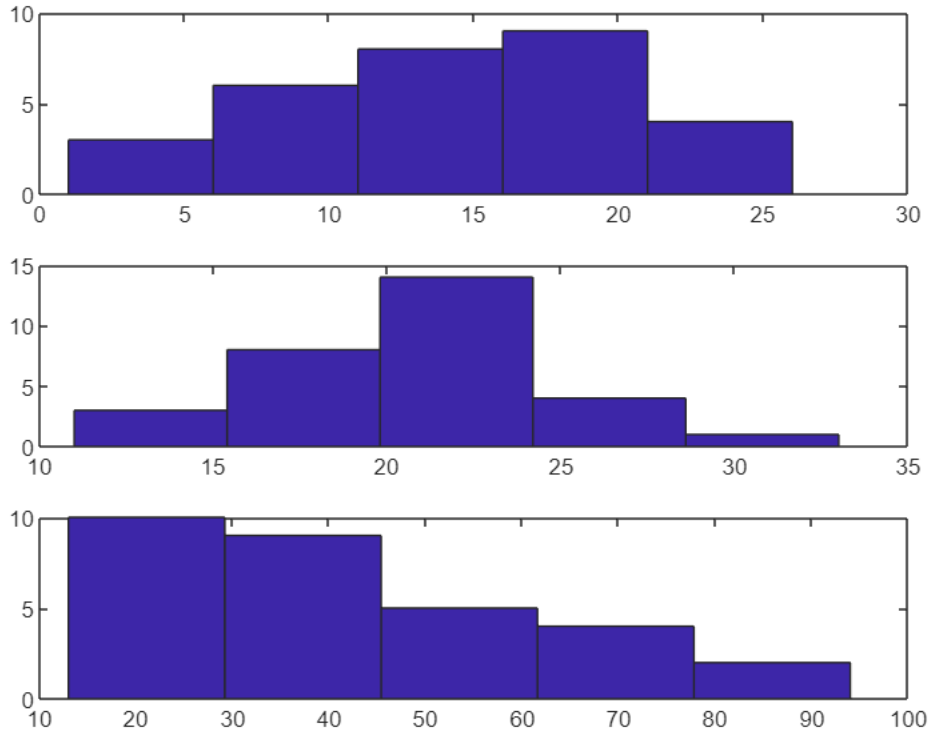
(2) 17, 24, 21, 22, 26, 22, 19, 21, 23, 11, 19, 14, 23, 25, 26, 15, 17, 26, 21, 18, 19, 21, 24, 18, 16, 20, 21, 20, 23, 33

(3) 56, 52, 13, 34, 33, 18, 44, 41, 48, 75, 24, 19, 35, 27, 46, 62, 71, 24, 66, 94, 40, 18, 15, 39, 53, 23, 41, 78, 15, 35

(a) For each data set, draw a histogram and determine whether the distribution is right-skewed, left-skewed, or symmetric.

(b) Compute sample means and sample medians. Do they support your findings about skewness and symmetry? How?

```
data1 = [19, 24, 12, 19, 18, 24, 8, 5, 9, 20, 13, 11, 1, 12, 11, 10, 22, 21, 7, 16, 15,
data2 = [17, 24, 21, 22, 26, 22, 19, 21, 23, 11, 19, 14, 23, 25, 26, 15, 17, 26, 21, 18
data3 = [56, 52, 13, 34, 33, 18, 44, 41, 48, 75, 24, 19, 35, 27, 46, 62, 71, 24, 66, 94
noofbins =5; subplot(311); hist(data1,noofbins); subplot(312); hist(data2,noofbins); su
```




```
[mean(data1) mean(data2) mean(data3); median(data1) median(data2) median(data3)]
```

```
ans = 2×3
   14.9667   20.8333   41.3000
   15.5000   21.0000   39.5000
```

Qus.4 The following data set represents the number of new computer accounts registered during ten consecutive days.

43, 37, 50, 51, 58, 105, 52, 45, 45, 10.

(a) Compute the mean, median, quartiles, and standard deviation.

(b) Check for outliers using the 1.5(IQR) rule.

(c) Delete the detected outliers and compute the mean, median, quartiles, and standard deviation again.

(d) Make a conclusion about the effect of outliers on basic descriptive statistics.

```
%(a) Compute the mean, median, quartiles, and standard deviation.
data = [43, 37, 50, 51, 58, 105, 52, 45, 45, 10];
[mean(data) quantile(data,0.25) quantile(data,0.50) quantile(data,0.75) std(data)]
```

```
ans = 1×5
   49.6000   43.0000   47.5000   52.0000   23.4767
```

```
newdata = data(find( data >= (quantile(data,0.25)-iqr(data))  & data <= (quantile(data,
```

```
newdata = 1×8
```

```
    43    37    50    51    58    52    45    45
```

```
%Make a conclusion about the effect of outliers on basic descriptive
%statistics. The std deviation is reduced significantly.
[mean(newdata) quantile(newdata,0.25) quantile(newdata,0.50) quantile(newdata,0.75) std
```

```
ans = 1×5
   47.6250   44.0000   47.5000   51.5000    6.4573
```

Qus.5 The following data set shows population of the United States (in million) since 1790,

Year 1790 1800 1810 1820 1830 1840 1850 1860 1870 1880 1890 1900

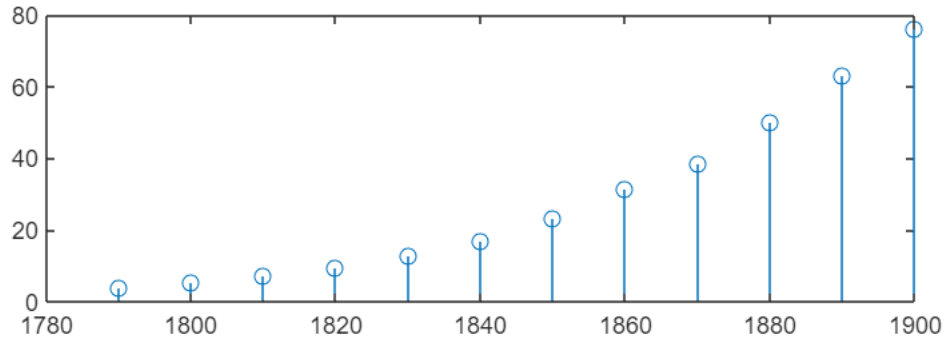
Population 3.9 5.3 7.2 9.6 12.9 17.1 23.2 31.4 38.6 50.2 63.0 76.2

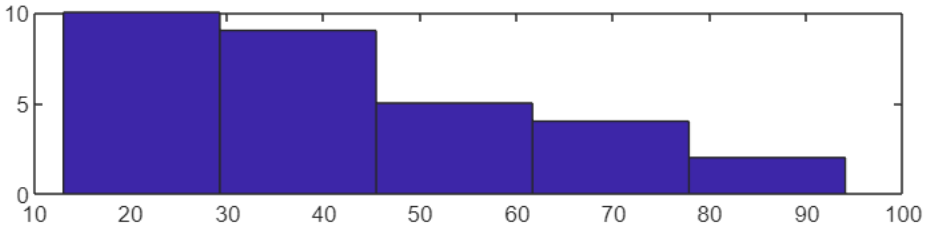
Year 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2010

Population 92.2 106.0 123.2 132.2 151.3 179.3 203.3 226.5 248.7 281.4 308.7

Construct a time plot for the U.S. population. What kind of trend do you see? What information can be extracted from this plot?

```
year1 = [1790 1800 1810 1820 1830 1840 1850 1860 1870 1880 1890 1900];
polulation1 = [3.9 5.3 7.2 9.6 12.9 17.1 23.2 31.4 38.6 50.2 63.0 76.2];

year2 = [1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2010];
polulation2 = [92.2 106.0 123.2 132.2 151.3 179.3 203.3 226.5 248.7 281.4 308.7];

subplot(211)
stem(year1,polulation1)
```






```
subplot(212)
stem(year2,polulation2)
```

Qus.6 Compute 10-year increments of the population growth x1 = 5.3−3.9, x2 = 7.2 − 5.3, etc

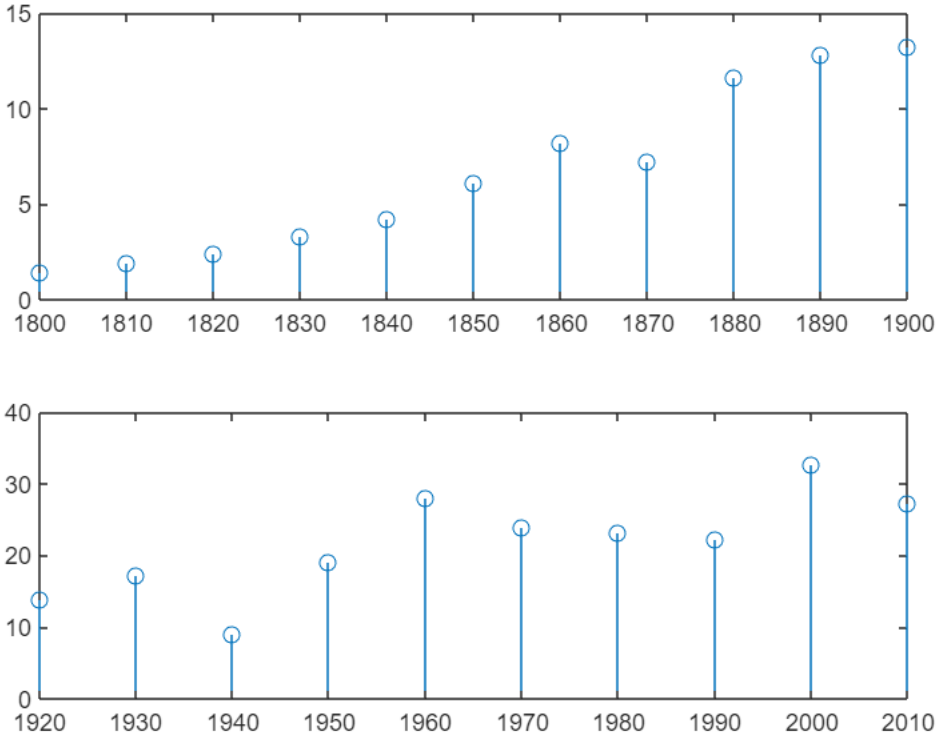
(a) Compute sample mean, median, and variance of 10-year increments. Discuss how the U.S. population changes during a decade.

(b) Construct a time plot of 10-year increments and discuss the observed pattern.

```
[mean(diff(polulation1)) median(diff(polulation1)) var(diff(polulation1)); mean(diff(po
```

```
ans = 2×3
    6.5727    6.1000   19.3582
   21.6500   22.7000   50.0583
```

```
subplot(211)
stem(year1(2:end),diff(polulation1))
subplot(212)
stem(year2(2:end),diff(polulation2))
```



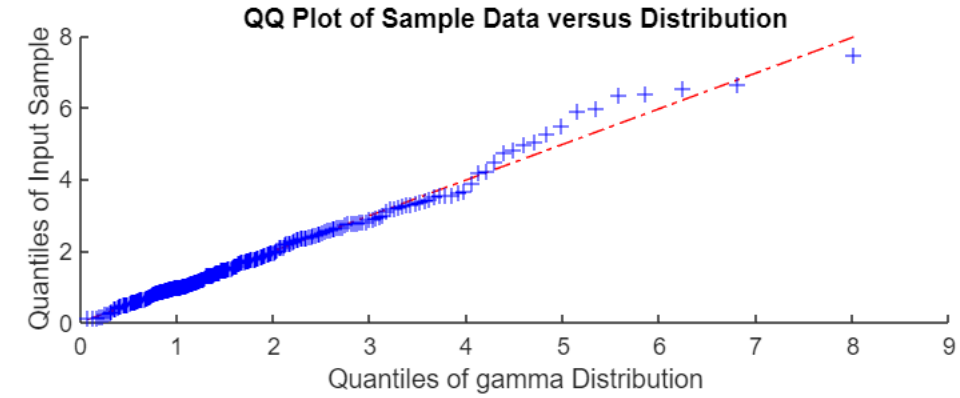
Qus.7 Do the following:

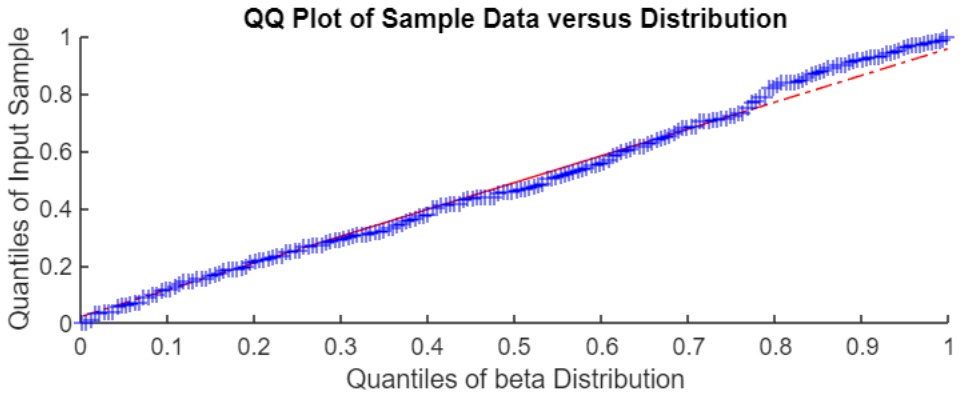
a) Generate 2 independent $N = 200$ samples X_1 and X_2 each from exponential random varaible with mean time $\frac{1}{\lambda} = 1$.

```
clear all
lambda=1; N = 200;
x1 = exprnd(1/lambda,[N 1]); x2 = exprnd(1/lambda,[N 1]);
```

b) Compute new samples $Y_1 = X_1 + X_2$, and $Y_2 = \frac{X_1}{X_1 + X_2}$ and visualize using quantile-quantile plot (qqplot) whether Y_1 and Y_4 to Gamma, and Beta distributions or not.

```
y1 = x1+x2; pd1 = fitdist(y1,'Gamma');
subplot(211); qqplot(y1,pd1)
y2 = x1./(x1+x2); pd2 = fitdist(y2,'Beta');
subplot(212); qqplot(y2,pd2)
```





Qus10a) Write a program that generate 200 samples of $Y = X_1 + X_2 + ...X_n$ where $X_1, X_2, ...X_n$ are independent exponential random variables with mean value $\frac{1}{\lambda} = \frac{1}{2}$, and n is not a constant but the 200 random samples, one for each Y. The distribution of n is geometric with the success probability $p = 0.05$

```
clear all
close all
lambda = 2; p = 0.05;
for i =1:200
    x(i,1) = sum(exprnd(1/lambda,[1,geornd(p)]));
end
```

Qus10b) Find the mean and standard deviation of Y. Are they coming out close to $\frac{1}{\lambda p}$?

```
[mean(x) sqrt(var(x)) 1/lambda/p]
```

```
ans = 1×3
    9.1484    9.2904   10.0000
```

Qus10c) Plot the histogram. Is the histogram is close to the exponential distribution? Do the following

(i) Find Maximum likelihood estimate using

[PointEstimate ConfidenceInterval]=mle(x,'distribution','exponential')

(ii) Fit the distribution to exponential without knowing the mean value

pd1 = fitdist(x,'exponential');

(iii) Make the distribution to exponential by assuming the mean value as $\frac{1}{\lambda p}$

pd2 = makedist('exponential','mu', 1/lambda/p);

(iv) Visualize whether pd1 and pd2 are close using histogram/pdf command and quantile-quantile plot (qqplot)

figure(1)

histogram(x,'normalization','pdf'); hold all;

ix = linspace(min(x),max(x));

iy1 = pdf(pd1,ix); iy2 = pdf(pd2,ix);

plot(ix,iy1,'r.'); plot(ix,iy2,'b-');

figure(2)

qqplot(x,pd1); hold all; qqplot(x,pd2);

figure(3)

qqplot(iy1,iy2);

```
[PointEstimate ConfidenceInterval]=mle(x,'distribution','exponential')
```

```
PointEstimate = 9.1484
ConfidenceInterval = 2×1
    8.0020
   10.5615
```

```
figure(1)
ix = linspace(min(x),max(x));
pd1 = fitdist(x,'exponential')
```

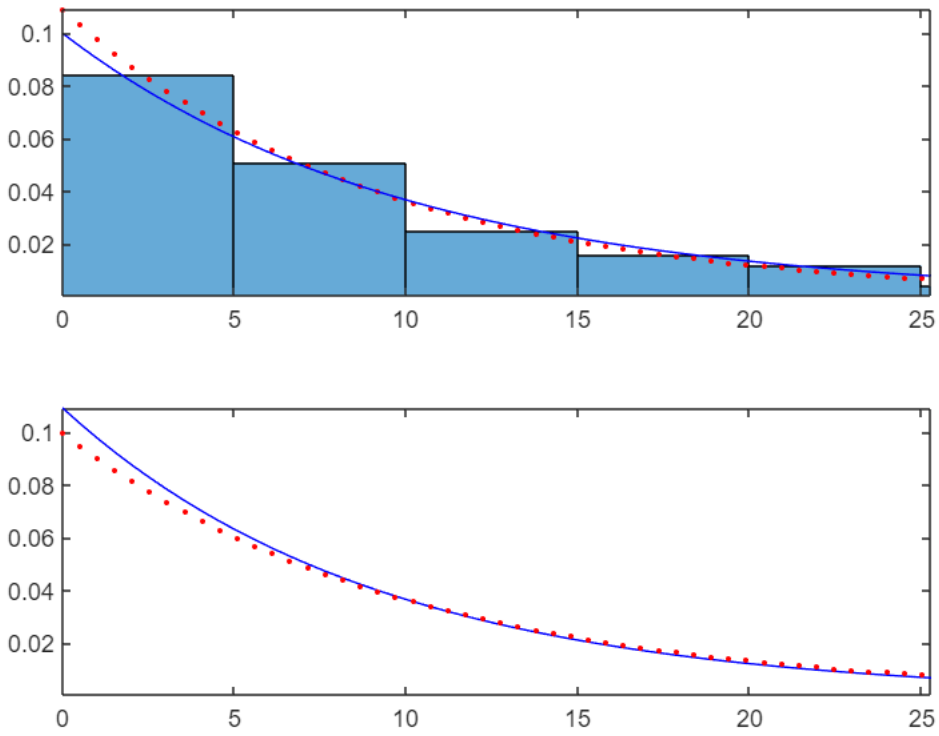
```
pd1 =
  ExponentialDistribution

  Exponential distribution
    mu = 9.1484   [8.00201, 10.5615]
```

```
iy1 = pdf(pd1,ix);
pd2 = makedist('exponential','mu', 1/lambda/p)
```

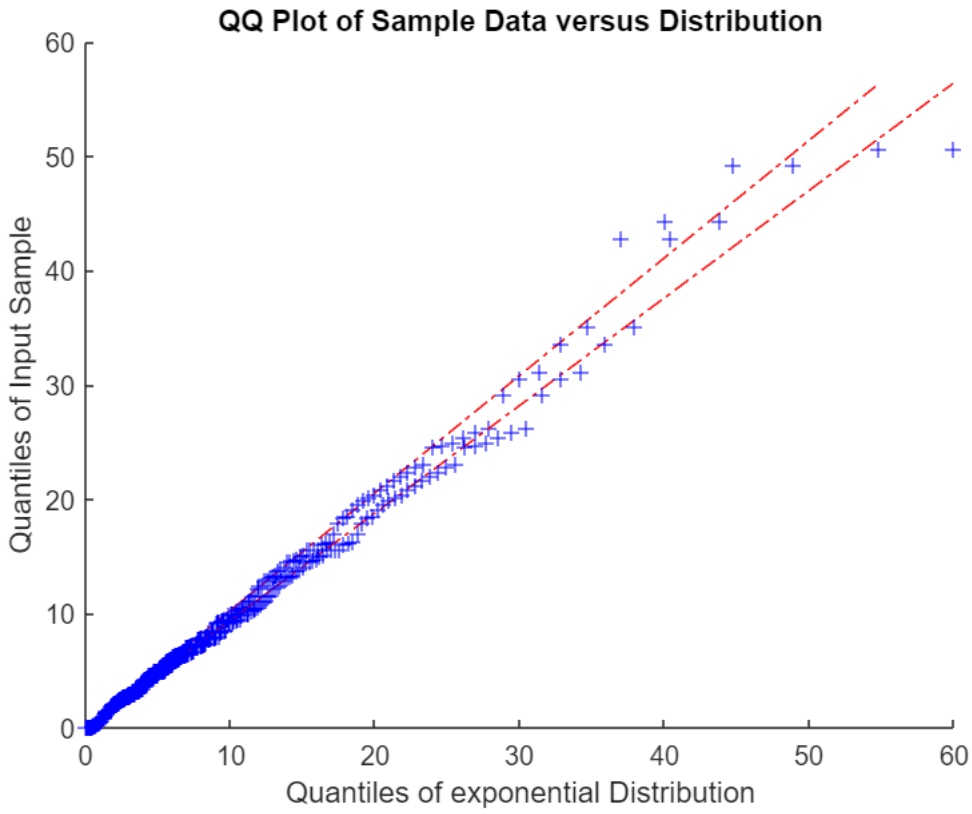
```
pd2 =
  ExponentialDistribution

  Exponential distribution
    mu = 10
```

```
iy2 = pdf(pd2,ix);

subplot(211)
```

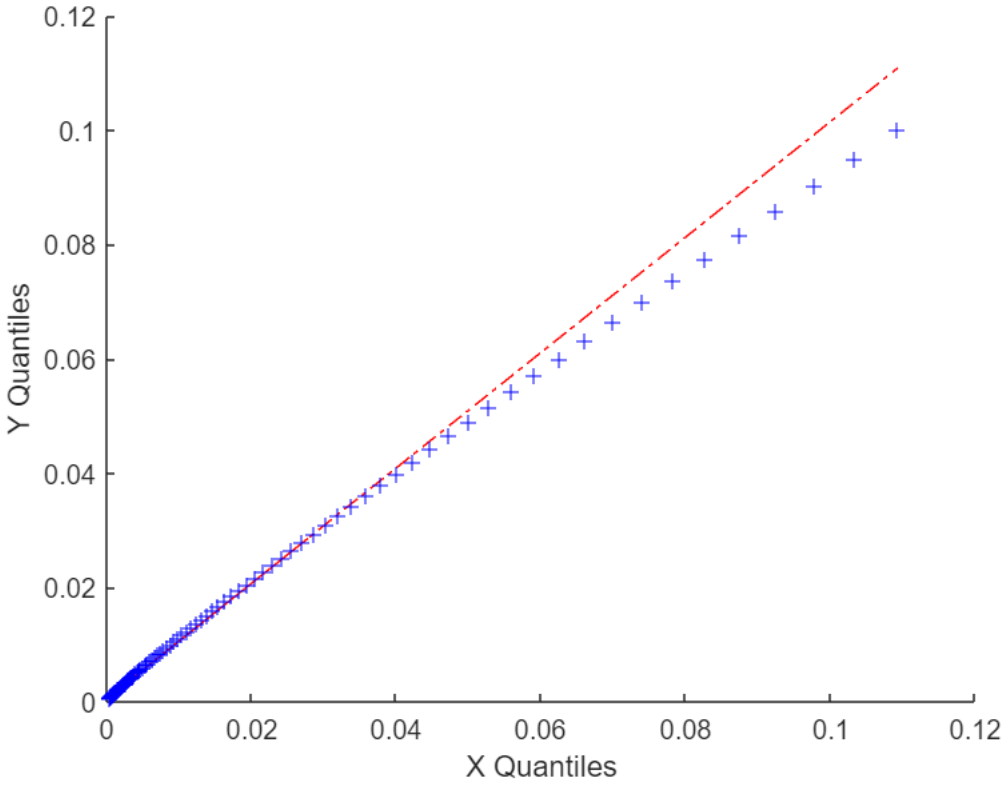
```
histogram(x,'normalization','pdf')
hold all
plot(ix,iy1,'r.');
plot(ix,iy2,'b-');
axis([min(ix) max(ix)/2 min(min(iy1),min(iy2)) max(max(iy1),max(iy2))])

subplot(212)
plot(ix,iy1,'b-');
hold all
plot(ix,iy2,'r.');
axis([min(ix) max(ix)/2 min(min(iy1),min(iy2)) max(max(iy1),max(iy2))])
```




```
figure(2)
qqplot(x,pd1)  % we use normplot for normal distribution. qqplot is for general distrib
hold all
qqplot(x,pd2)
```

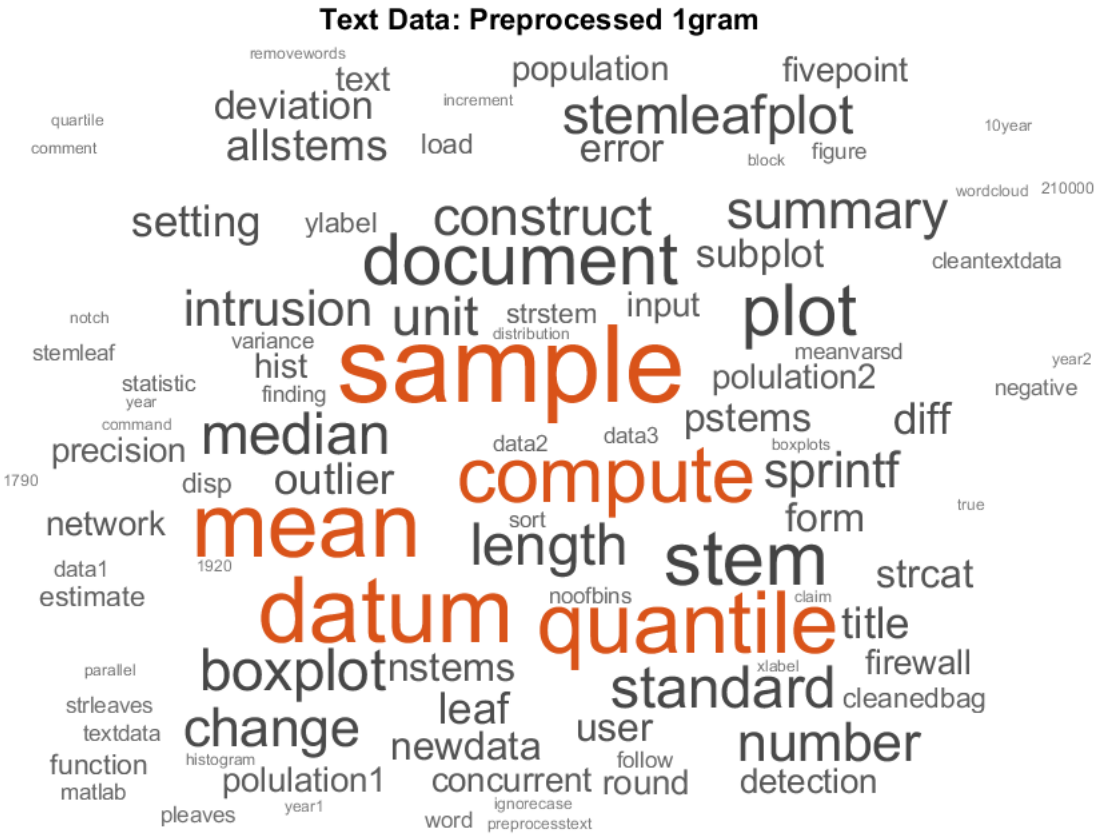



```
figure(3)
qqplot(iy1,iy2)
```

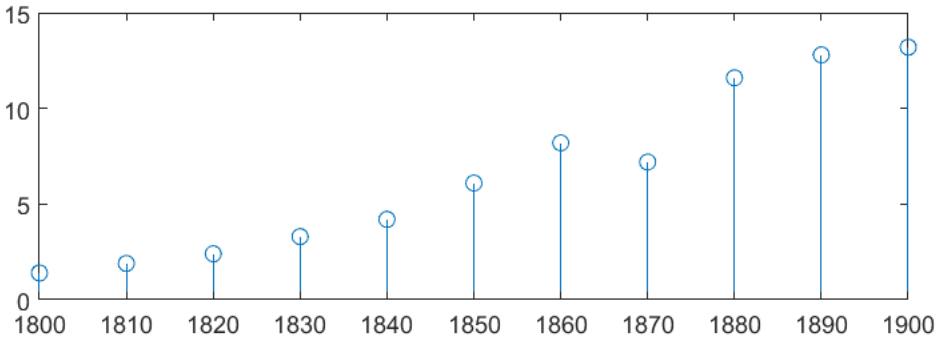


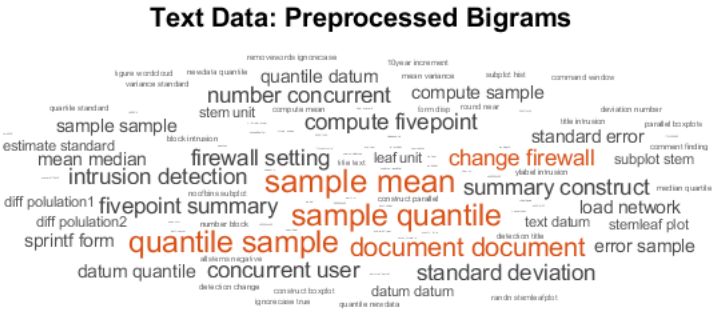
Processsing Tutorial7b_MA201_2021_12_15.pdf file using NGRAM, LDA (Latent Dirichlet Allocation) -

```
filename = "Tutorial7a_MA201_2021_12_01.pdf";
str = extractFileText(filename);
textData = split(str,newline);
cleanedDocuments = preprocessText(textData);
RemoveTheseWords = ["probability"];
cleanedDocuments = removeWords(cleanedDocuments,RemoveTheseWords, 'IgnoreCase',true);
cleanedBag = bagOfWords(cleanedDocuments);
[cleanedBag,idx] = removeEmptyDocuments(cleanedBag);
figure(1)
wordcloud(cleanedBag,'Shape','rectangle');
title("Text Data: Preprocessed 1gram")
```

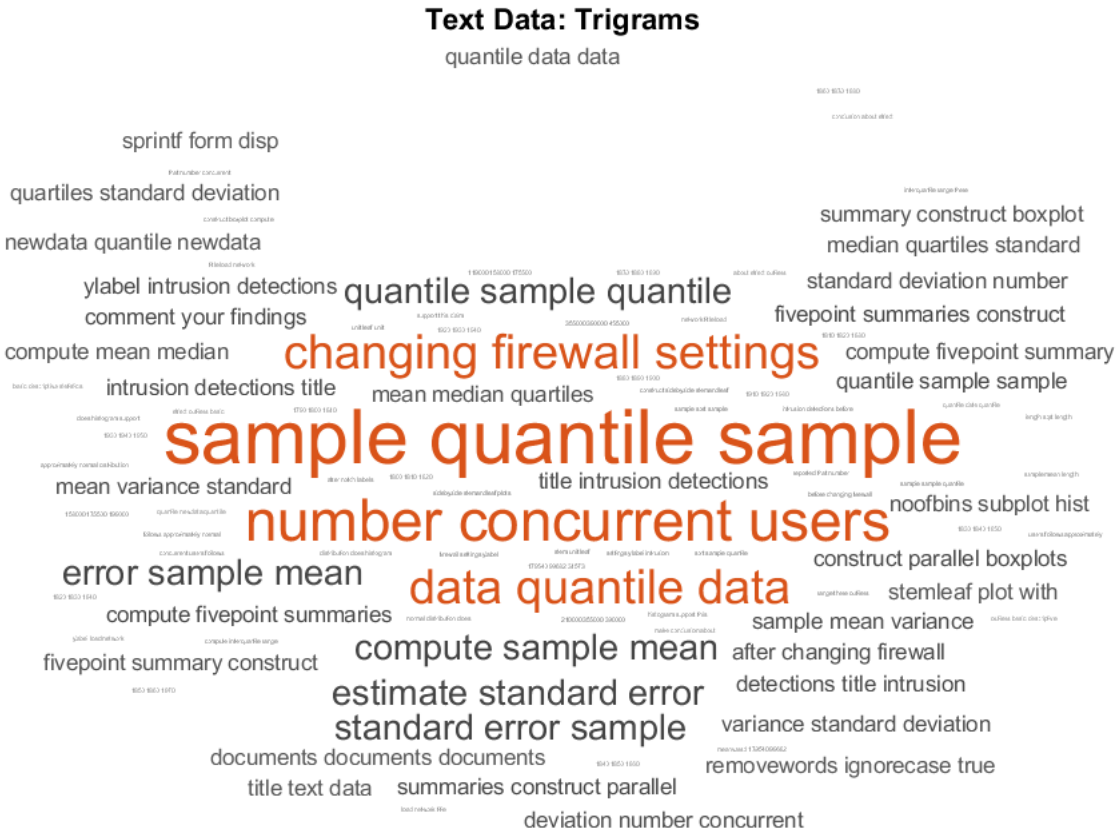



```
cleanedDocuments = preprocessText(str);
RemoveTheseWords = ["probability"];
cleanedDocuments = removeWords(cleanedDocuments,RemoveTheseWords, 'IgnoreCase',true);
bag = bagOfNgrams(cleanedDocuments);
cleanedBag = bagOfWords(cleanedDocuments);
figure(2)
wordcloud(bag);
title("Text Data: Preprocessed Bigrams")
```






```
cleanTextData = lower(str);
cleanedDocuments = tokenizedDocument(cleanTextData);
cleanedDocuments = erasePunctuation(cleanedDocuments);
%cleanedDocuments = removeStopWords(cleanedDocuments);
cleanedDocuments = removeShortWords(cleanedDocuments,3);
cleanedDocuments = removeLongWords(cleanedDocuments,15);
cleanedDocuments = addPartOfSpeechDetails(cleanedDocuments);
documents = normalizeWords(cleanedDocuments,'Style','lemma');
RemoveTheseWords = ["probability"];
cleanedDocuments = removeWords(cleanedDocuments,RemoveTheseWords, 'IgnoreCase',true);
bag = bagOfNgrams(cleanedDocuments,'NGramLengths',3);
figure(3)
wordcloud(bag);
title("Text Data: Trigrams")
```




```
function documents = preprocessText(textData)

% Convert the text data to lowercase.
cleanTextData = lower(textData);

% Tokenize the text.
documents = tokenizedDocument(cleanTextData);

% Erase punctuation.
documents = erasePunctuation(documents);

% Remove a list of stop words.
documents = removeStopWords(documents);

% Remove words with 2 or fewer characters, and words with 15 or greater
% characters.
documents = removeShortWords(documents,3);
documents = removeLongWords(documents,15);

% Lemmatize the words.
documents = addPartOfSpeechDetails(documents);
documents = normalizeWords(documents,'Style','lemma');

end

function stemleafplot(v,p)
% Plots stem and leaf plot to command window
```

```
%
% stemleafplot(v)
% stemleafplot(v,p)
%
% STEMLEAFPLOT plots stem-leaf plots of the input V to the command window.
% Leaf precision may be defined by the user. Note that inputs will be
% rounded to the nearest leaf unit (http://en.wikipedia.org/wiki/Stemplot).
%
% INPUT
%   V   : Array of numerical inputs (NaN values are ignored)
%
% OPTION
%   P   : Leaf precision (defined as integer power of 10)
%         Stem precision (by default) is 10^(P+1).
%         P is automatically rounded at the beginning of the function.
%         Leaf and stem units are printed at the bottom of the graph.
%         Examples: P = -3 rounds V to the nearest 10^-3 = 0.001
%                   P = 3 rounds V to the nearest 10^3 = 1000
%         [DEFAULT: P = 0]
%
% OUTPUT
%   Command window output
%
% EXAMPLES
%   % Stem-leaf plot of V with unit precision
%   V = 10.*randn(1,50);
%   stemleafplot(V)
%
%   % Stem-leaf plot of V with precision of 0.1
%   V = randn(1,50);
%   stemleafplot(V,-1)
%
%   % Stem-leaf plot of V with precision of 100
%   V = 5000.*randn(1,50);
%   stemleafplot(V,2)
%
% Jered Wells
% 01/28/2011
% jered [dot] wells [at] duke [dot] edu
%
% v1.2 (02/14/2012)
%
if ~isnumeric(v); error 'Input V must be numeric'; end
if ~exist('p','var'); p = 0; elseif isempty(p); p = 0; end
if ~isnumeric(p); error 'Input P must be an integer'; end
p = round(p);
% Condition V
v = v(~isnan(v));
v = v(:);
v = roundn(v,p);
% Organize stems and leaves
allstems = floor(v./10^(p+1));
allleaves = round(abs(v./10^p));
nstems = allstems(allstems<0)+1;     % Negative stems
```

```
nstems = nstems(:);
pstems = allstems(~(allstems<0));    % Positive stems
pstems = pstems(:);
nleaves = allleaves(allstems<0);    % Negative leaves
nleaves = nleaves(:);
pleaves = allleaves(~(allstems<0)); % Negative leaves
pleaves = pleaves(:);
dig = ceil(max(log10(abs(allstems))))+1;    % Max # of digits in stem
form = strcat(['%' num2str(dig+1) 'i']);    % Format string for SPRINTF
% Plot negative stems
if ~isempty(nstems)
    for ii = min(nstems(:)):0
        strstem = sprintf(form,ii);
        if ii==0; strstem(end-1:end) = '-0'; end
        strleaves = sprintf('%2i',mod(sort(nleaves(nstems==ii)),10));
        s = strcat([strstem ' |' strleaves]);
        disp(s)
    end % NSTEMS
end % IF
% Plot positive stems
if ~isempty(pstems)
    for ii = 0:max(pstems(:))
        strstem = sprintf(form,ii);
        strleaves = sprintf('%2i',mod(sort(pleaves(pstems==ii)),10));
        s = strcat([strstem ' |' strleaves]);
        disp(s)
    end % PSTEMS
end % IF
% Print out key and units
form = strcat(['%.' num2str(max(0,-p)) 'f']);
s = strcat(['key: 36|5 = ' sprintf(form,36*10^(p+1)+5*10^p)]);
disp(s)
s = strcat(['stem unit: ' sprintf(form,10^(p+1))]);
disp(s)
s = strcat(['leaf unit: ' sprintf(form,10^p)]);
disp(s)
end % MAIN
```

Mention Boxplot Outlier

```
function MentionBoxplotOutlier(c)
figure
e = eps(max(c(:)));
% x0=10;
% y0=10;
% width=1920;
% height=1440;
%set(gcf,'position',[x0,y0,width,height])
boxplot( c,'Notch','on')
h = flipud(findobj(gcf,'tag','Outliers')); % flip order of handles
for jj = 1 : length( h )
    x =  get( h(jj), 'XData' );
    y =  get( h(jj), 'YData' );
    for ii = 1 : length( x )
```

```
        if not( isnan( x(ii) ) )
            ix = find( abs( c(:,jj)-y(ii) ) < e );
            text( x(ii), y(ii), sprintf( '\\leftarrow%03s', string(c(ix))))
        end
    end
end
end
```

Qus1. Estimate the unknown parameter θ from a sample

3, 3, 3, 3, 3, 7, 7, 7

drawn from a discrete distribution with the probability mass function

P(3) = θ

P(7) = 1 − θ**.**

Compute two estimators of θ**:**

(a) the method of moments estimator;

(b) the maximum likelihood estimator. Also,

(c) Estimate the standard error of each estimator of θ**.**

(d) Run the matlab command to find the mle directly.

Hint - (a) Convert it to Bernoulli RV Y as $Y = \frac{-X+7}{4} \quad \Rightarrow X = -4Y + 7 \quad \Rightarrow E[X] = -4E[Y] + 7 = 7 - 4\theta$

The sample mean is $\overline{X} = \frac{9}{2}$. After Equating $7 - 4\hat{\theta}_{mm} = \frac{9}{2}$, we get $\hat{\theta}_{mm} = \frac{5}{8}$

```
clear all
```

```
Function definitions in a script must appear at the end of the file.
Move all statements after the "MentionBoxplotOutlier" function definition to before the first local
function definition.
```

```
X = [3 3 3 3 3 7 7 7]';
mean(X)
syms theta
solve(3*theta + 7*(1-theta)==mean(X),theta)
```

(b) The MLE for the discrete distribution is evaluated using mode statistics. Since mode(X) = 3 which occurs 5 times out of 8, therefore $\hat{\theta}_{mle} = \frac{5}{8}$

(c) Since both the estimates are same, the standard error for both the estimate will be the same. Moreover, $X = -4Y + 7$ where Y is a Bernoulli Distribution with parameter θ

$$std(\hat{\theta}) = \frac{s}{\sqrt{n}} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = \sqrt{\frac{5\times 3}{8\times 8\times 8}} = 0.1712$$

```
sqrt(5/8*3/8/8)
```

(d) Run the matlab command to find the mle directly.

```
[mlex mleci] = mle((-X+7)/4,'Distribution','Bernoulli')
```

or

```
fitdist((-X+7)/4,'Binomial')
```

Qus2. The number of times a computer code is executed until it runs without errors has a Geometric distribution with unknown parameter p. For 5 independent computer projects, a student records the following numbers of runs:

3 7 5 3 2

Estimate p

(a) by the method of moments;

(b) by the method of maximum likelihood.

(c) Run the matlab command to find the mle directly.

Hint: The pmf for the geometric distribution is given as $p_X(x) = (1-p)^{x-1}p$ for $x = 1, 2, 3...$

(a) $E[X] = \frac{1}{p}$, therefore $\frac{1}{\hat{p}_{mm}} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \Rightarrow \hat{p}_{mm} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{5}{20} = 0.25$

(b) For MLE, the Likelihood function can be written as follows:

$$\mathbf{p}_X(x) = \prod_{i=1}^{N} (1-p)^{x_i - 1} p = p^N \prod_{i=1}^{N} (1-p)^{x_i - 1}$$

$$\log \mathbf{p}_X(x) = N \log p + \log(1-p)\left[\sum_{i=1}^{N} x_i - N\right]$$

After differentiating wrt p, and substituing to 0, we get:

$$\frac{N}{p} = \frac{\left[\sum_{i=1}^{N} x_i - N\right]}{1-p} \quad \Rightarrow \hat{p}_{mle} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{5}{20} = 0.25$$

It is, therefore the same as moment method. Please note that the mle for exponential and geometric r.v. are the same.

(c) Run the matlab command to find the mle directly.

```
clear all
X = [3 7 5 3 2]';
[mlex mleci]=mle(X-1,'Distribution','Geometric')
```

Qus3. Use method of moments and method of maximum likelihood to estimate

(a) parameters a and b if a sample from Uniform(a, b) distribution is observed;

(b) parameter λ if a sample from Exponential(λ) distribution is observed;

(c) parameter μ if a sample from Normal(μ, σ) distribution is observed, and we already know σ;

(d) parameter σ if a sample from Normal(μ, σ) distribution is observed, and we already know μ;

(e) parameters μ and σ if a sample from Normal(μ, σ) distribution is observed, and both μ and σ are unknown.

Hint (a) For moment Method $\hat{X} = \frac{a+b}{2}, \quad \hat{\sigma} = \frac{a-b}{2\sqrt{3}} \quad \Rightarrow \hat{a} = \hat{X} - \hat{\sigma}\sqrt{3} \quad \hat{b} = \hat{X} + \hat{\sigma}\sqrt{3}$ where $\hat{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and $\hat{\sigma} = \frac{\sum_{i=1}^{n} (X_i - \hat{X})^2}{n-1}$

Similarly for MLE $\hat{a} = \min(X_i)$ and $\hat{b} = \max(X_i)$

DO NOT DERIVE THE FORMULAS. JUST KEEP A EYE ON THEM.

SNo	Method	$\widehat{a}$	$\widehat{b}$	$\widehat{\lambda}$	$\widehat{\mu}$	$\widehat{\sigma^2}$
(a)	$Uniform(a,b)/MM$	$\widehat{X}-\widehat{\sigma}\sqrt{3}$	$\widehat{X}+\widehat{\sigma}\sqrt{3}$	–	–	–
(a)	$Uniform(a,b)/MLE$	$\min(X_i)$	$\max(X_i)$	–	–	–
(b)	$Exponential(\lambda)/MM$	–	–	$\frac{n}{\sum_{i=1}^{n} X_i}$	–	–
(b)	$Exponential(\lambda)/MLE$	–	–	$\frac{n}{\sum_{i=1}^{n} X_i}$	–	–
(c)	$Normal(\mu,\sigma^2)(\sigma\,known)/MM$	–	–	–	$\frac{\sum_{i=1}^{n} X_i}{n}$	–
(c)	$Normal(\mu,\sigma^2)(\sigma\,known)/MLE$	–	–	–	$\frac{\sum_{i=1}^{n} X_i}{n}$	–
(d)	$Normal(\mu,\sigma^2)(\mu\,known)/MM$	–	–	–	–	$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}$
(d)	$Normal(\mu,\sigma^2)(\mu\,known)/MLE$	–	–	–	–	$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}$
(e)	$Normal(\mu,\sigma^2)(\mu,\sigma\,unknown)/MM$	–	–	–	$\frac{\sum_{i=1}^{n} X_i}{n}$	$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}$
(e)	$Normal(\mu,\sigma^2)(\mu,\sigma\,unknown)/MLE$	–	–	–	$\frac{\sum_{i=1}^{n} X_i}{n}$	$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n-1}$

Qus4. A sample of 3 observations (X1 = 0.4, X2 = 0.7, X3 = 0.9) is collected from a continuous distribution with density

$$f(x)=\begin{cases}\theta x^{\theta-1} & \text{if } 0<x<1\\ 0 & \text{otherwise}\end{cases}$$

Estimate θ by your favorite method. (Moment Method is easier).

Hint - Moment Method - $E[X]=1-\frac{1}{1+\theta}=\frac{\theta}{1+\theta}=\frac{0.4+0.7+0.9}{3}=\frac{2}{3}\Rightarrow\widehat{\theta}_{mm}=2$

```
clear all
syms theta positive
syms x
```

```
int(x*theta*x^(theta-1),x,0,1)
X = [0.4 0.7 0.9]';
mean(X)
%hence use solve function to find theta
solve(int(x*theta*x^(theta-1),x,0,1)==mean(X),theta)
```

Hint - Maximum Likelihood Estimate Method

After solving we will get $\hat{\theta}_{mle} = \frac{n}{\log\left(\prod_{i=1}^{n} X_i\right)} = 2.1765$

```
syms x1 x2 x3 positive
assume(x1<1 & x2<1 & x3<1)
f = @(x)(theta*x^(theta-1));
solv = solve(diff(f(x1)*f(x2)*f(x3),theta)==0,theta,ReturnConditions=true)
x1 = X(1); x2 = X(2); x3 = X(3);
vpa(subs(solv.theta))
```

Qus5. A sample (X1, ..., X10) is drawn from a distribution with a probability density function

$$\frac{1}{2}\left(\frac{1}{2}e^{-\frac{x}{\theta}} + \frac{1}{10}e^{-\frac{x}{10}}\right), 0 < x < \infty$$

The sum of all 10 observations equals 150.

(a) Estimate θ by the method of moments.

(b) Estimate the standard error of your estimator in (a).

Hint (a) - We write density $f_X(x)$ as follows:

$$f_X(x) = \frac{1}{2}\left(\frac{1}{2}e^{-\frac{x}{\theta}} + \frac{1}{10}e^{-\frac{x}{10}}\right) = \frac{1}{2}\left(\frac{\theta}{2}\times\frac{1}{\theta}e^{-\frac{x}{\theta}} + \frac{1}{10}e^{-\frac{x}{10}}\right) = \frac{1}{2}\left(\frac{\theta}{2}\times\exp\left(\lambda=\frac{1}{\theta}\right) + \exp\left(\lambda=\frac{1}{10}\right)\right)$$

Therefore $E[X] == \frac{1}{2}\left(\frac{\theta}{2}\times\theta + 10\right) = \frac{\theta^2}{4} + 5 = \frac{150}{10} = 15 \quad \Rightarrow \theta = \sqrt{40}$

```
clear all
syms theta x positive
n = 15;
Moment1stOrder = int(x/2*(1/2*exp(-x/theta)+1/10*exp(-x/10)),x,0,inf)
solve(int(x/2*(1/2*exp(-x/theta)+1/10*exp(-x/10)),x,0,inf)==15,theta)
```

(b) Since only sample mean statistics is provided in the question, we cannot find the sample standard deviation. Therefore, we find the variance of the pdf.

```
Moment2ndOrder = int(x^2/2*(1/2*exp(-x/theta)+1/10*exp(-x/10)),x,0,inf);
Variance = Moment2ndOrder - Moment1stOrder^2;
EstimateStderror = sqrt(Variance/n)
theta = sqrt(40);
```

```
vpa(subs(theta))
```

Qus6. Verify columns 3-5 in Table.

Null hypothesis	Parameter, estimator	If H_0 is true:		Test statistic
H_0	θ, $\hat{\theta}$	$\mathbf{E}(\hat{\theta})$	$\mathrm{Var}(\hat{\theta})$	$Z = \dfrac{\hat{\theta} - \theta_0}{\sqrt{\mathrm{Var}(\hat{\theta})}}$
One-sample Z-tests for means and proportions, based on a sample of size n				
$\mu = \mu_0$	μ, $\bar{X}$	μ_0	$\dfrac{\sigma^2}{n}$	$\dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
$p = p_0$	p, $\hat{p}$	p_0	$\dfrac{p_0(1-p_0)}{n}$	$\dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
Two-sample Z-tests comparing means and proportions of two populations, based on independent samples of size n and m				
$\mu_X - \mu_Y = D$	$\mu_X - \mu_Y$, $\bar{X} - \bar{Y}$	D	$\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{m}$	$\dfrac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$
$p_1 - p_2 = D$	$p_1 - p_2$, $\hat{p}_1 - \hat{p}_2$	D	$\dfrac{p_1(1-p_1)}{n} + \dfrac{p_2(1-p_2)}{m}$	$\dfrac{\hat{p}_1 - \hat{p}_2 - D}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}}$
$p_1 = p_2$	$p_1 - p_2$, $\hat{p}_1 - \hat{p}_2$	0	$p(1-p)\left(\dfrac{1}{n} + \dfrac{1}{m}\right)$, where $p = p_1 = p_2$	$\dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}}$ where $\hat{p} = \dfrac{n\hat{p}_1 + m\hat{p}_2}{n+m}$

Qus7. In order to ensure efficient usage of a server, it is necessary to estimate the mean number of concurrent users. According to records, the average number of concurrent users at 100 randomly selected times is 37.7, with a standard deviation σ = 9.2.

(a) Construct a 90% confidence interval for the expectation of the number of concurrent users.

(b) At the 1% significance level, do these data provide significant evidence that the mean number of concurrent users is greater than 35?

(c) What is the P-Value?

(a) $\left[\hat{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \quad \hat{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right] = [37.7 - z_{0.05}0.92, \quad 37.7 + z_{0.05}0.92]$

```
clear all
muhat = 37.7; sigma = 9.2; n = 100;
muhat+ norminv([0.05 0.95])*sigma/sqrt(n);
```

or

```
ci = .9; %confidence interval (two sided)
alpha = 1 -ci;
muhat+ norminv([alpha/2 1-alpha/2])*sigma/sqrt(n)
```

(b) We will conduct one-sided Z test:

(i) TEST STATISTICS -

```
testvalue = 35;
Z = (muhat- testvalue)/(sigma/sqrt(n))
```

(ii) Generate Hypothesis:

$H_0 : \mu = 35$

$H_A : \mu > 35$

```
alpha = 0.01;
z_alpha = norminv(1-alpha) %-norminv(alpha)
h = Z>z_alpha %reject the null hypothesis if true.
```

Reject null hypothesis even with 1% level of significance.

(iii) P_Value = P(Z>ZObs) = P(Z>2.9348)

```
P_Value = normcdf(Z,0,1,"Upper")
```

Strong evidence against null hypothesis.

Qus8. Installation of a certain hardware takes random time with a standard deviation of 5 minutes.

(a) A computer technician installs this hardware on 64 different computers, with the average installation time of 42 minutes. Compute a 95% confidence interval for the population mean installation time.

(b) Suppose that the population mean installation time is 40 minutes. A technician installs the hardware on your PC. What is the probability that the installation time will be within the interval computed in (a)?

(c) A manager questions the assumptions. Her pilot sample of 40 installation times has a sample standard deviation of s = 6.2 min, and she says that it is significantly different from the assumed value of σ = 5 min. Do you agree with the manager? Conduct the suitable test of a standard deviation. (Qus20) (Chi-square)

```
clear all
sigma = 5; muhat = 42; n =64;
ci = .95; %confidence interval (two sided)
alpha = 1-ci;
muhat+ norminv([alpha/2 1-alpha/2])*sigma/sqrt(n)
mu = 40;
zinterval= (muhat + norminv([alpha/2 1-alpha/2])*sigma/sqrt(n) -mu)/sigma
P = normcdf(zinterval(2)) - normcdf(zinterval(1))
```

$H_0 : \sigma^2 = \sigma_0^2$

$H_A : \sigma^2 > \sigma_0^2$

compute the χ^2 – statistc.

$\chi^2_{obs} = \frac{(n-1)s^2}{\sigma_0^2}$

```
n = 40; s =6.2;
chi2obs = (n-1)*s^2/sigma^2
P_Value = chi2cdf(chi2obs,n-1,"upper")
```

Low P_Value (Null hypothesis can only be accepted at 1% level of significance or lower. Sufficient evidence for H_A).

Qus9. Salaries of entry-level computer engineers have Normal distribution with unknown mean and variance. Three randomly selected computer engineers have salaries (in $ 1000s): 30, 50, 70

(a) Construct a 90% confidence interval for the average salary of an entry-level computer engineer.

(b) Does this sample provide a significant evidence, at a 10% level of significance, that the average salary of all entry-level computer engineers is different from $80,000? Explain.

(c) Looking at this sample, one may think that the starting salaries have a great deal of variability. Construct a 90% confidence interval for the standard deviation of entrylevel salaries.

```
clear all
x = [30 50 70]';
n = length(x);
muhat = mean(x);
stdhat = std(x);
ci = 0.9;
alpha = 1 - ci;

ConfInterval = muhat+ tinv([alpha/2 1-alpha/2],n-1)*stdhat/sqrt(n)
mu = 80;
T = (muhat - mu)/(stdhat/sqrt(n)) % T statistics.
AcceptanceRegion = tinv([alpha/2 1-alpha/2],n-1)
% Null hypothesis is not rejected. No significant evidence for salary
```

```
% different from $80000.


Chi2OneMinusAlphaBy2 = min(chi2inv((1-alpha/2),n-1),chi2inv(alpha/2,n-1));
Chi2AlphaBy2 = max(chi2inv((1-alpha/2),n-1),chi2inv(alpha/2,n-1));
CIstd = sqrt([std(x)^2*(n-1)/Chi2AlphaBy2  std(x)^2*(n-1)/Chi2OneMinusAlphaBy2])
```

Qus10. We have to accept or reject a large shipment of items. For quality control purposes, we collect a sample of 200 items and find 24 defective items in it.

(a) Construct a 96% confidence interval for the proportion of defective items in the whole shipment.

(b) The manufacturer claims that at most one in 10 items in the shipment is defective. At the 4% level of significance, do we have sufficient evidence to disprove this claim? Do we have it at the 15% level?

```
clear all
n1=200;
p1hat= 24/n1;
ci =0.96; alpha = 1-ci;
p1hat + norminv([alpha/2 1-alpha/2])*sqrt(p1hat*(1-p1hat)/n1)
z_obs = (1/10 - p1hat)/sqrt(p1hat*(1-p1hat)/n1)
P_Value = normcdf(z_obs)
norminv([0.04 0.15 P_Value 0.35])
```

Qus11. Refer to Exercise 9.10. Having looked at the collected sample, we consider an alternative supplier. A sample of 150 items produced by the new supplier contains 13 defective items. Is there significant evidence that the quality of items produced by the new supplier is higher than the quality of items in Exercise 9.10? What is the P-value?

```
n2 = 150;
p2hat= 13/n2;
ci =0.96; alpha = 1-ci;
Zobs = (p1hat-p2hat-0)/sqrt(p1hat*(1-p1hat)/n1 + p2hat*(1-p2hat)/n2) % H0: p2=p1 HA: p2
normcdf(Zobs,"upper")
```

Qus12. An electronic parts factory produces resistors. Statistical analysis of the output suggests that resistances follow an approximately Normal distribution with a standard deviation of 0.2 ohms. A sample of 52 resistors has the average resistance of 0.62 ohms.

(a) Based on these data, construct a 95% confidence interval for the population mean resistance.

(b) If the actual population mean resistance is exactly 0.6 ohms, what is the probability that an average of 52 resistances is 0.62 ohms or higher?

Qus16. A sample of 250 items from lot A contains 10 defective items, and a sample of 300 items from lot B is found to contain 18 defective items.

(a) Construct a 98% confidence interval for the difference of proportions of defective items.

(b) At a significance level α = 0.02, is there a significant difference between the quality of the two lots?

```
clear all
n1=250; p1hat= 10/n1;
n2=300; p2hat= 18/n2;
ci = 0.98; alpha = 1-ci;
p1hat - p2hat + norminv([alpha/2 1-alpha/2])*sqrt(p1hat*(1-p1hat)/n1 + p2hat*(1-p2hat)/
```

This interval contains 0, hence no significance difference between the quality of two lots at 2% level of significance.

Qus17. A news agency publishes results of a recent poll. It reports that candidate A leads candidate B by 10% because 45% of the poll participants supported Ms. A whereas only 35% supported Mr. B. What margin of error should be reported for each of the listed estimates, 10%, 35%, and 45%? Notice that 900 people participated in the poll, and the reported margins of error typically correspond to 95% confidence intervals.

```
clear all
n=900; ci =.95; alpha = 1- ci;
p1hat = 0.45;
norminv(1-alpha/2)*sqrt(p1hat*(1-p1hat)/n)*100
p2hat = 0.35;
norminv(1-alpha/2)*sqrt(p2hat*(1-p2hat)/n)*100
norminv(1-alpha/2)*sqrt(p1hat*(1-p1hat)/n+p2hat*(1-p2hat)/n)*100
```

Qus23. Anthony says to Eric that he is a stronger student because his average grade for the first six quizzes is higher. However, Eric replies that he is more stable because the variance of his grades is lower. The actual scores of the two friends (presumably, independent and normally distributed) are in the table.

	Quiz1	Quiz2	Quiz3	Quiz4	Quiz5	Quiz6
Anthony	85	92	97	65	75	96
Eric	81	79	76	84	83	77

(a) Is there significant evidence to support Anthony's claim? State H_0 and H_A. Test equality of variances and choose a suitable two-sample t-test. Then conduct the test and state conclusions.

(b) Is there significant evidence to support Eric's claim? State H_0 and H_A and conduct the test. For each test, use the 5% level of significance.

```
clear all
x = [85 92 97 65 75 96]';
y = [81 79 76 84 83 77]';
n = length(x); m = length(y);
muxhat = mean(x); stdxhat = std(x);
muyhat = mean(y); stdyhat = std(y);
[muxhat muyhat]
[stdxhat stdyhat]
% (a) H0: mu_x = mu_y % HA: mu_x > mu_y
```

```
v = (stdxhat^2/n+stdyhat^2/m)^2/((stdxhat^2/n)^2/(n-1)+(stdyhat^2/m)^2/(m-1));
T = (muxhat - muyhat)/sqrt(stdxhat^2/n+stdyhat^2/m)
alpha = 0.05;
tinv([1- alpha],v)
h = T>tinv([1- alpha],v) % reject null hypothesis if true
P_Value = tcdf(T,v,"Upper")
% (b) H0: sigma_x^2 = sigma_y^2 % HA: sigma_x^2 > sigma_y^2
EstVarRat = stdxhat^2/stdyhat^2;
PopuVarRat = 1;
F = EstVarRat/PopuVarRat;
alpha =0.05;
fcdf(1 - alpha,n-1,m-1,"Upper")
h = F>fcdf(1 - alpha,n-1,m-1,"Upper") % reject null hypothesis if true
FObs = F;
PValue = fcdf(FObs,n-1,m-1,"Upper") %significant evidence at 1% to reject the null hypo
```

Qus24. Recall Exercise 2s3. Results essentially show that a sample of six quizzes was too small for Anthony to claim that he is a stronger student. We realize that each student has his own population of grades, with his own mean μ_i and variance σ_i^2 . The observed quiz grades are sampled from this population, and they are different due to all the uncertainty and random factors involved when taking a quiz. Let us estimate the population parameters with some confidence.

(a) Construct a 90% confidence interval for the population mean score for each student.

(b) Construct a 90% confidence interval for the difference of population means. If you have not completed Exercise 23(a), start by testing equality of variances and choosing the appropriate method.

(c) Construct a 90% confidence interval for the population variance of scores for each student.

(d) Construct a 90% confidence interval for the ratio of population variances.

```
ci = .9; alpha = 1- ci;
%(a) T distribution
ConfIntervalmuX = muxhat + tinv([alpha/2 1-alpha/2],n+m-2)*stdxhat/sqrt(n)
ConfIntervalmuY = muyhat + tinv([alpha/2 1-alpha/2],n+m-2)*stdyhat/sqrt(m)
% (b)Test from exercise 9.23 asserts inequality of variance $sigmaxhat != sigmayhat$
ConfIntervalDiffmeanXY = muxhat - muyhat + tinv([alpha/2 1-alpha/2],v)*sqrt(stdxhat^2/n
% (c) Chi2 Distribution
ConfIntervalVarX = stdxhat^2*(n-1) ./ chi2inv([1- alpha/2 alpha/2],n-1)
ConfIntervalVarY = stdyhat^2*(m-1) ./ chi2inv([1- alpha/2 alpha/2],m-1)

% (d) F Distribution
ConfIntervalRatioVarXY = stdxhat^2/stdyhat^2 * [1/finv(1- alpha/2,n-1,m-1) finv(1- alph
```

Processsing Tutorial7c_MA201_2021_12_25.pdf file using NGRAM, LDA (Latent Dirichlet Allocation) -

```
filename = "Tutorial7c_MA201_2021_12_25.pdf";
str = extractFileText(filename);
textData = split(str,newline);
cleanedDocuments = preprocessText(textData);
RemoveTheseWords = ["probability"];
cleanedDocuments = removeWords(cleanedDocuments,RemoveTheseWords, 'IgnoreCase',true);
```

```
cleanedBag = bagOfWords(cleanedDocuments);
[cleanedBag,idx] = removeEmptyDocuments(cleanedBag);
figure(1)
wordcloud(cleanedBag,'Shape','rectangle');
title("Text Data: Preprocessed 1gram")
cleanedDocuments = preprocessText(str);
RemoveTheseWords = ["probability"];
cleanedDocuments = removeWords(cleanedDocuments,RemoveTheseWords, 'IgnoreCase',true);
bag = bagOfNgrams(cleanedDocuments);
cleanedBag = bagOfWords(cleanedDocuments);
figure(2)
wordcloud(bag);
title("Text Data: Preprocessed Bigrams")
cleanTextData = lower(str);
cleanedDocuments = tokenizedDocument(cleanTextData);
cleanedDocuments = erasePunctuation(cleanedDocuments);
%cleanedDocuments = removeStopWords(cleanedDocuments);
cleanedDocuments = removeShortWords(cleanedDocuments,3);
cleanedDocuments = removeLongWords(cleanedDocuments,15);
cleanedDocuments = addPartOfSpeechDetails(cleanedDocuments);
documents = normalizeWords(cleanedDocuments,'Style','lemma');
RemoveTheseWords = ["probability"];
cleanedDocuments = removeWords(cleanedDocuments,RemoveTheseWords, 'IgnoreCase',true);
bag = bagOfNgrams(cleanedDocuments,'NGramLengths',3);
figure(3)
wordcloud(bag);
title("Text Data: Trigrams")
```

```
function documents = preprocessText(textData)

% Convert the text data to lowercase.
cleanTextData = lower(textData);

% Tokenize the text.
documents = tokenizedDocument(cleanTextData);

% Erase punctuation.
documents = erasePunctuation(documents);

% Remove a list of stop words.
documents = removeStopWords(documents);

% Remove words with 2 or fewer characters, and words with 15 or greater
% characters.
documents = removeShortWords(documents,3);
documents = removeLongWords(documents,15);

% Lemmatize the words.
documents = addPartOfSpeechDetails(documents);
documents = normalizeWords(documents,'Style','lemma');

end
```