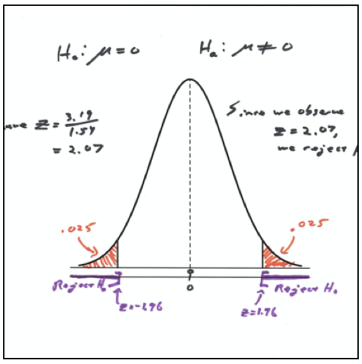
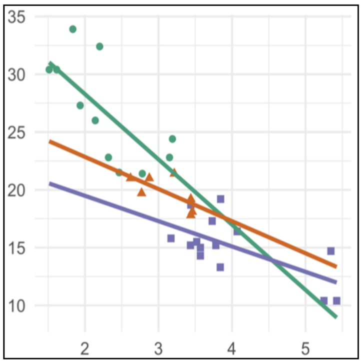
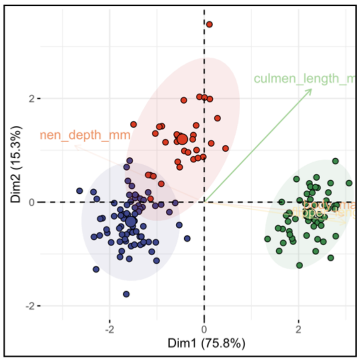
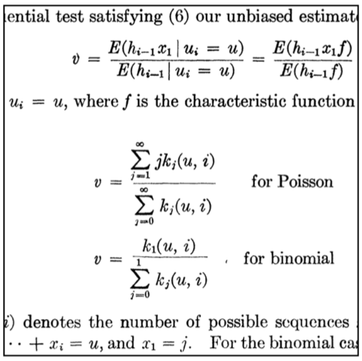








ential test satisfying (6) our unbiased estimat

$$v = \frac{E(h_{i-1}x_1 \mid u_i = u)}{E(h_{i-1} \mid u_i = u)} = \frac{E(h_{i-1}x_1 f)}{E(h_{i-1} f)}$$

$u_i = u$, where f is the characteristic function

$$v = \frac{\sum_{j=1}^{\infty} j k_j(u, i)}{\sum_{j=0}^{\infty} k_j(u, i)} \quad \text{for Poisson}$$

$$v = \frac{k_1(u, i)}{\sum_{j=0}^{1} k_j(u, i)} \quad \text{for binomial}$$

i) denotes the number of possible sequences
$\cdots + x_i = u$, and $x_1 = j$. For the binomial ca

2023-2 Astronomical Statistical Analysis Method Homework Report

2023-2_AstroStatistics_HW_MY.docx (ver. 5)

23/09/29 MY

REVISION LOG

Ver 1. 23.09.08 Make document form and sections. HW1 Problem A in progress.
Ver 2. 23.09.14 HW1 Problem A in progress.
Ver 3. 23.09.17 HW1 Details in progress.
Ver 4. 23.09.18 HW1 Complete.
Ver 5. 23.09.29 HW2 Complete.

Table of Contents

1. Introduction

This report was written for an assignment for Professor Arman Shafieloo's Astronomical Statistical Analysis Method class held in the second semester of 2023. These assignments were written in the Python language, and the codes used in the exercises can be found in the Github repository (https://github.com/mmingyeong/2023-2_AstroStatistics_HW.git).

1.1 Reference Documents

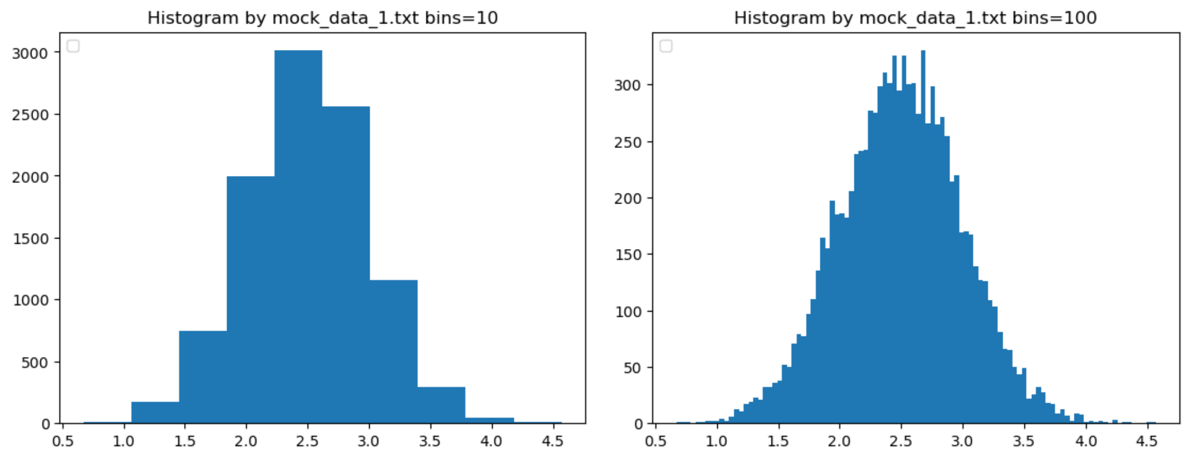
2. Homework 1

Reports by Wednesday 20th September, 10 AM

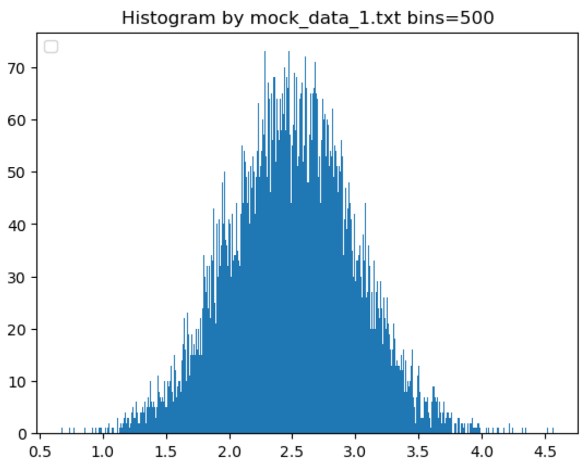
2.1 Problem A

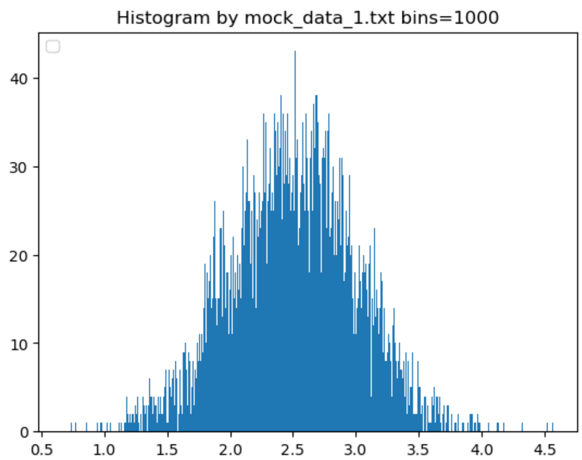
a. For the 4 data sets of mock_data_1.txt, mock_data_2.txt, mock_data_3.txt, mock_data_4.txt (10000 values in each set) attached to this email plot the binned data. Try to choose a reasonable bin size. What is your visual interpretation?

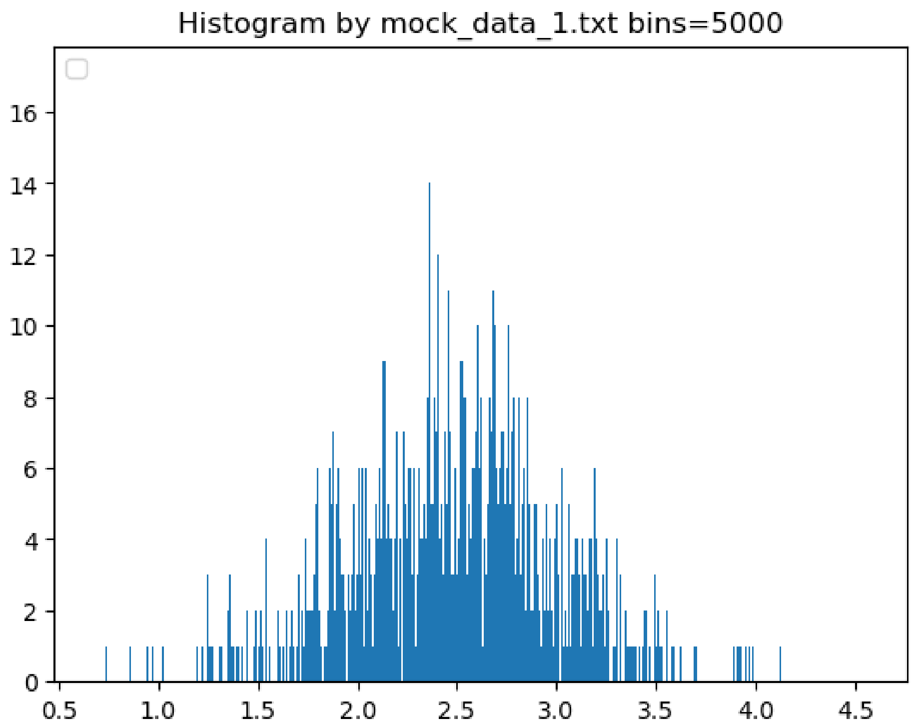
First, the minimum data value of mock_data_1.txt is 0.6734531856690171 and the maximum is 4.568540601217464. Therefore, the data range is approximately 0.67 < x < 4.57. The data was divided into binsizes of 10, 100, 500, 1000, and 5000, respectively, and a histogram was drawn according to these binsizes.



When a histogram is drawn with a binsize of 10 or 100, a bell-shaped distribution is visible to some extent, but because the binsize is too small, the data distribution appears lumpy rather than detailed, making it difficult to capture the features of 10,000 data.

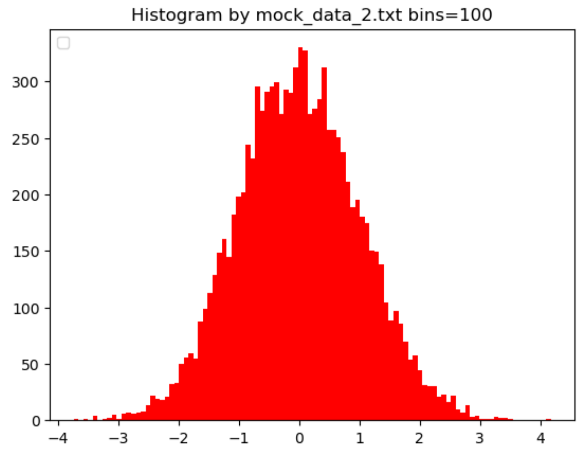


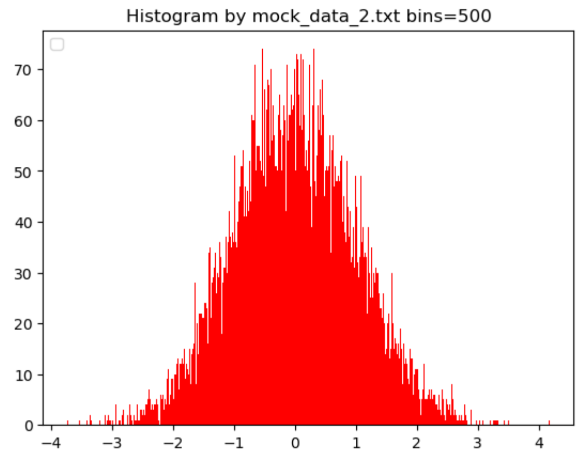


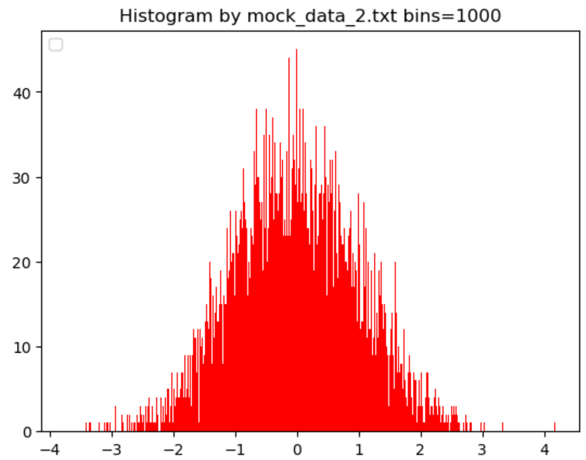


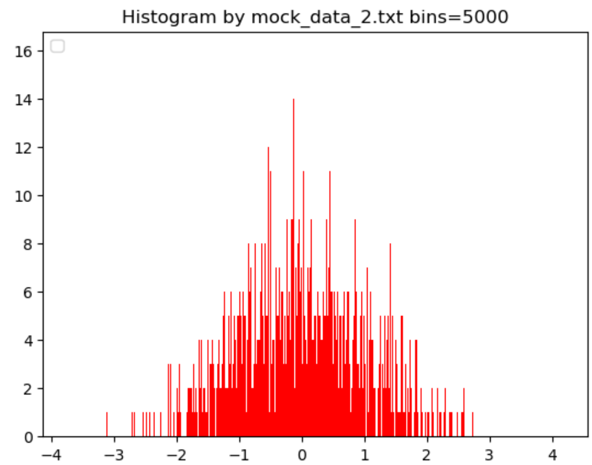
Also, the bell-shaped distribution is clearly observed in binsize=5000, but it does not seem necessary to do this in such detail. This is because as the binsize increases, computational efficiency decreases. Thus, binsize=500 seems most appropriate. It can also be seen that the distribution is somewhat symmetrical.

This is a histogram of mock_data_2.tx data. The minimum data value of mock_data_2.txt is -3.735476599796913 and the maximum is 4.17922911127687. Therefore, the data range is approximately - 3.74 < x < 4.18. Likewise, binsize=500 seems to be the most appropriate and it can be confirmed that it achieves a symmetrical distribution.

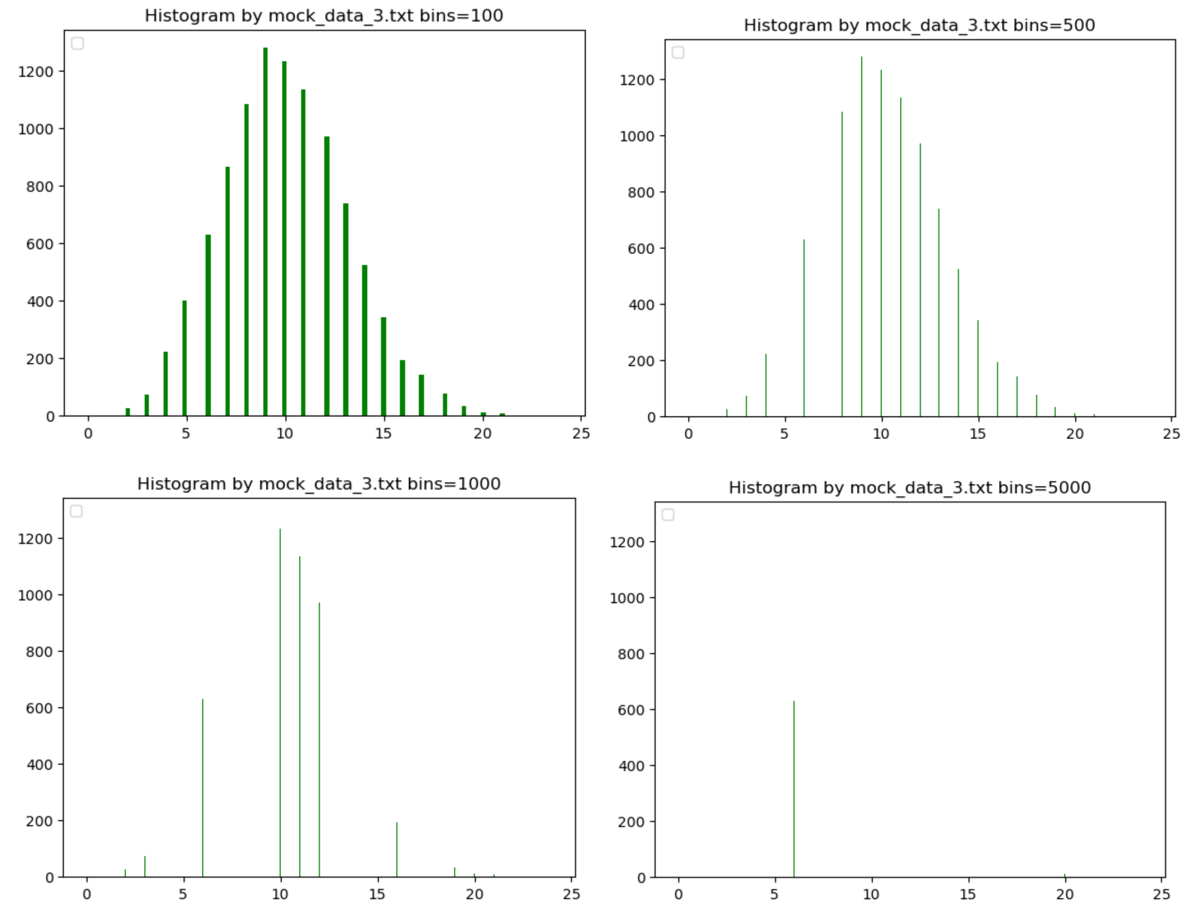




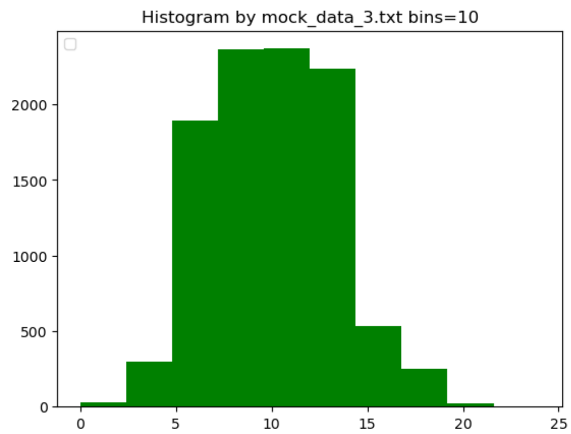


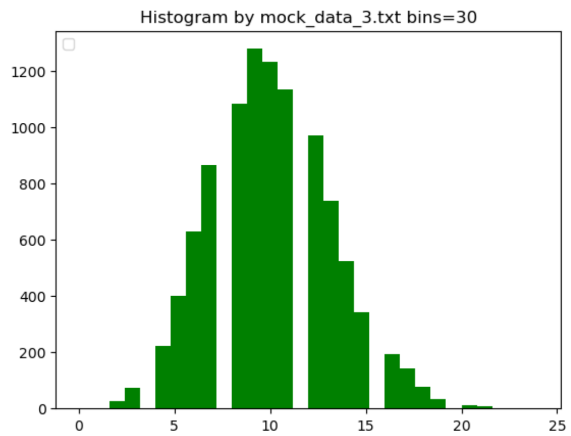


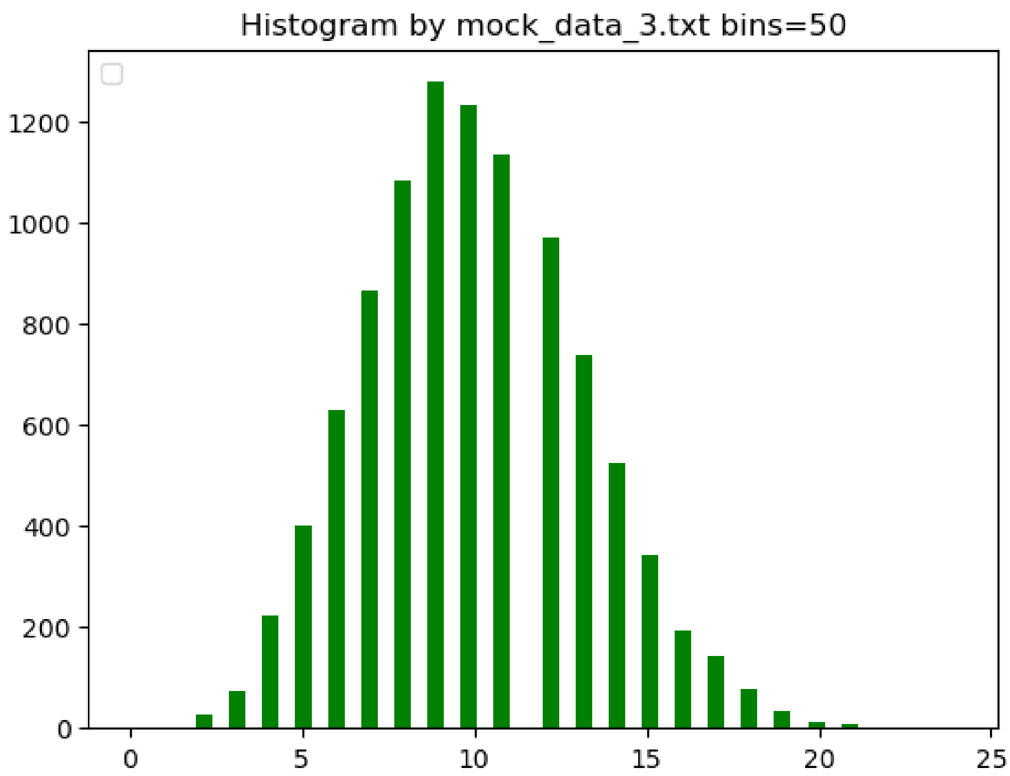
This is a histogram of mock_data_3.tx data. The minimum data value of mock_data_3.txt is 0.0 and the maximum is 24.0. Therefore, the data range is approximately 0 < x < 24. However, when mock_data_3.txt data was set to a high binsize like the above datasets, it was difficult to identify data features and distribution.



When the binsize of the mock_data_3.txt data was set very small, the data distribution could be seen more properly.

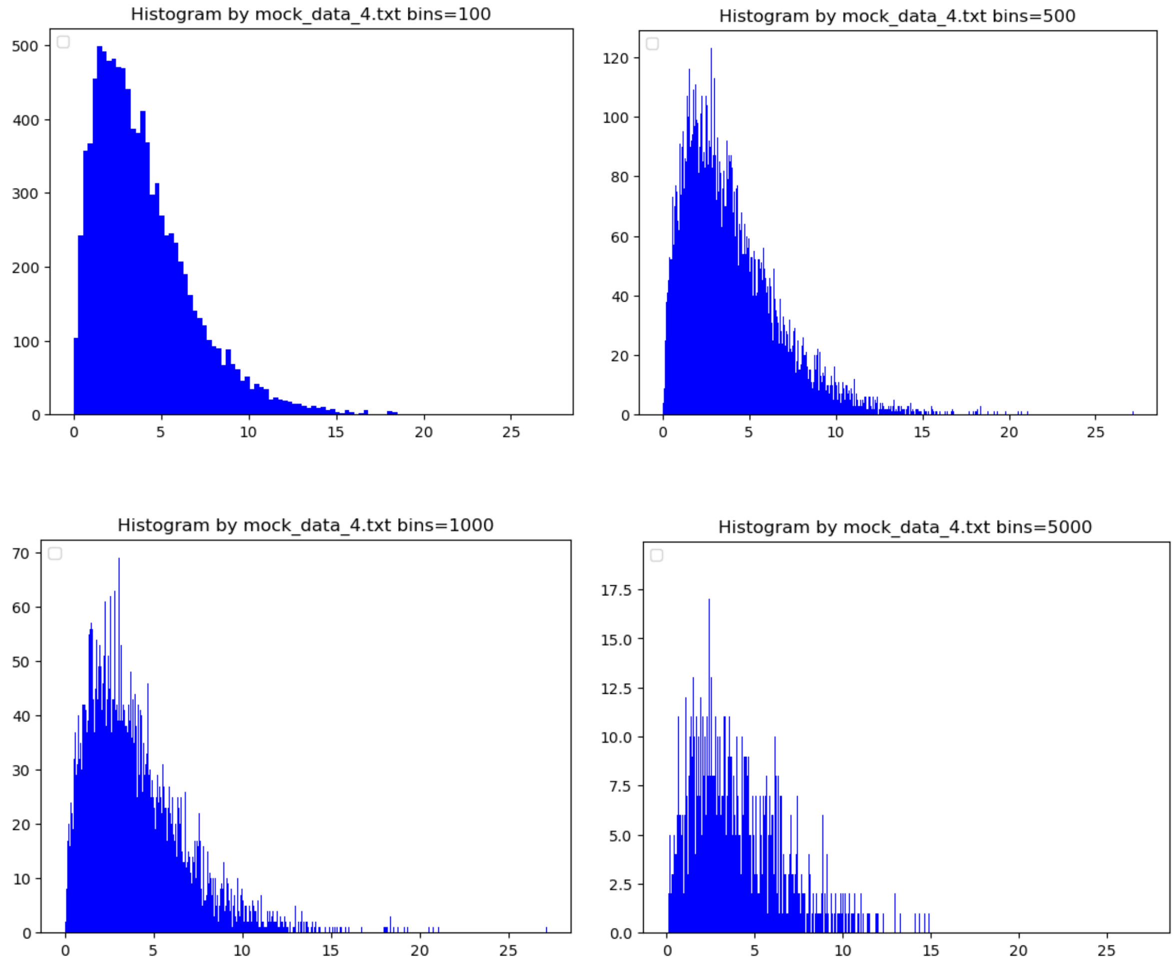






Therefore, binsize=50 seems most appropriate for mock_data_3.txt data. It can be seen that the distribution is somewhat symmetrical.

This is a histogram of mock_data_4.txt data. The minimum data value of mock_data_4.txt is 0.0077590982505106884 and the maximum is 27.191265850251494. Therefore, the data range is approximately 0.007 < x < 27.19.



For mock_data_4.txt data, binsize=500 seems most appropriate. Unlike the data we looked at earlier, it has an asymmetric distribution and the distribution of data values is skewed to the left.

b. Then calculate the following for each dataset:

1. Mean
2. Geometric mean (can you?!)
3. Median
4. Mode
5. Variance
6. Standard deviation
7. Skewness
8. Kurtosis

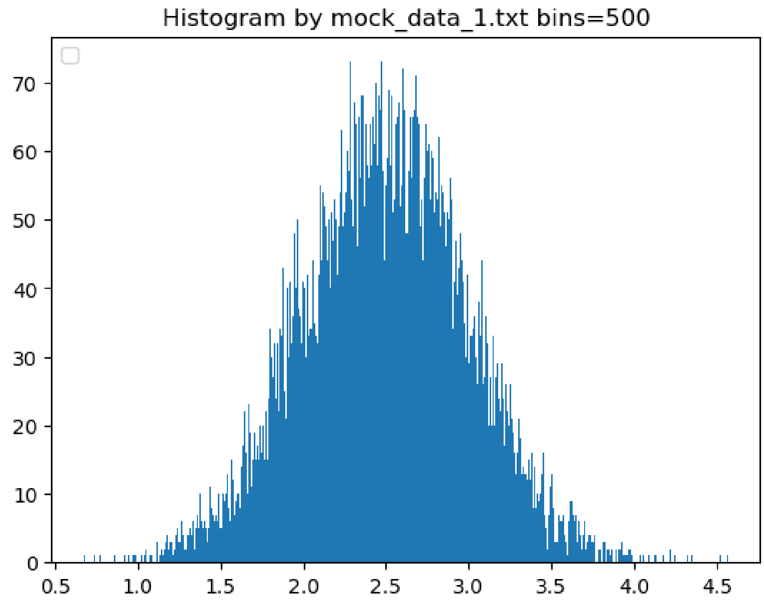
<calculate the following for mock_data_1.txt dataset>
mean: 2.498734561848049
geo_mean: inf
median: 2.502429616649482
mode: Not exist
variance: 0.24858959854287538
StandardDeviation: 0.4985876036795092
skewness: 0.011738658428985149
kurtosis: 3.0241187574775616

<calculate the following for mock_data_2.txt dataset>
mean: -0.00014846001134048625
geo_mean: 0.0
median: -0.012107505572214567
mode: Not exist
variance: 1.0147123901919113
StandardDeviation: 1.0073293355163997
skewness: 0.06107852058808343
kurtosis: 2.957203337827449

<calculate the following for mock_data_3.txt dataset>
mean: 9.9772
geo_mean: nan
median: 10.0
mode: 9.0
variance: 9.996280160000028
StandardDeviation: 3.161689447115265
skewness: 0.26248765623225234
kurtosis: 2.993012413793572

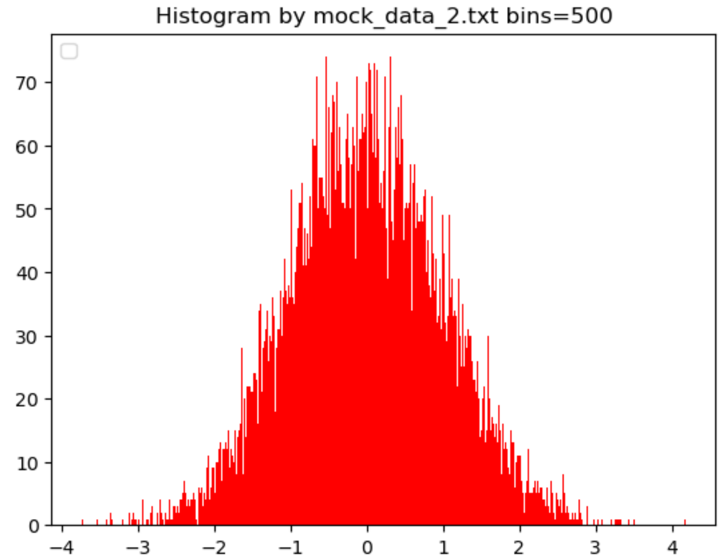
<calculate the following for mock_data_4.txt dataset>
mean: 3.9935028796220444
geo_mean: inf
median: 3.3652271344676468
mode: Not exist
variance: 7.9649963543468925
StandardDeviation: 2.822232512453021
skewness: 1.385146148973597
kurtosis: 5.917344575128322

c. **How do you interpret your results for each dataset?**



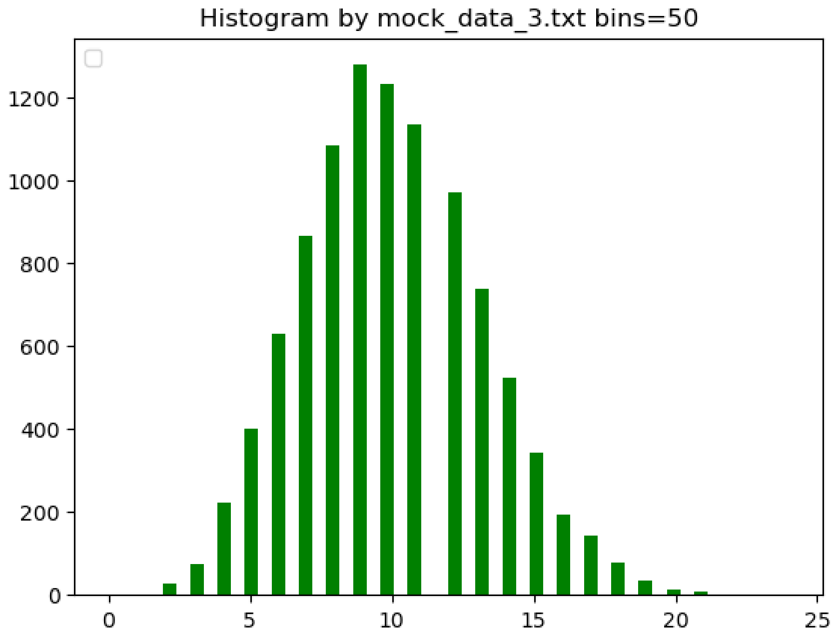
<calculate the following for mock_data_1.txt dataset>
mean: 2.498734561848049
geo_mean: inf
median: 2.502429616649482
mode: Not exist
variance: 0.24858959854287538
StandardDeviation: 0.4985876036795092
skewness: 0.011738658428985149
kurtosis: 3.0241187574775616

The shape of the graph closely resembles a normal distribution, with the mean and median appearing as nearly identical values. There is no mode in the dataset, and the data is distributed predominantly around the median, resulting in low values for variance and standard deviation. Due to its close symmetry like the normal distribution, the skewness value is low, and the kurtosis value is close to 3.



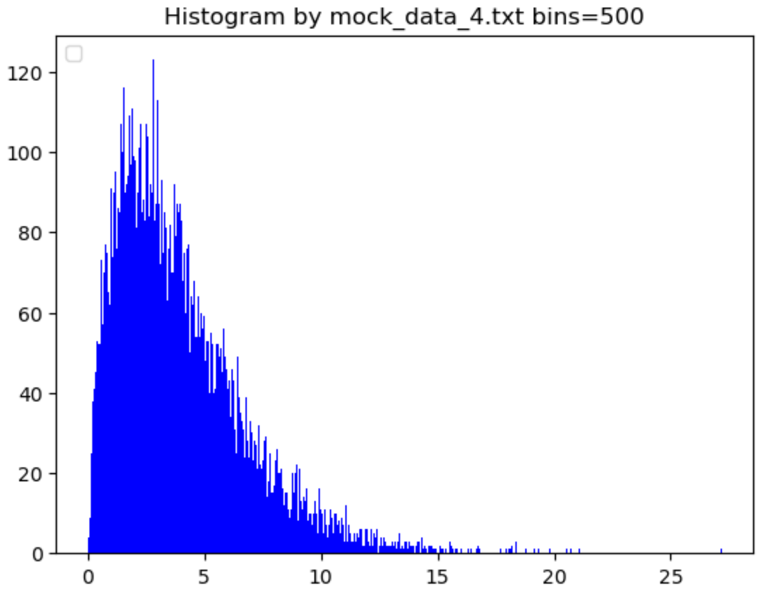
<calculate the following for mock_data_2.txt dataset>
mean: -0.00014846001134048625
geo_mean: 0.0
median: -0.012107505572214567
mode: Not exist
variance: 1.0147123901919113
StandardDeviation: 1.0073293355163997
skewness: 0.06107852058808343
kurtosis: 2.957203337827449

Same as the previous graph, the shape of the graph closely resembles a normal distribution, with the mean and median appearing as nearly identical values. There is no mode in the dataset, and the data is distributed predominantly around the median, resulting in low values for variance and standard deviation. However, mock_data_2.txt has larger data distribution. Due to its close symmetry like the normal distribution, the skewness value is low, and the kurtosis value is close to 3.



<calculate the following for mock_data_3.txt dataset>
mean: 9.9772
geo_mean: nan
median: 10.0
mode: 9.0
variance: 9.996280160000028
StandardDeviation: 3.161689447115265
skewness: 0.26248765623225234
kurtosis: 2.993012413793572

Same as the previous graphs, the shape of the graph closely resembles a normal distribution, with the mean, median and mode appearing as nearly identical values. Because the deviation between the values is very large, variance and standard deviation has higher values than the previous graphs. Due to its close symmetry like the normal distribution, the skewness value is low, and the kurtosis value is close to 3.



<calculate the following for mock_data_4.txt dataset>
mean: 3.9935028796220444
geo_mean: inf
median: 3.3652271344676468
mode: Not exist
variance: 7.9649963543468925
StandardDeviation: 2.822232512453021
skewness: 1.385146148973597
kurtosis: 5.917344575128322

This graph has a noticeably different shape from the previous graphs. Because the graph shape is skewed to the left, the mean value is greater than the median value. It has a data range similar to mock_data_3.txt, so the variance and standard deviation appear similarly large. Because it is biased to the left, the skewness is large and the kurtosis value also appears far from 3.

2.2 Problem B

a. Divide each data set of question A, to 10 equal sub-sets (set 1 from 1st data to 1000th data, set 2 from 1001th data to 2000th data,.. etc) and calculate again the quantities of question A for each sub-set.

mock_data_1.txt subset

<calculate the following for subset dataset 1>
mean: 2.4858052924884353
geo_mean: inf
median: 2.48434469054677
mode: Not exist
variance: 0.2460277164518672
StandardDeviation: 0.49601181079876233
skewness: 0.09575763322813595
kurtosis: 2.933979479546409

<calculate the following for subset dataset 2>
mean: 2.508803241529174
geo_mean: inf
median: 2.5192608614404106
mode: Not exist
variance: 0.24178454859888737
StandardDeviation: 0.49171592266153774
skewness: -0.16177283808103388
kurtosis: 3.015858588822157

<calculate the following for subset dataset 3>
mean: 2.5164763500653033
geo_mean: inf
median: 2.522969711873708
mode: Not exist
variance: 0.2613159954489752
StandardDeviation: 0.5111907622883802
skewness: -0.10972560898072747
kurtosis: 2.6653909040907684

<calculate the following for subset dataset 4>
mean: 2.5192874477611675

geo_mean: inf
median: 2.5173376337607114
mode: Not exist
variance: 0.25669233102900546
StandardDeviation: 0.5066481333519403
skewness: 0.05126125048111259
kurtosis: 3.091474810026859

<calculate the following for subset dataset 5>
mean: 2.501467188947697
geo_mean: inf
median: 2.527756796997764
mode: Not exist
variance: 0.23381757391905703
StandardDeviation: 0.4835468683789163
skewness: -0.12599935871928417
kurtosis: 2.8844542102131303

<calculate the following for subset dataset 6>
mean: 2.47812099736616
geo_mean: inf
median: 2.4596428577546554
mode: Not exist
variance: 0.23060461757044598
StandardDeviation: 0.48021309600056306
skewness: 0.14122203703015895
kurtosis: 3.179038627921752

<calculate the following for subset dataset 7>
mean: 2.5055902633947813
geo_mean: inf
median: 2.5076739036864955
mode: Not exist
variance: 0.2644109140431965
StandardDeviation: 0.5142090178547986
skewness: 0.09983087668837057
kurtosis: 3.1233245298773853

<calculate the following for subset dataset 8>
mean: 2.5000958749608344
geo_mean: inf
median: 2.515827399729301
mode: Not exist

variance: 0.2674099625419615
StandardDeviation: 0.5171169718177517
skewness: 0.0021901101485645707
kurtosis: 3.0850576896986617

<calculate the following for subset dataset 9>
mean: 2.5153192975116463
geo_mean: inf
median: 2.5161935225437055
mode: Not exist
variance: 0.23935946041142148
StandardDeviation: 0.48924376379410445
skewness: 0.12405175635471381
kurtosis: 3.210962902545103

<calculate the following for subset dataset 10>
mean: 2.4563796644553118
geo_mean: inf
median: 2.4595249556942624
mode: Not exist
variance: 0.24091689802994648
StandardDeviation: 0.49083286160356715
skewness: -0.022172506614848 94
kurtosis: 3.029646535931132

mock_data_2.txt subset

<calculate the following for subset dataset 1>
mean: -0.017234850653007113
geo_mean: (0.5554722930067865+0.0017450734160479972j)
median: -0.06088746684570992
mode: Not exist
variance: 1.0054032895333112
StandardDeviation: 1.0026980051507588
skewness: 0.1256775113989512
kurtosis: 2.917384736093718

<calculate the following for subset dataset 2>
mean: -0.005514814730216284
geo_mean: (0.5029972884092292+0.0015802177847535 76j)
median: -0.0007231537857209261

mode: Not exist
variance: 0.9164713136630653
StandardDeviation: 0.9573250825414872
skewness: 0.10851278475070665
kurtosis: 2.8731127957814326

<calculate the following for subset dataset 3>
mean: 0.021434642379263923
geo_mean: (0.5177432020737232+0.0016265435911983143j)
median: 0.018176061314683224
mode: Not exist
variance: 1.0512664484158538
StandardDeviation: 1.0253128539211112
skewness: -0.029296983645157778
kurtosis: 3.0112792470586336

<calculate the following for subset dataset 4>
mean: -0.01441729294097353
geo_mean: (0.555041572800208+0.0017437202641595636j)
median: 0.03706489418181301
mode: Not exist
variance: 1.0709708181777777
StandardDeviation: 1.0348771995641695
skewness: -0.08711926902096166
kurtosis: 2.8960359564319726

<calculate the following for subset dataset 5>
mean: 0.03525934741458711
geo_mean: (0.5571589377723062+0.00175037218428483j)
median: 0.014315905223084516
mode: Not exist
variance: 1.0217508345353703
StandardDeviation: 1.010816914448591
skewness: 0.07419974465953064
kurtosis: 2.836334808216459

<calculate the following for subset dataset 6>
mean: 0.002635414741401607
geo_mean: (0.5142688104379393+0.0016156284320501839j)
median: -0.038115578711360884
mode: Not exist
variance: 0.9777788482647767
StandardDeviation: 0.9888270062375808

skewness: 0.2234355924130936
kurtosis: 3.2777085418582783

<calculate the following for subset dataset 7>
mean: -0.007814462932440858
geo_mean: 0.5264949549891179
median: -0.0016091662703748419
mode: Not exist
variance: 1.0078098329139222
StandardDeviation: 1.003897321897973
skewness: 0.007730463230614208
kurtosis: 3.11328753985225

<calculate the following for subset dataset 8>
mean: 0.018970019053735732
geo_mean: (0.5295171952467829+0.0016635328033439254j)
median: -0.0016478097571734924
mode: Not exist
variance: 1.120304508842907
StandardDeviation: 1.0584443815538478
skewness: 0.02025136276425956
kurtosis: 2.9890796757236084

<calculate the following for subset dataset 9>
mean: 0.004575316502281025
geo_mean: (0.5583489209564183+0.0017541106390129631j)
median: -0.016634261253034202
mode: Not exist
variance: 1.0132062899577226
StandardDeviation: 1.0065814869933396
skewness: 0.07248166673867679
kurtosis: 2.7288161344525212

<calculate the following for subset dataset 10>
mean: -0.03937791894803653
geo_mean: 0.5486979753647335
median: -0.08670029357455472
mode: Not exist
variance: 0.9579245340634231
StandardDeviation: 0.9787361922721685
skewness: 0.1245789388595356
kurtosis: 2.845764721071054

mock_data_3.txt subset

<calculate the following for subset dataset 1>
mean: 9.978
geo_mean: inf
median: 10.0
mode: 9.0
variance: 10.149516000000036
StandardDeviation: 3.185830503965965
skewness: 0.17015433293519475
kurtosis: 2.8161261786548226

<calculate the following for subset dataset 2>
mean: 10.014
geo_mean: inf
median: 10.0
mode: 10.0
variance: 10.103803999999995
StandardDeviation: 3.1786481403263234
skewness: 0.2965969497879963
kurtosis: 3.067806147609972

<calculate the following for subset dataset 3>
mean: 9.938
geo_mean: inf
median: 10.0
mode: 10.0
variance: 10.224156000000022
StandardDeviation: 3.19752341664608
skewness: 0.35415205724655435
kurtosis: 3.168935577381038

<calculate the following for subset dataset 4>
mean: 10.083
geo_mean: inf
median: 10.0
mode: 9.0
variance: 9.850110999999999
StandardDeviation: 3.1384886490156356
skewness: 0.3130493006853107
kurtosis: 3.0128084384313234

<calculate the following for subset dataset 5>
mean: 9.953
geo_mean: inf
median: 10.0
mode: 11.0
variance: 10.188791000000037
StandardDeviation: 3.1919885651424313
skewness: 0.25396947484105503
kurtosis: 2.9941662968791776

<calculate the following for subset dataset 6>
mean: 9.857
geo_mean: inf
median: 10.0
mode: 9.0
variance: 9.61855099999992
StandardDeviation: 3.101378886882401
skewness: 0.2211234361783912
kurtosis: 2.8450217050961526

<calculate the following for subset dataset 7>
mean: 9.867
geo_mean: nan
median: 10.0
mode: 9.0
variance: 9.643310999999996
StandardDeviation: 3.1053680941234583
skewness: 0.2241042855603896
kurtosis: 2.875336729361191

<calculate the following for subset dataset 8>
mean: 10.027
geo_mean: inf
median: 10.0
mode: 11.0
variance: 9.598270999999945
StandardDeviation: 3.0981076482265664
skewness: 0.20841474962740483
kurtosis: 2.6403572911394093

<calculate the following for subset dataset 9>
mean: 10.036
geo_mean: inf

median: 10.0
mode: 9.0
variance: 10.464704000000031
StandardDeviation: 3.2349194734954425
skewness: 0.2888345795414705
kurtosis: 3.1146793328096556

<calculate the following for subset dataset 10>
mean: 10.019
geo_mean: inf
median: 10.0
mode: 9.0
variance: 10.072639000000027
StandardDeviation: 3.173742113026833
skewness: 0.27918118996106395
kurtosis: 3.287002584913436

mock_data_4.txt subset

<calculate the following for subset dataset 1>
mean: 3.899484570156791
geo_mean: inf
median: 3.3461399098108657
mode: Not exist
variance: 7.46962286833211
StandardDeviation: 2.7330610802417334
skewness: 1.3603894543607413
kurtosis: 5.69797433439668

<calculate the following for subset dataset 2>
mean: 4.012035972929117
geo_mean: inf
median: 3.4409184533404726
mode: Not exist
variance: 7.080075905385176
StandardDeviation: 2.660841202587102
skewness: 1.1453658145349797
kurtosis: 4.809593758015298

<calculate the following for subset dataset 3>
mean: 4.102864181002587
geo_mean: inf
median: 3.297276524474827
mode: Not exist
variance: 9.755844721293272
StandardDeviation: 3.1234347634124315
skewness: 1.6742002079310436
kurtosis: 7.689490559018485

<calculate the following for subset dataset 4>
mean: 4.03050623324052
geo_mean: inf
median: 3.448125593404048
mode: Not exist
variance: 7.754486977373974
StandardDeviation: 2.7846879497304493
skewness: 1.2711493942805294
kurtosis: 5.036320380968432

<calculate the following for subset dataset 5>
mean: 4.008557014331891
geo_mean: inf
median: 3.5006349256193596
mode: Not exist
variance: 7.639402977855046
StandardDeviation: 2.7639469925914004
skewness: 1.2335040158445711
kurtosis: 4.995749759633996

<calculate the following for subset dataset 6>
mean: 4.131723711735572
geo_mean: inf
median: 3.603778942530715
mode: Not exist
variance: 8.17781573999677
StandardDeviation: 2.859688049420211
skewness: 1.2010593388935613
kurtosis: 5.0476291669554465

<calculate the following for subset dataset 7>
mean: 3.9425572795357304
geo_mean: inf

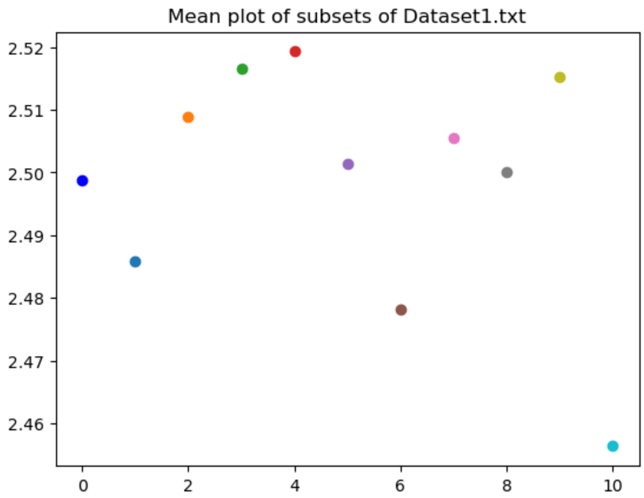
median: 3.2927628000239144
mode: Not exist
variance: 7.552513016221989
StandardDeviation: 2.7481835848832934
skewness: 1.3610156175275385
kurtosis: 5.70076039108694

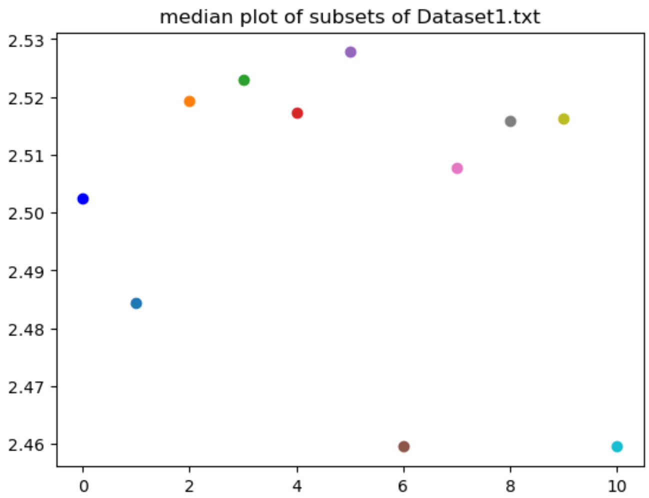
<calculate the following for subset dataset 8>
mean: 3.948764711101466
geo_mean: inf
median: 3.2456282812941413
mode: Not exist
variance: 8.693234018845574
StandardDeviation: 2.948429076448266
skewness: 1.5658697600594647
kurtosis: 6.283655341592659

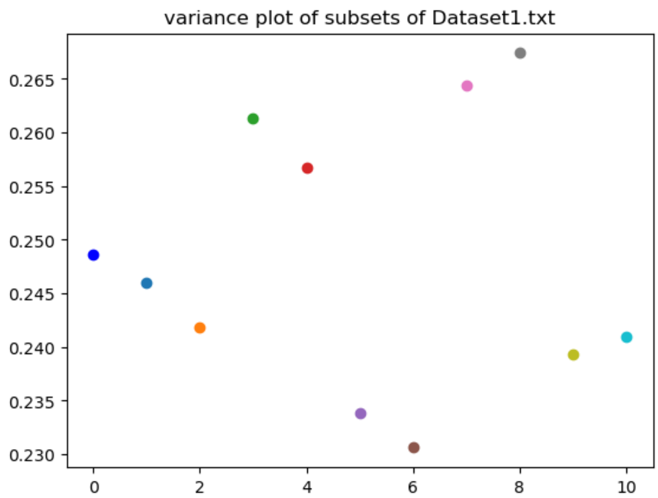
<calculate the following for subset dataset 9>
mean: 3.8718893209438985
geo_mean: inf
median: 3.238889322199208
mode: Not exist
variance: 7.545008106119426
StandardDeviation: 2.7468178145118083
skewness: 1.3498644743368338
kurtosis: 5.802680935519187

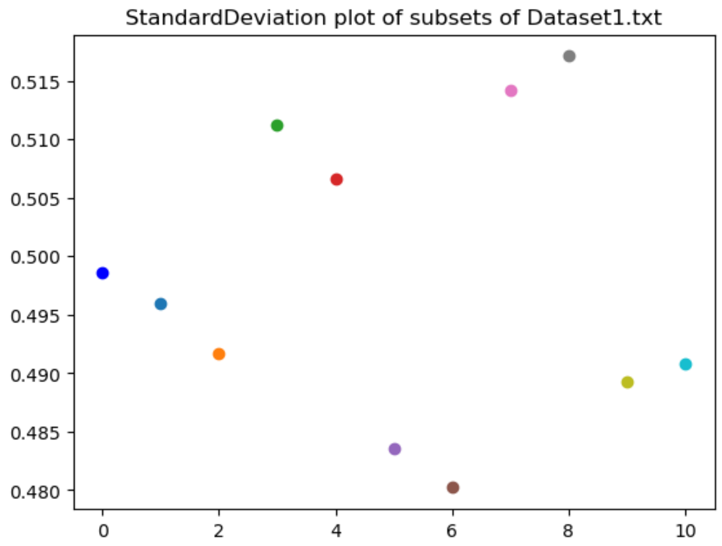
<calculate the following for subset dataset 10>
mean: 3.986645801243016
geo_mean: inf
median: 3.4223497343013882
mode: Not exist
variance: 7.920681691093742
StandardDeviation: 2.8143705674792976
skewness: 1.4449958076654776
kurtosis: 6.225988522370861

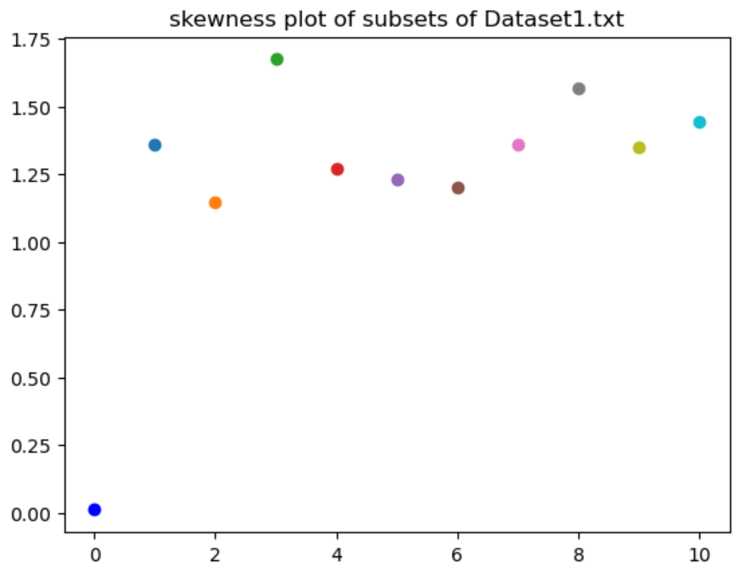
b. What do you learn from this practice?

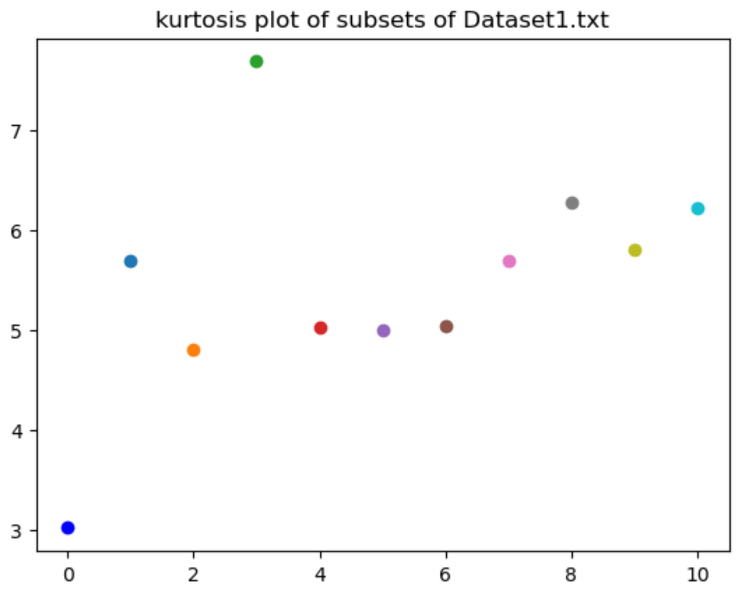


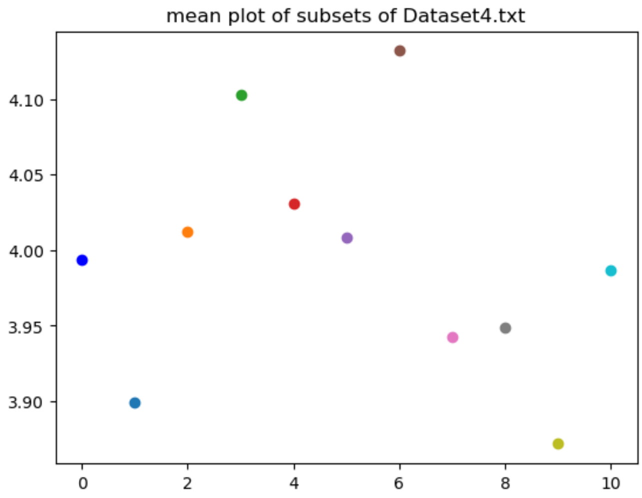


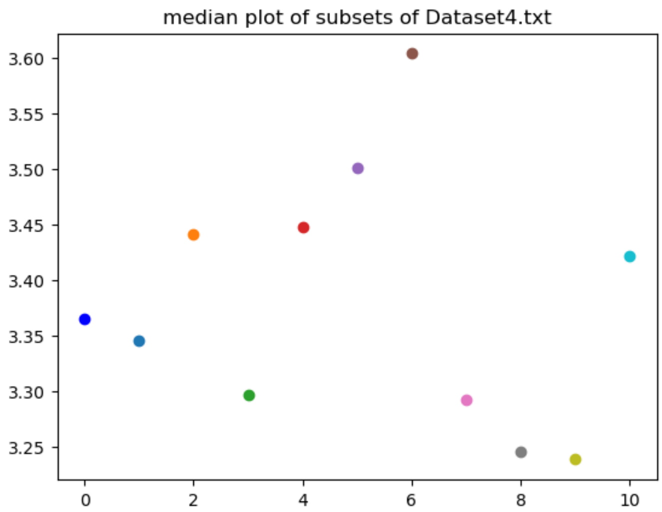


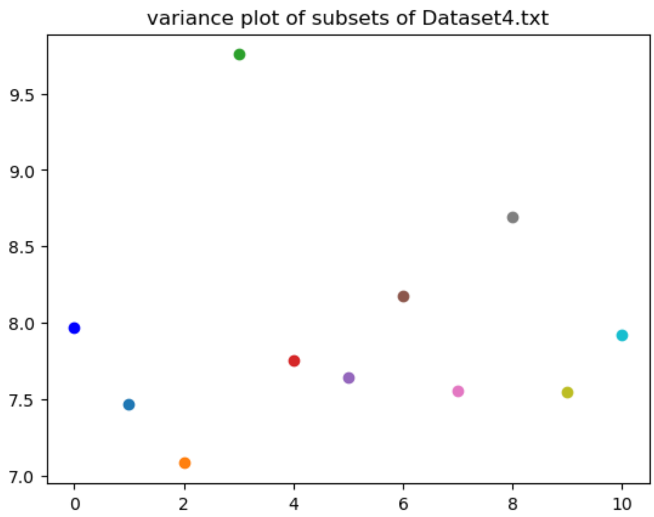

StandardDeviation plot of subsets of Dataset1.txt
0.515
0.510
0.505
0.500
0.495
0.490
0.485
0.480
0
2
4
6
8
10

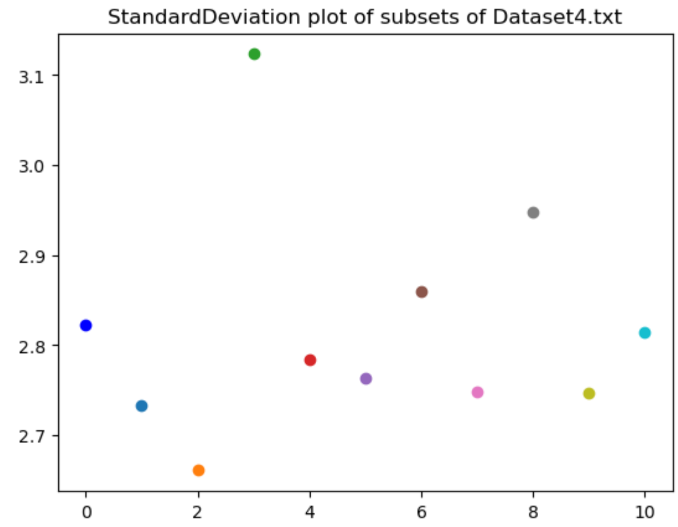

skewness plot of subsets of Dataset1.txt
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
2
4
6
8
10

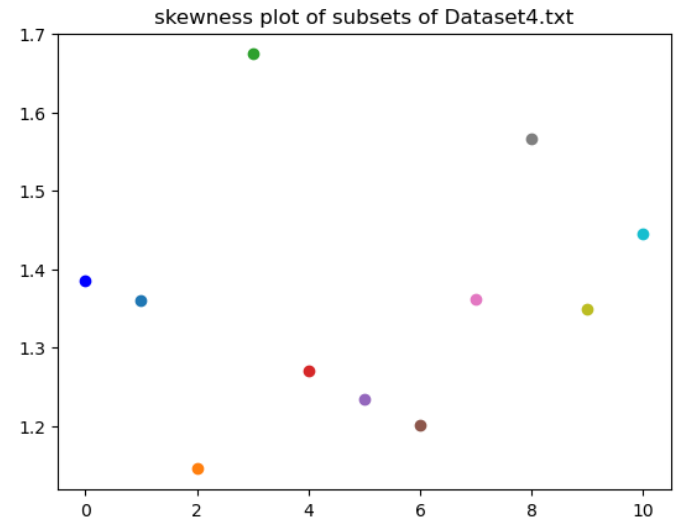

kurtosis plot of subsets of Dataset1.txt
7
6
5
4
3
0
2
4
6
8
10

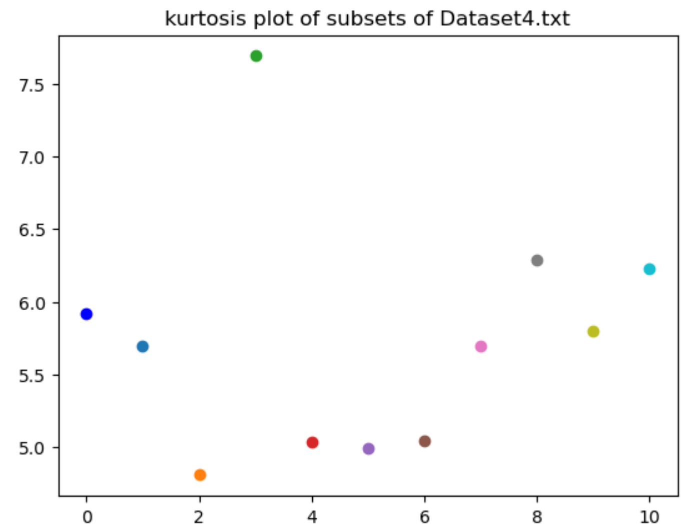

mean plot of subsets of Dataset4.txt
4.10
4.05
4.00
3.95
3.90
0
2
4
6
8
10


median plot of subsets of Dataset4.txt
3.60
3.55
3.50
3.45
3.40
3.35
3.30
3.25
0
2
4
6
8
10


variance plot of subsets of Dataset4.txt
9.5
9.0
8.5
8.0
7.5
7.0
0
2
4
6
8
10

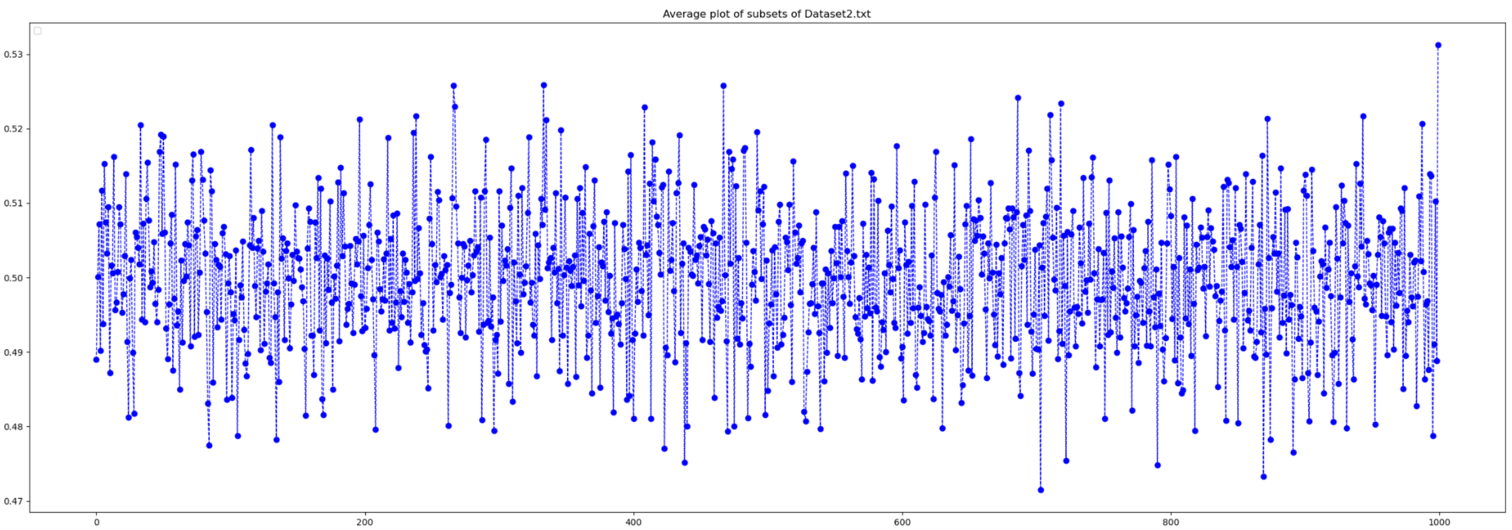






In general, the statistical calculation values of the subsets show similar patterns.

2.3 Problem C



a. Divide the Dataset2.txt to 1000 subsets and calculate the average value of each subset. Note that this is a large data set with 1,000,000 numbers. Plot the distribution of these 1000 average values. How does this distribution looks like? Calculate the quantities of Q.A for these 1000 averaged values.

<calculate the following for subset Dataset2>
mean: 0.4998205448001 24
geo_mean: 0.4997322092717983
median: 0.5000277711329199
mode: Not exist
variance: 8.826945052491794e-05
StandardDeviation: 0.009395182303974625
skewness: 0.003853383734848837 8
kurtosis: 2.9775116178788372

All values do not deviate significantly from the median. Therefore, the variance and standard deviation are very small. Although it is not bell-shaped, it has a graph that is almost like a straight line, so the skewness is very small because it is symmetrical.

b. What do you learn from this practice?

If the values are all similar to the mean and median, the variance and standard deviation become very small.

2.4 Problem D (Geometric mean)

Find out the 3D objects which have the maximum and minimum surface to volume ratio.

The surface-to-volume ratio of a 3D object is a measure of how much surface area the object has in relation to its volume. In general, objects with a higher surface-to-volume ratio have more surface area relative to their size, while objects with a lower surface-to-volume ratio have less surface area relative to their size.

To find 3D objects with the maximum and minimum surface-to-volume ratios, let's consider two examples:

1. Maximum Surface-to-Volume Ratio:
   - A sphere has the maximum surface-to-volume ratio among regular 3D shapes. This is because a sphere has the smallest possible surface area for a given volume. The formula for the surface area (A) and volume (V) of a sphere are as follows:
     - Surface Area (A) = 4πr^2
     - Volume (V) = (4/3)πr^3
   - Where "r" is the radius of the sphere.

Thus, Surface-to-Volume Ratio of sphere is 3/r.

2. Minimum Surface-to-Volume Ratio:
   - A cube has the minimum surface-to-volume ratio among regular 3D shapes. This is because a cube has a relatively larger surface area compared to its volume. The formula for the surface area (A) and volume (V) of a cube are as follows:
     - Surface Area (A) = 6s^2
     - Volume (V) = s^3
   - Where "s" is the length of one side of the cube.

Thus, Surface-to-Volume Ratio of sphere is 6/r.

So, in summary:
- The 3D object with the maximum surface-to-volume ratio is a sphere.
- The 3D object with the minimum surface-to-volume ratio is a cube.

2.5 Problem E

Show that Variance can be written as $V=\langle x^2\rangle-\langle x\rangle^2$ $\langle\rangle$:mean

To show that the variance of a random variable X can be written as $V(X) = E(X^2) - [E(X)]^2$, where $V(X)$ represents the variance of X and $E(X)$ represents the expected value (or mean) of X, we'll use the properties of variance and expected value.

The variance of a random variable X is defined as:

$$V(X) = E[(X - E(X))^2]$$

Now, let's expand the square inside the expectation:

$$V(X) = E[X^2 - 2X*E(X) + (E(X))^2]$$

Now, using the linearity of expectation, we can split this into three separate expectations:

$$V(X) = E(X^2) - 2E(X*E(X)) + E((E(X))^2)$$

Now, let's focus on the second term in the equation, which is $2E(X*E(X))$. $E(X)$ is a constant with respect to the expectation over X. Therefore, we can pull it out of the expectation:

$$2E(X*E(X)) = 2E(X) * E(X) = 2(E(X))^2$$

So, the equation becomes:

$$V(X) = E(X^2) - 2(E(X))^2 + E((E(X))^2)$$

Now, notice that the third term, $E((E(X))^2)$, is just a constant (the square of the mean of X). Therefore, it doesn't depend on X, and its expectation is just itself:

$$E((E(X))^2) = (E(X))^2$$

Now, we can substitute this back into our equation:

$$V(X) = E(X^2) - 2(E(X))^2 + (E(X))^2$$

Now, simplify the equation:

$$V(X) = E(X^2) - (E(X))^2$$

This is the desired result, which shows that the variance of X, $V(X)$, can be written as $V(X) = E(X^2) - (E(X))^2$.

2.6 Problem F

Show that skewness can be written as 1/sigma^3 [<x^3>-3<x><x^2>+2<x>^3]

To show that skewness can be written as:

$\text{Skewness} = 1/\sigma^3 [E(X^3) - 3E(X)E(X^2) + 2(E(X))^3]$

where Skewness represents the skewness of the random variable X, σ is the standard deviation of X, and E(X) represents the expected value (or mean) of X, we'll use the properties of skewness and expected value.

The skewness of a random variable X is defined as:

$\text{Skewness} = E[(X - E(X))^3] / \sigma^3$

Now, let's expand the cube inside the expectation:

$\text{Skewness} = E[X^3 - 3X^2E(X) + 3XE(X)^2 - (E(X))^3] / \sigma^3$

Now, using the linearity of expectation, we can split this into four separate expectations:

$\text{Skewness} = (1/\sigma^3) [E(X^3) - 3E(X^2)E(X) + 3E(X)E(X)^2 - (E(X))^3]$

Now, let's simplify this expression:

$\text{Skewness} = (1/\sigma^3) [E(X^3) - 3E(X^2)E(X) + 2E(X)E(X)^2]$

Notice that the third term, $3E(X)E(X)^2$, can be simplified further:

$3E(X)E(X)^2 = 3(E(X))^3$

So, the equation becomes:

$\text{Skewness} = (1/\sigma^3) [E(X^3) - 3E(X^2)E(X) + 2(E(X))^3]$

This is the desired result, which shows that the skewness of X can be written as $\text{Skewness} = 1/\sigma^3 [E(X^3) - 3E(X^2)E(X) + 2(E(X))^3]$.

Reference

- "Probability and Statistics" by Morris H. DeGroot and Mark J. Schervish.
- "Mathematical Statistics and Data Analysis" by John A. Rice.
- "Introduction to Probability" by Joseph K. Blitzstein and Jessica Hwang.

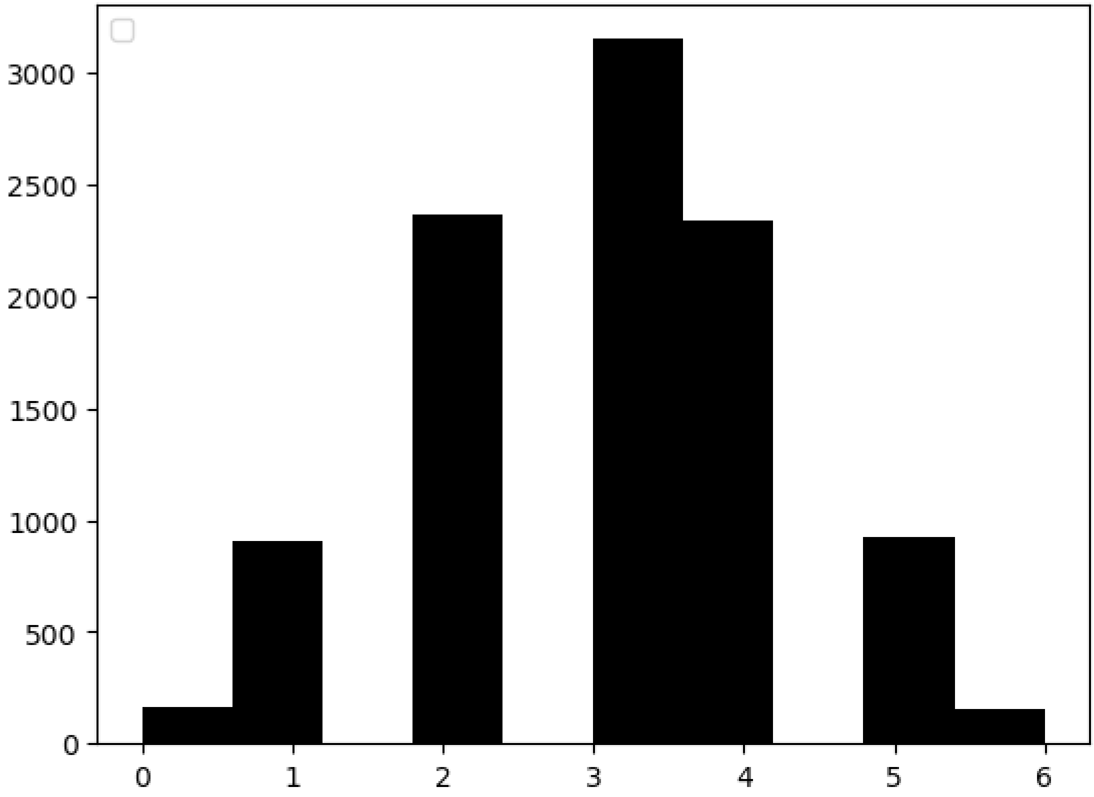
Homework 2

3.1 Problem 1

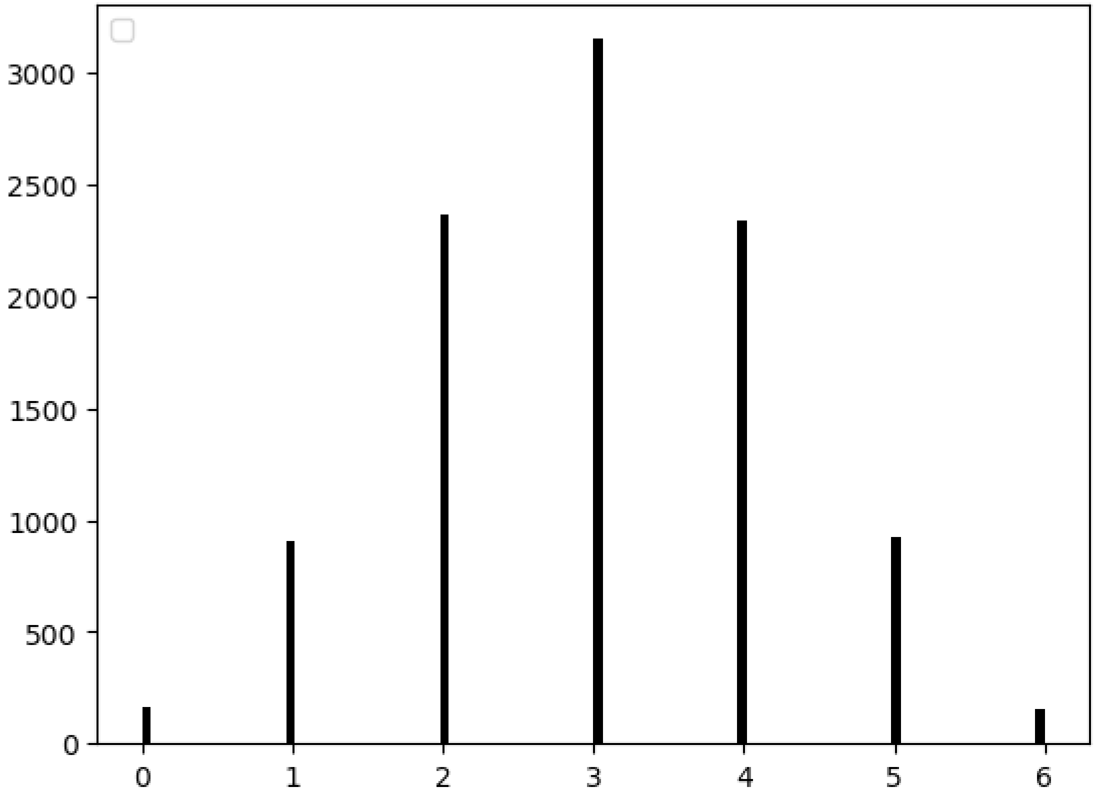
The zip file contains 9 files. For the mock_data_1 to mock_data 6, plot their histograms (play with bin sizes and find an appropriate bin-size) and try to have a guess from what distribution these data are drawn. Can you also guess the parameters of the distributions?

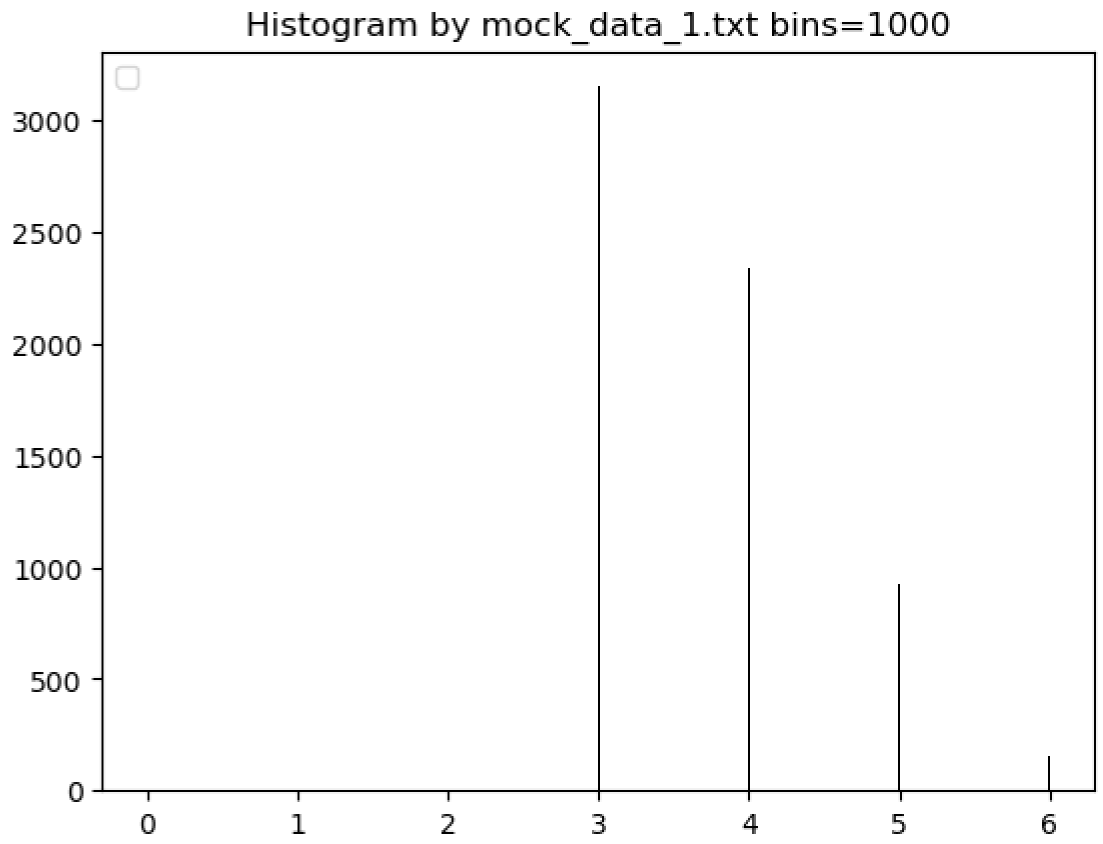
First, I tried to set the bins as 10, 100, 1000.

Histogram by mock_data_1.txt bins=10

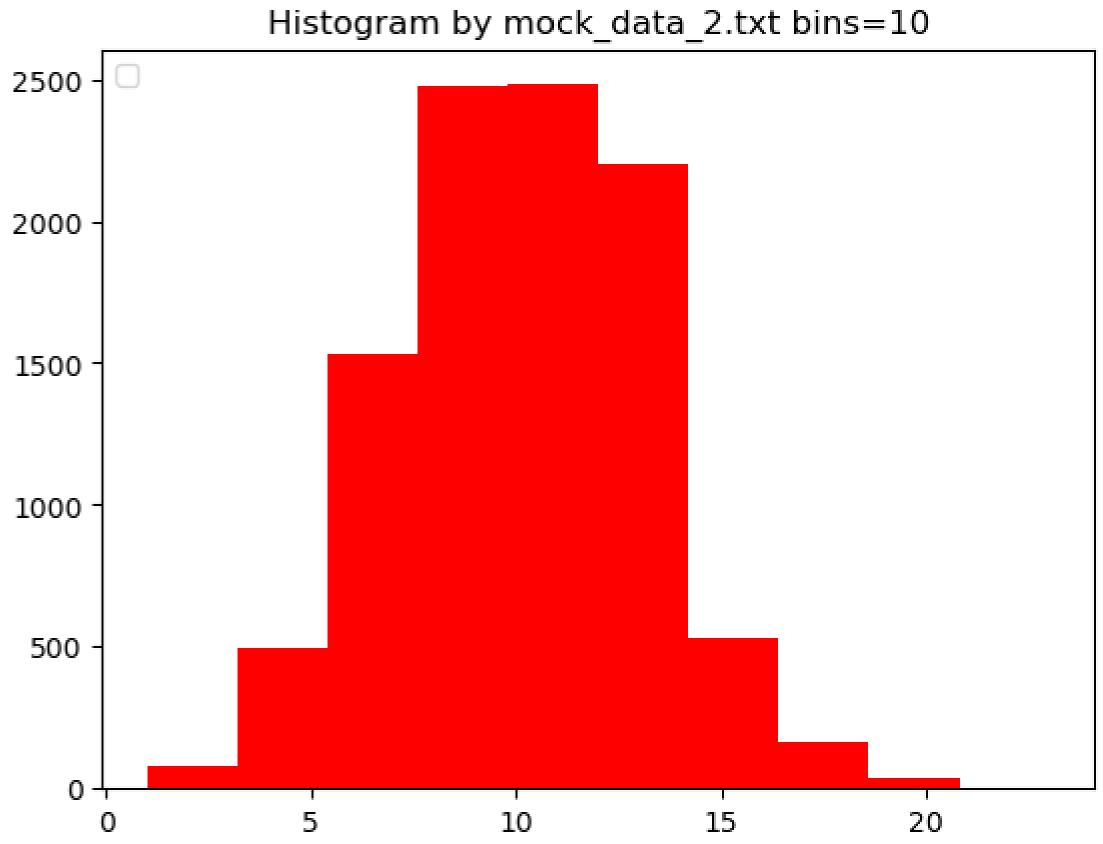


Histogram by mock_data_1.txt bins=100

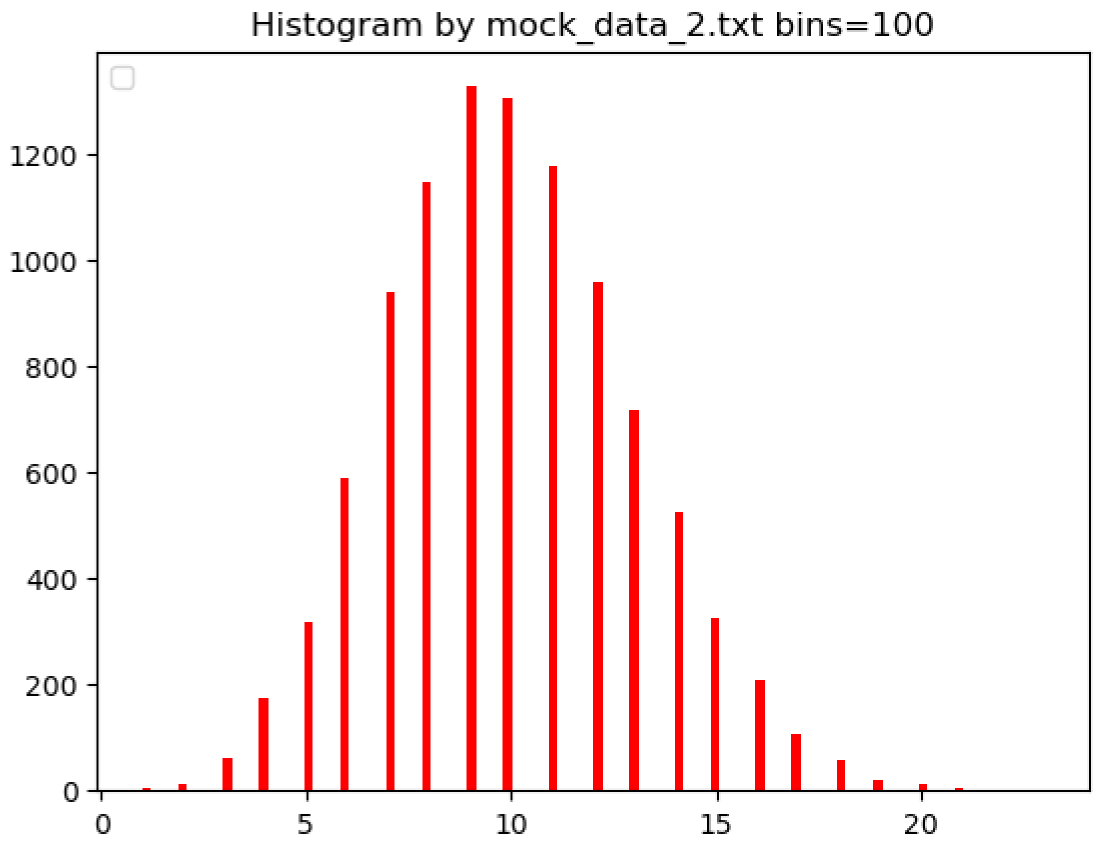


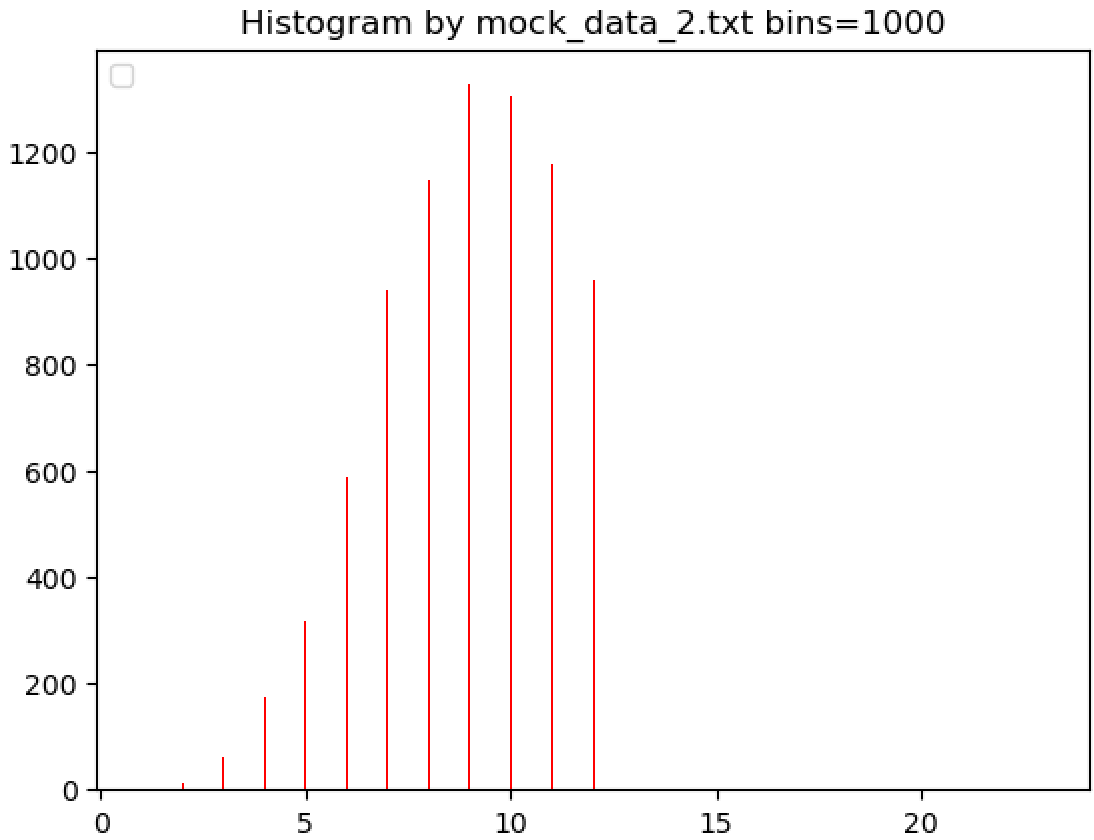


Histogram by mock_data_1.txt bins=1000
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6

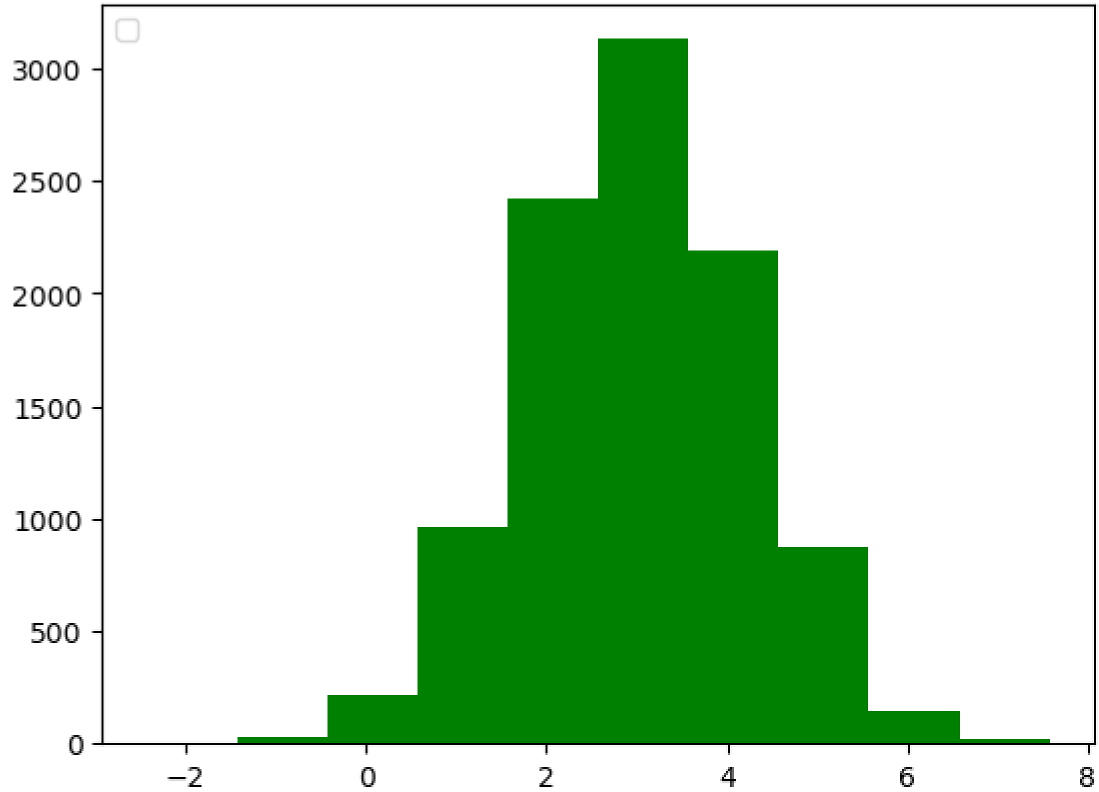


Histogram by mock_data_2.txt bins=10
2500
2000
1500
1000
500
0
0
5
10
15
20

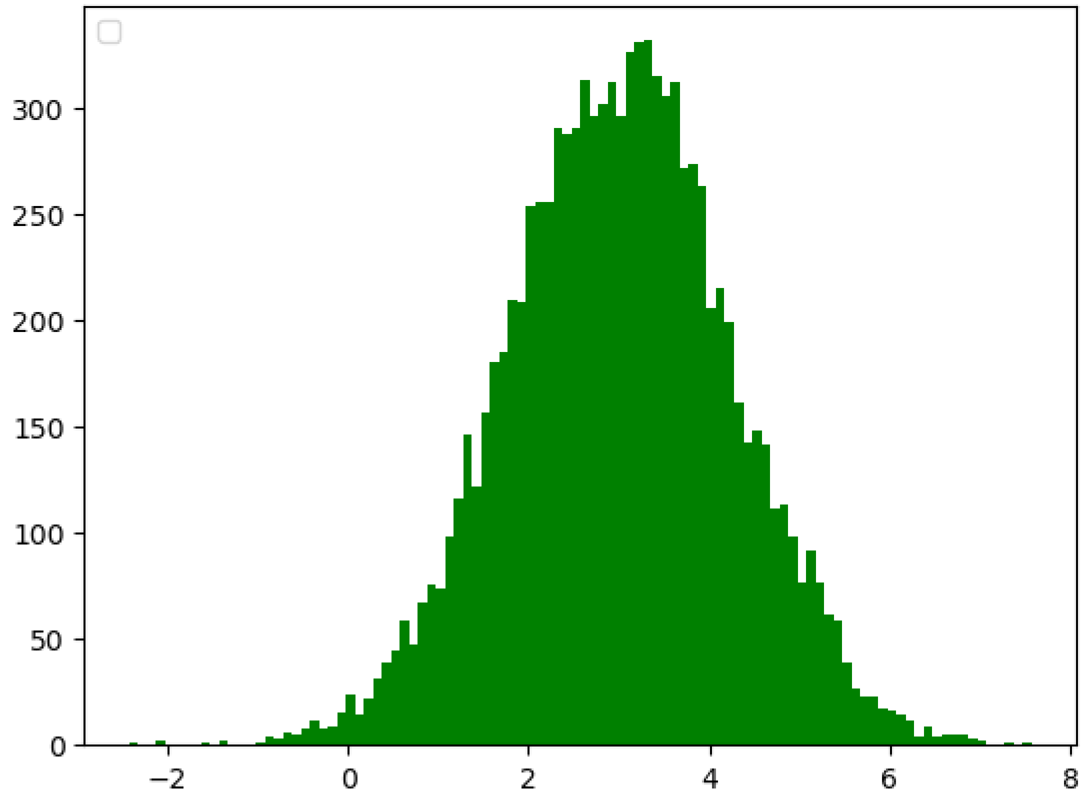

Histogram by mock_data_2.txt bins=100
0
200
400
600
800
1000
1200
0
5
10
15
20

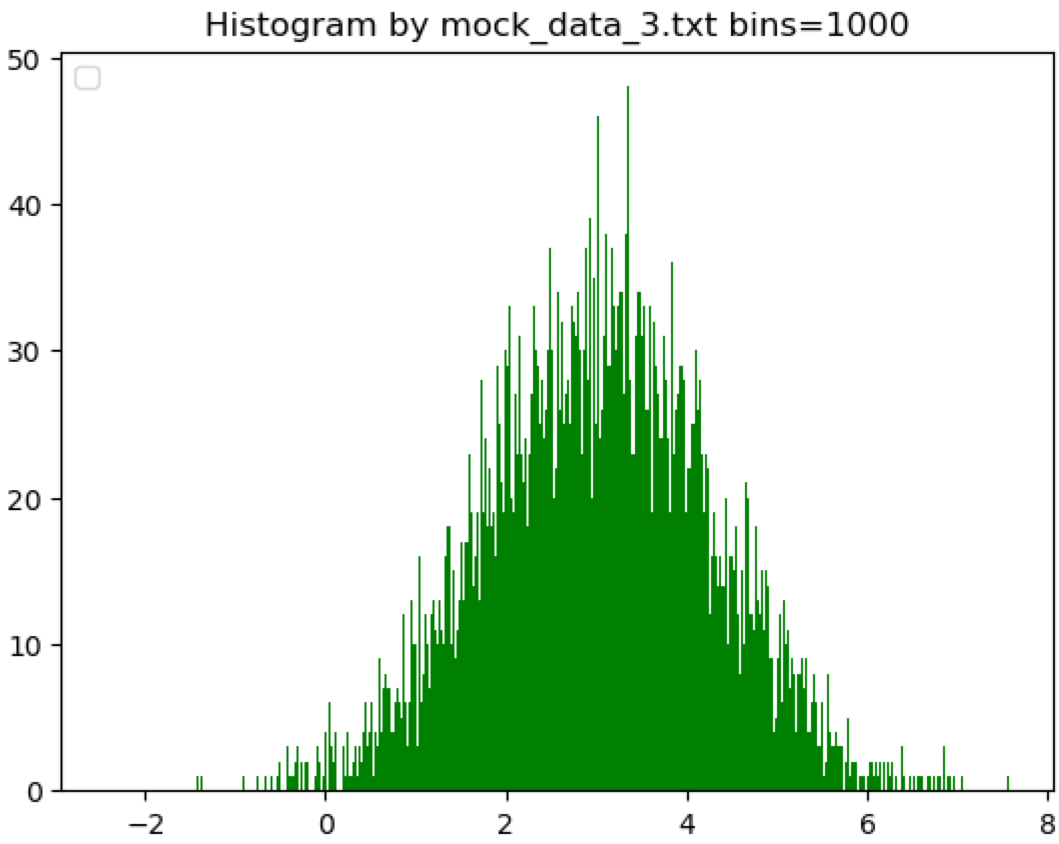

Histogram by mock_data_2.txt bins=1000
0
200
400
600
800
1000
1200
0
5
10
15
20

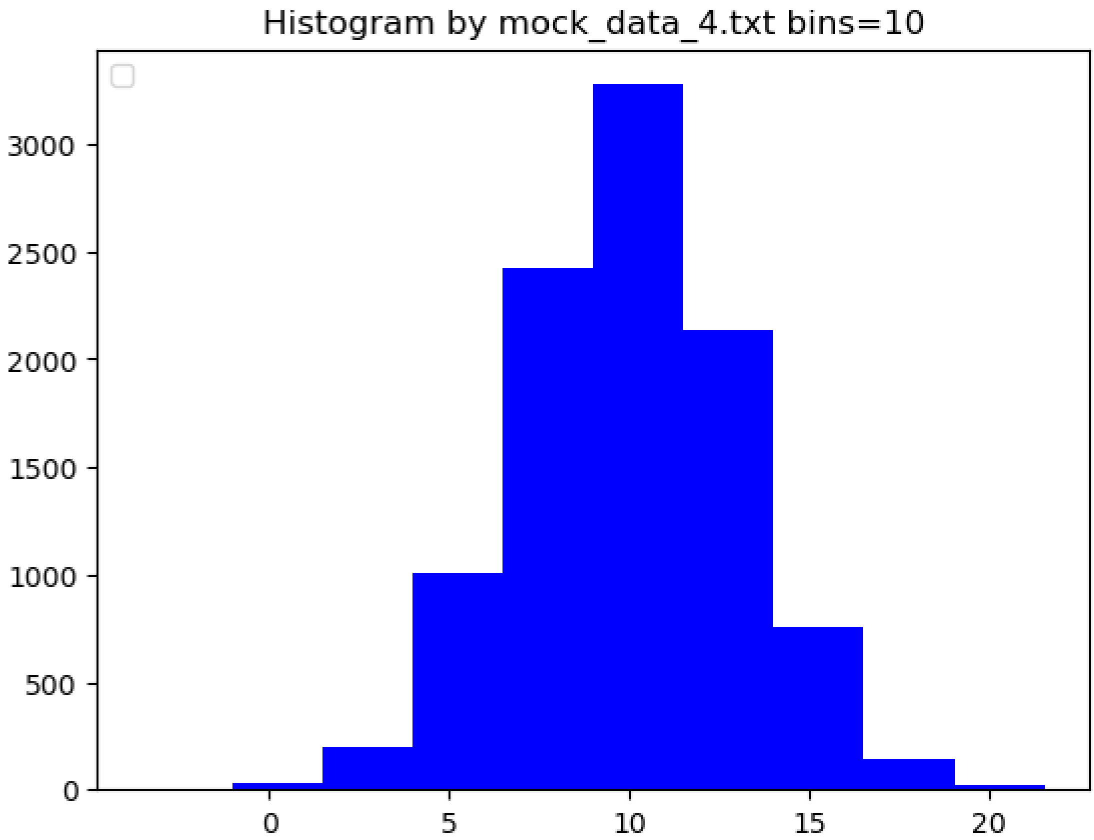
Histogram by mock_data_3.txt bins=10

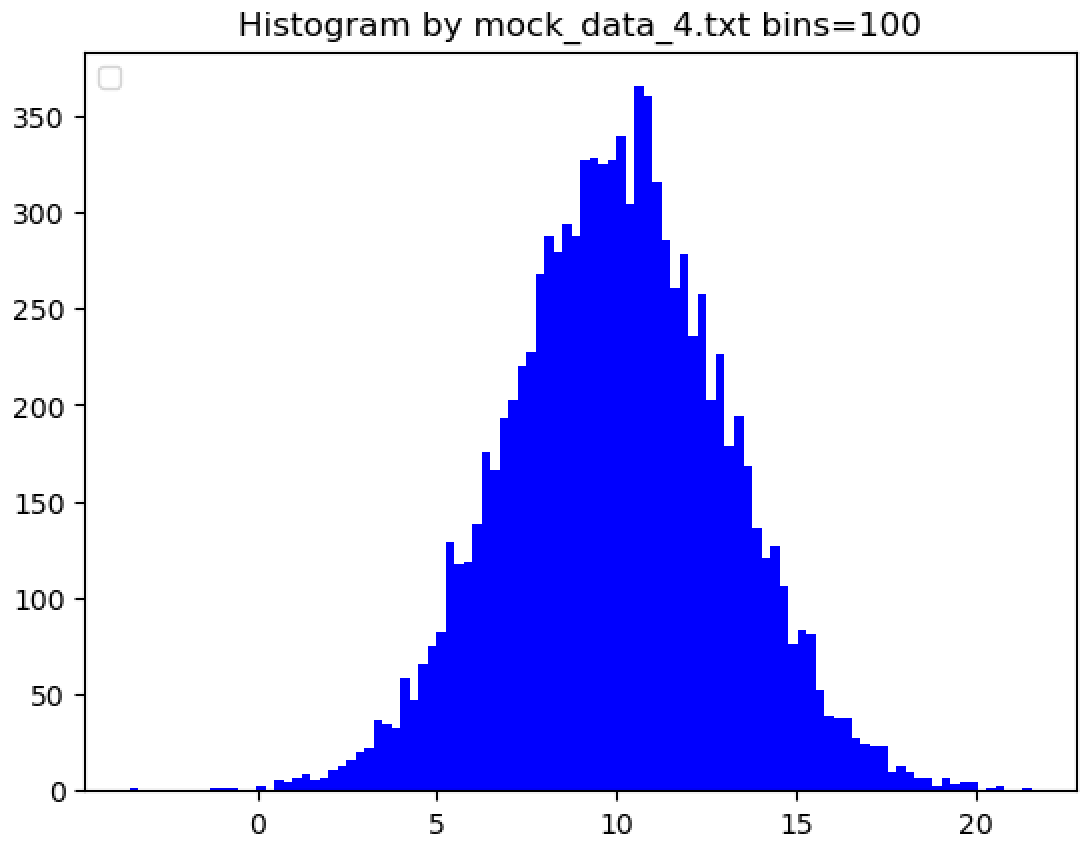


Histogram by mock_data_3.txt bins=100

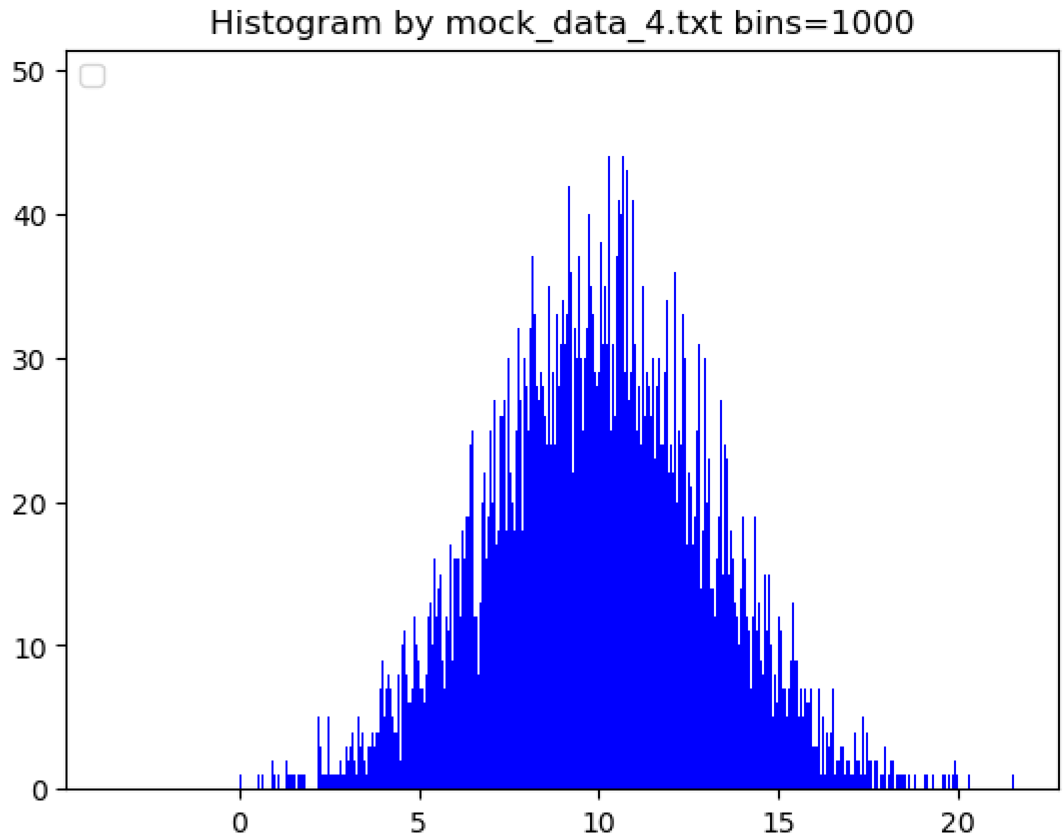



Histogram by mock_data_3.txt bins=1000
50
40
30
20
10
0
−2
0
2
4
6
8


Histogram by mock_data_4.txt bins=10
3000
2500
2000
1500
1000
500
0
0
5
10
15
20

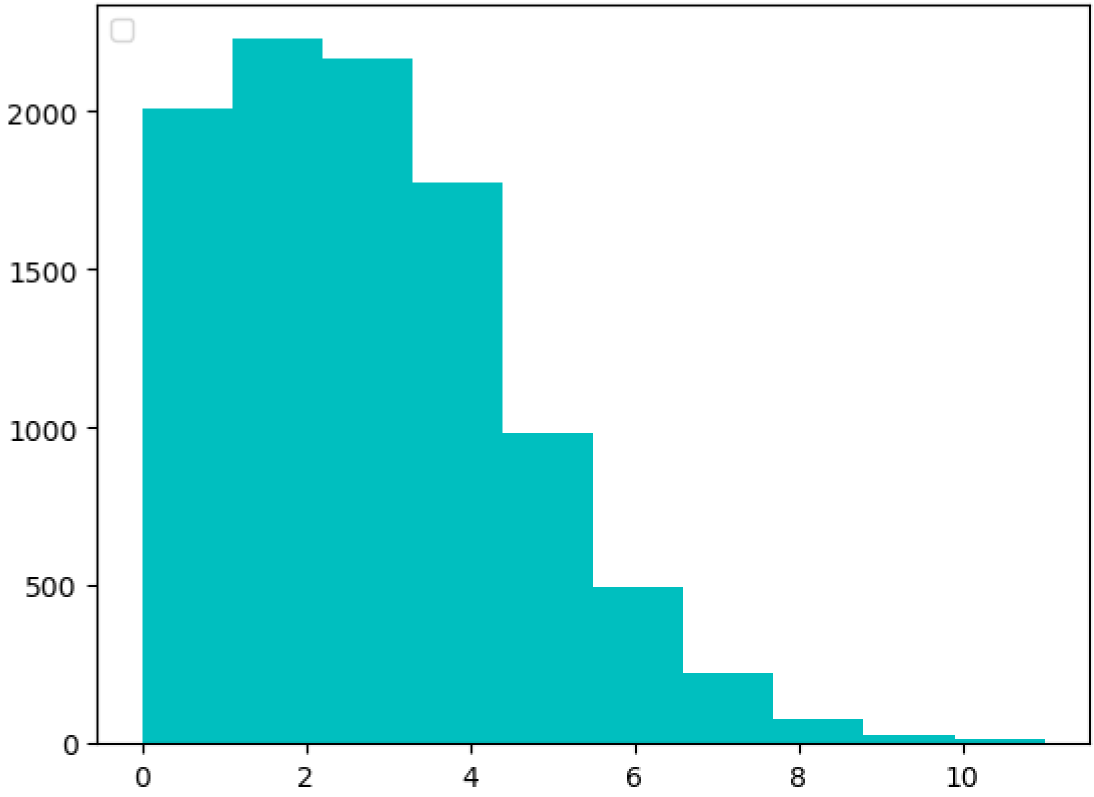


Histogram by mock_data_4.txt bins=100
350
300
250
200
150
100
50
0
0
5
10
15
20

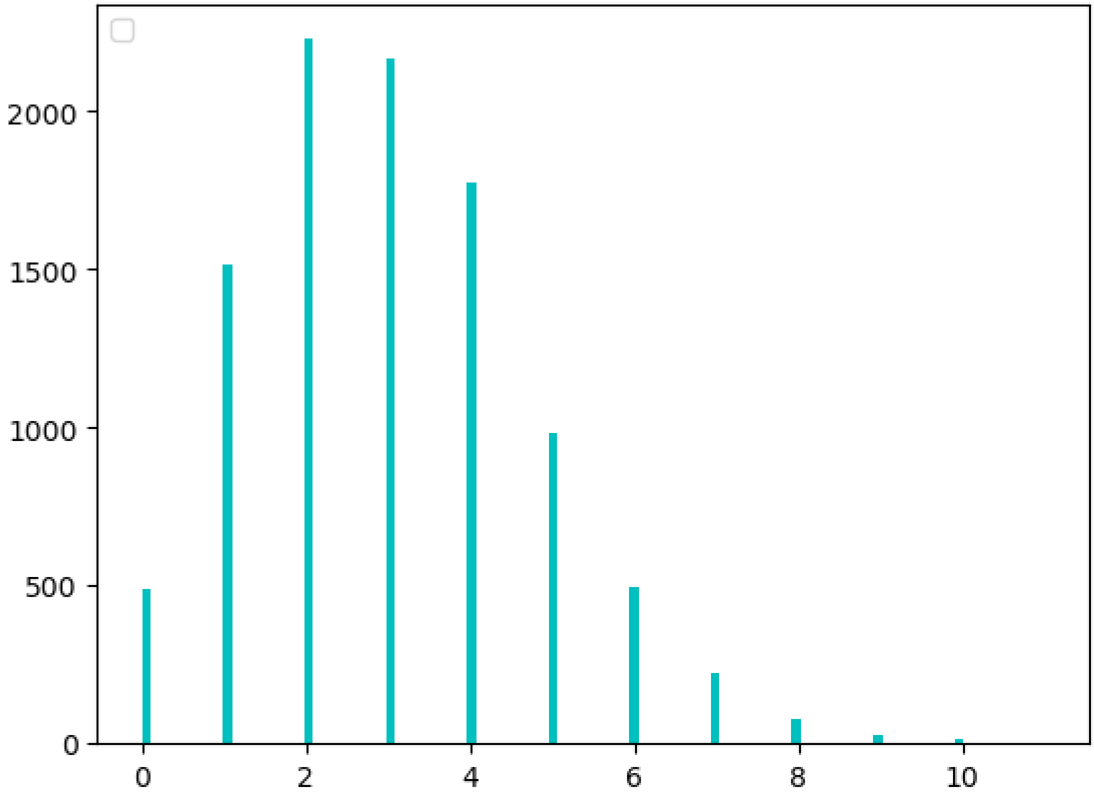


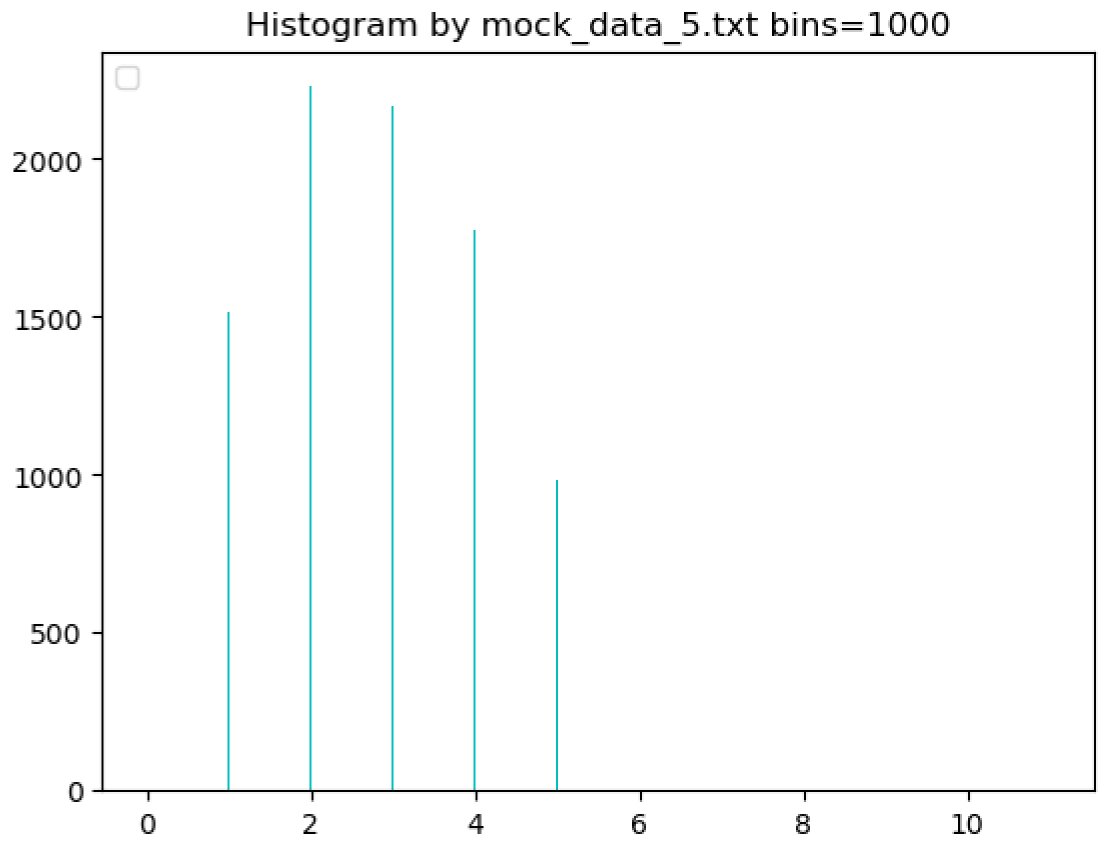
Histogram by mock_data_4.txt bins=1000
50
40
30
20
10
0
0
5
10
15
20

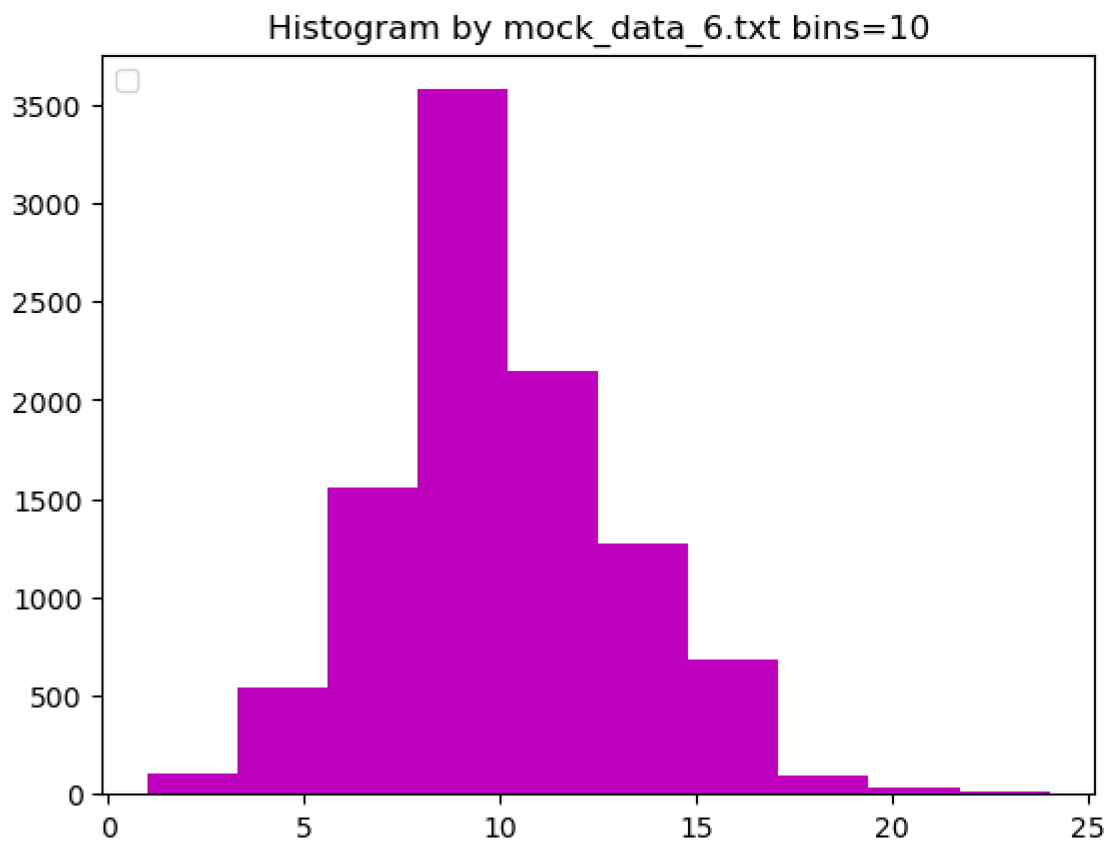
Histogram by mock_data_5.txt bins=10

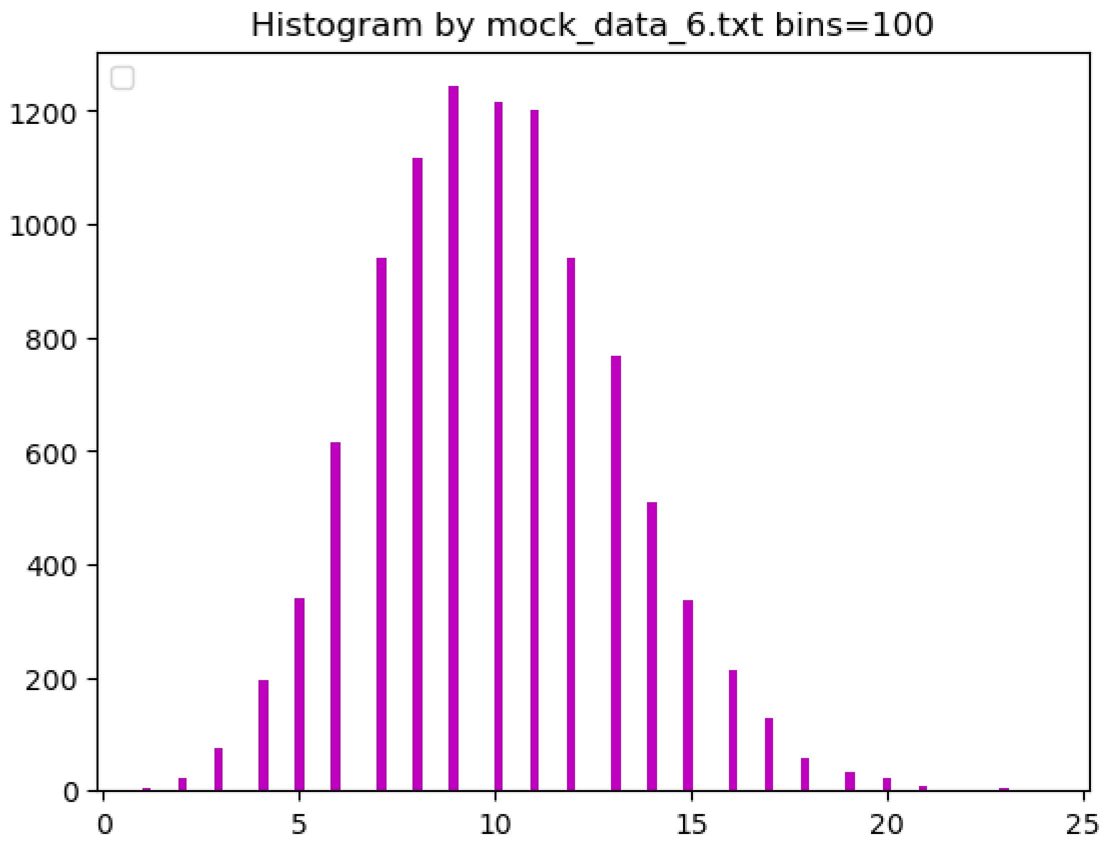


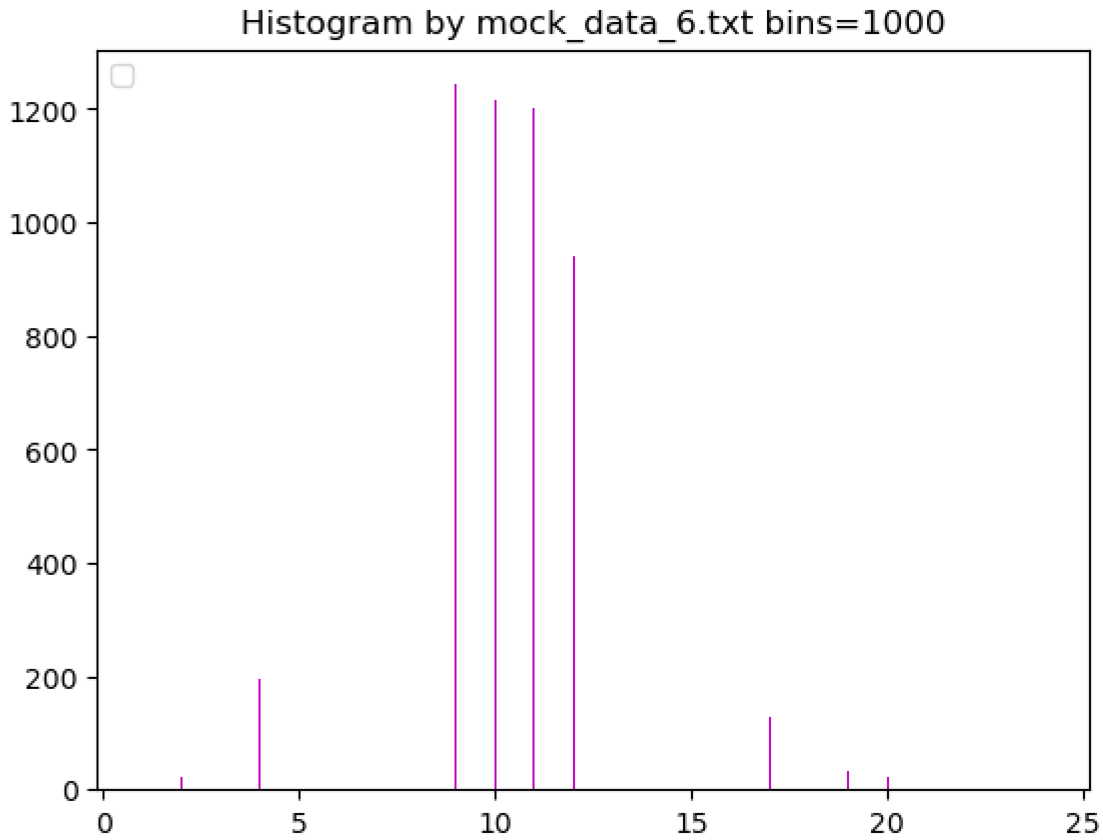
Histogram by mock_data_5.txt bins=100



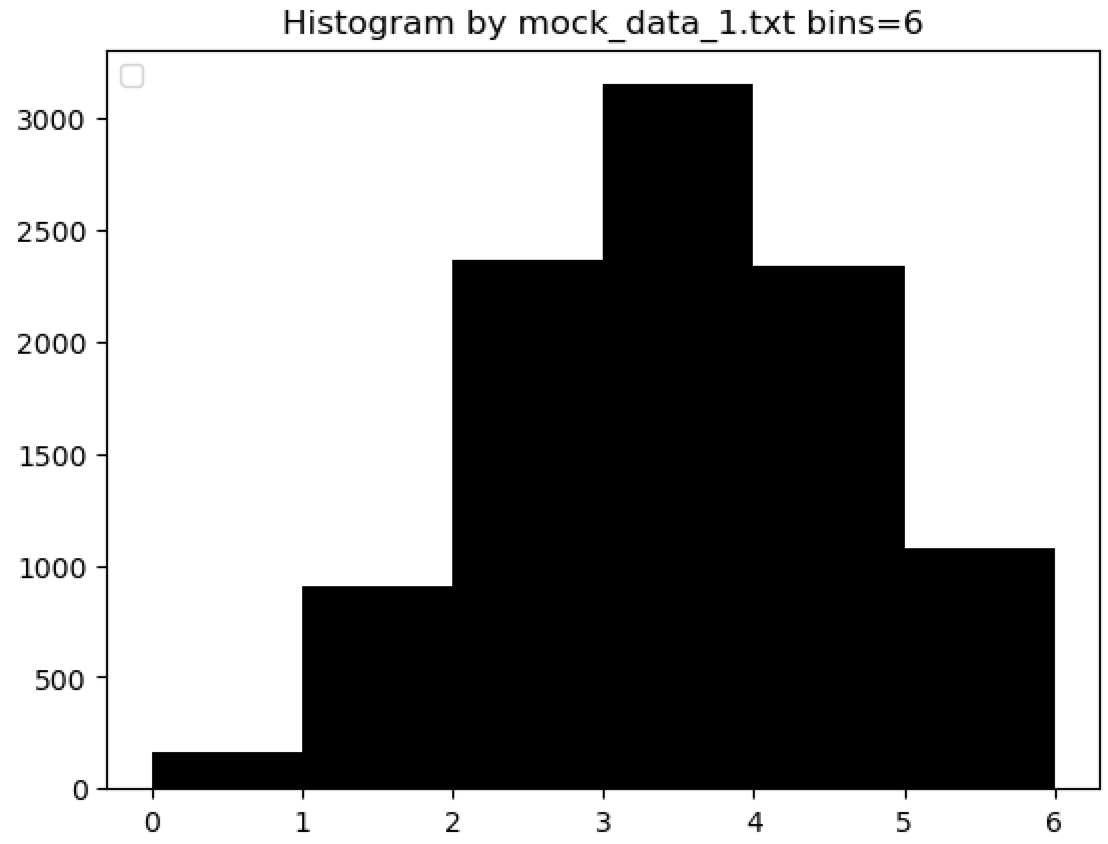

Histogram by mock_data_5.txt bins=1000
0
500
1000
1500
2000
0
2
4
6
8
10


Histogram by mock_data_6.txt bins=10
0
500
1000
1500
2000
2500
3000
3500
0
5
10
15
20
25


Histogram by mock_data_6.txt bins=100
0
200
400
600
800
1000
1200
0
5
10
15
20
25

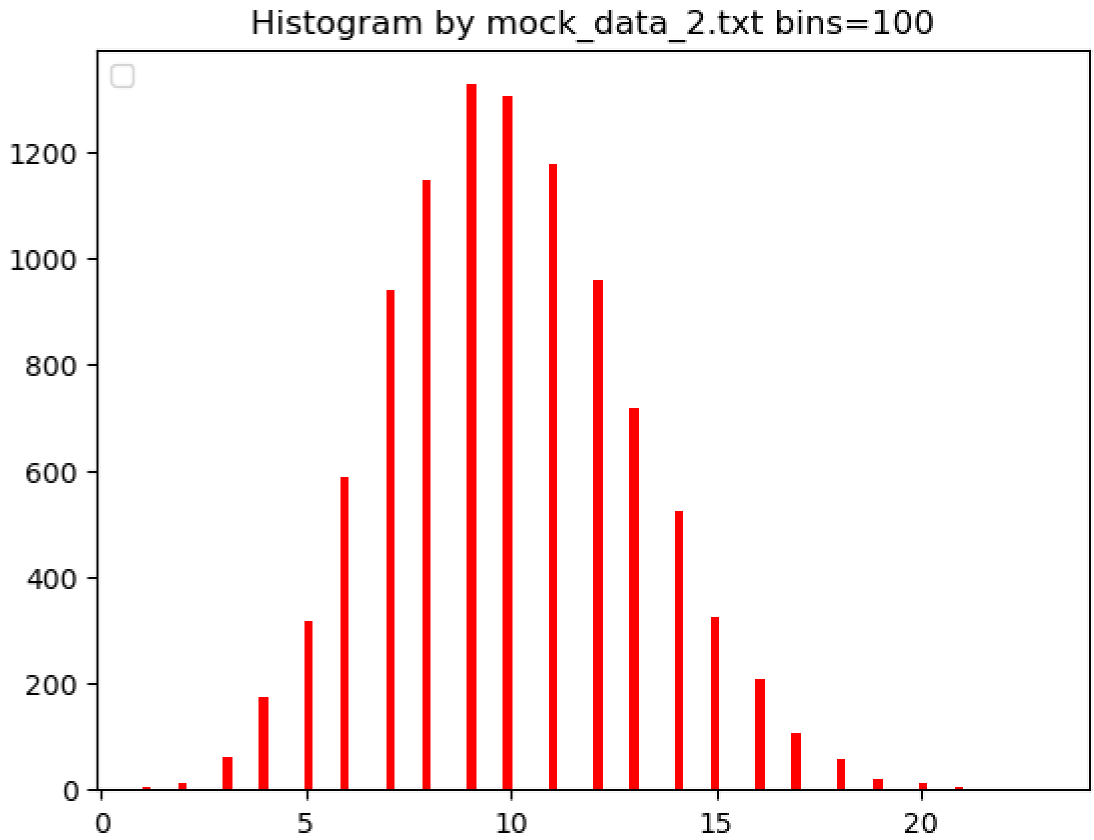

Histogram by mock_data_6.txt bins=1000
0
200
400
600
800
1000
1200
0
5
10
15
20
25

I found proper binsize like this.
In a binomial distribution, the mean value is always greater than the variance.
In a Poisson distribution, the mean and variance are the same.
And when the skewness is 0, it can be said to be a normal distribution.



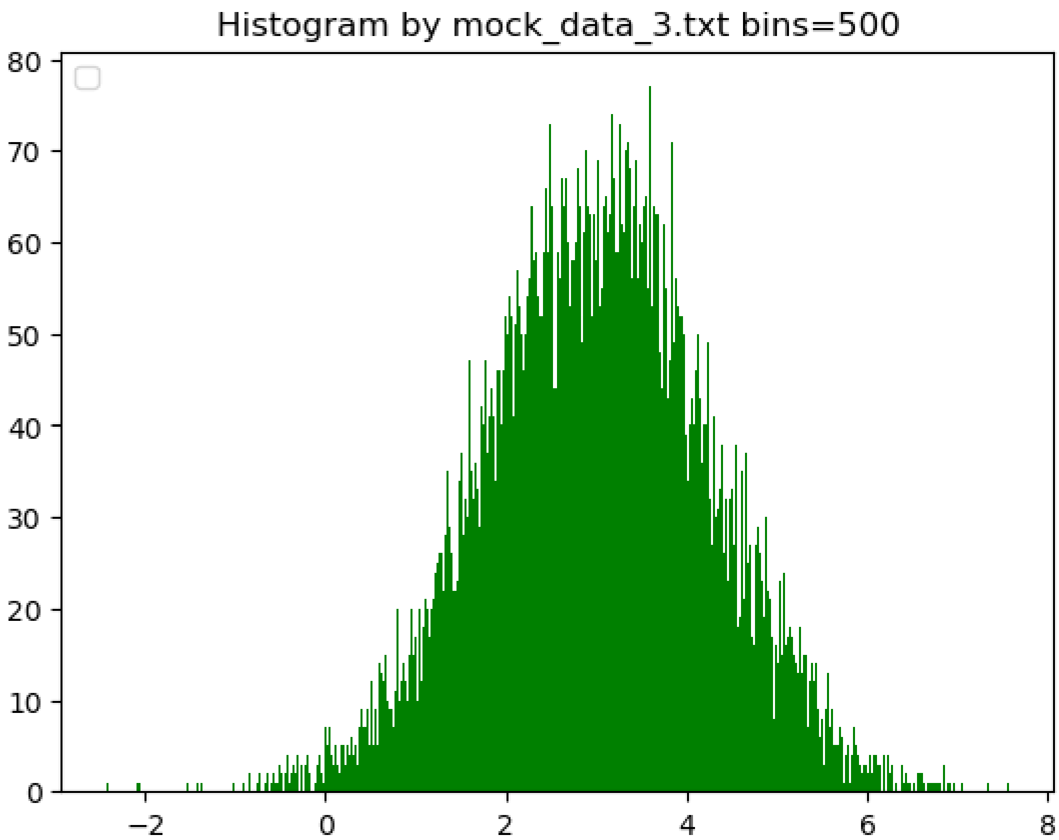
mean: 2.9974
variance: 1.4895932400000274
sigma: 1.2204889348126133
skewness: -0.007580259668532239
kurtosis: 2.700359911030378

-> **binomial**, Poisson, **normal**



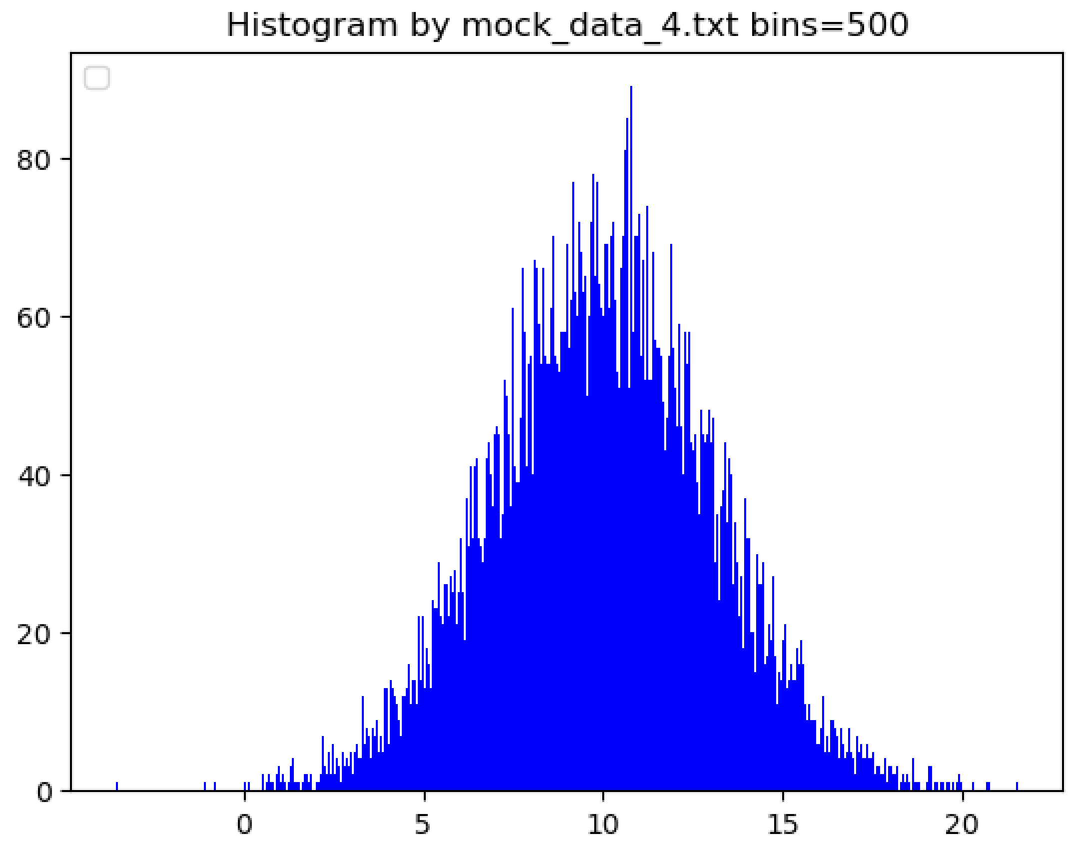

```
mean: 9.9796
variance: 9.051183840000164
sigma: 3.0085185457298023
skewness: 0.27205886498457466
kurtosis: 2.9861741651309344
```

-> **binomial**, Poisson, normal



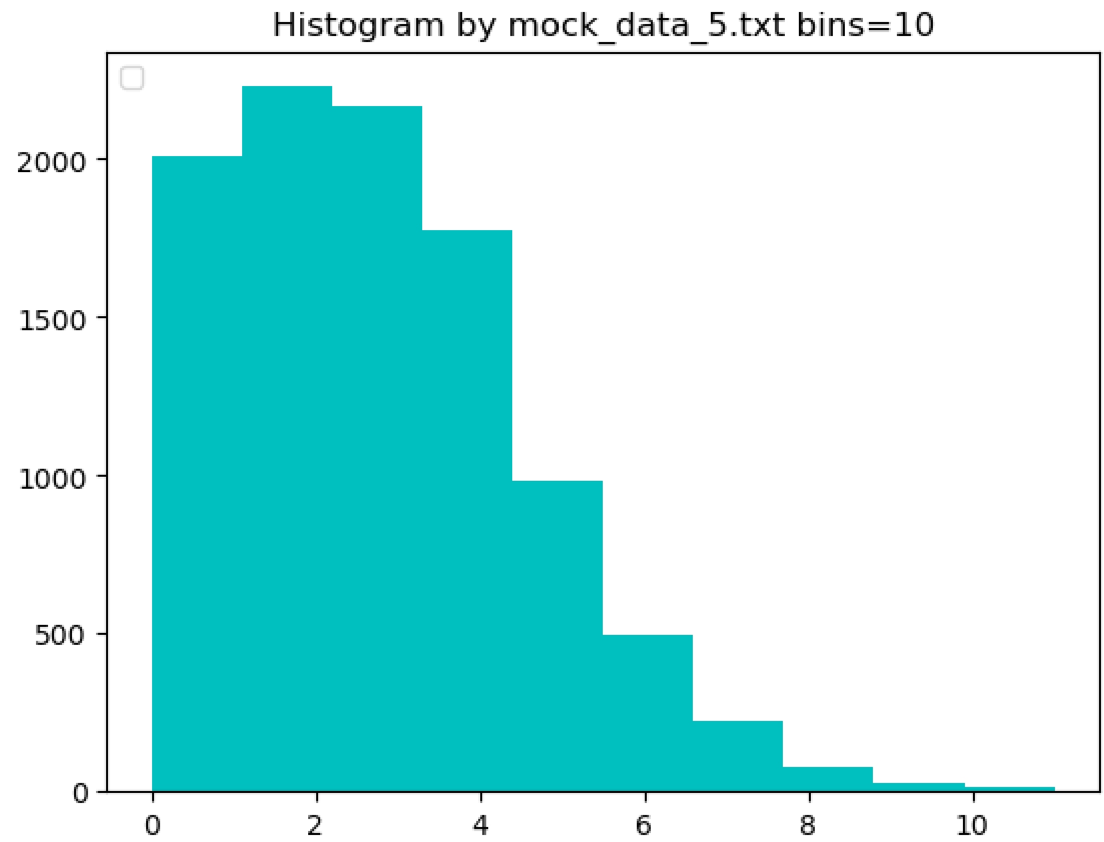

```
mean: 3.0071696146999463
variance: 1.5297783821647475
sigma: 1.2368421007407322
skewness: -0.027915247484694903
kurtosis: 3.0571697317007396
```

-> **binomial**, Poisson, normal



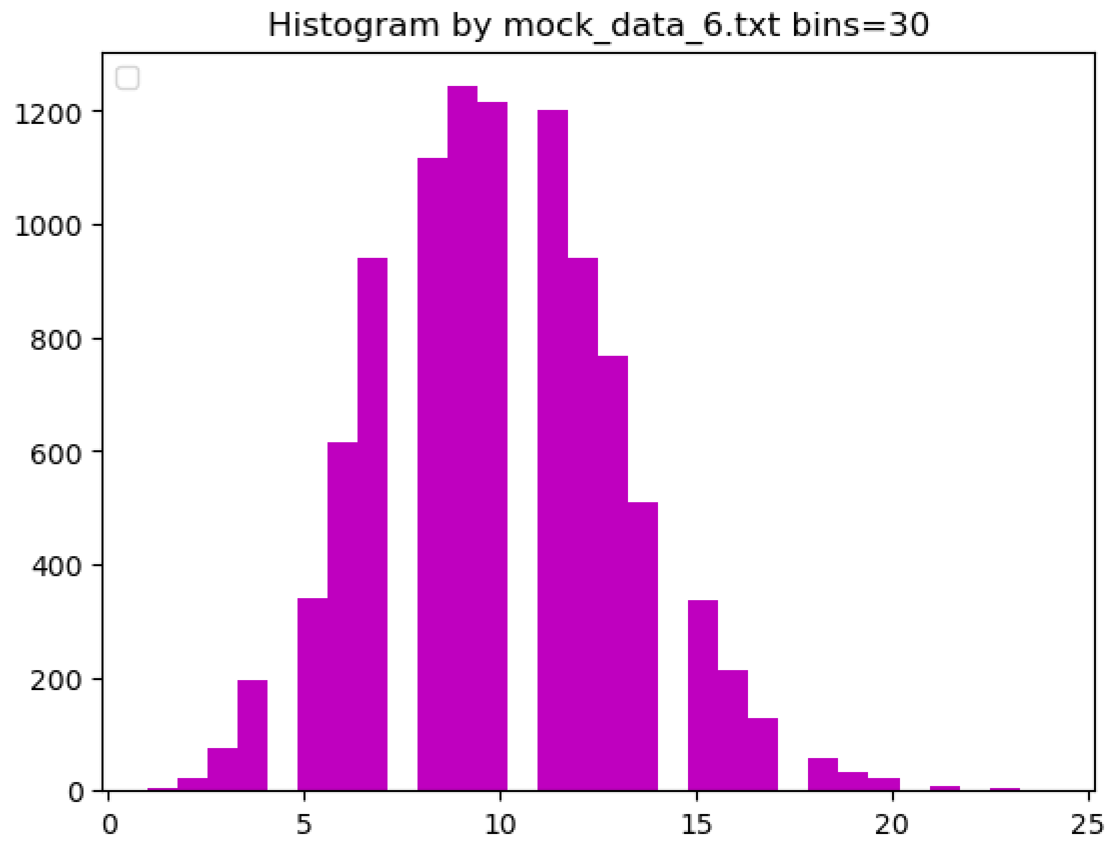

```
mean: 10.0122697691655
variance: 9.20384118425951
sigma: 3.0337833120148034
skewness: 0.006159587863107182
kurtosis: 3.062575878556151
```

-> **binomial**, Poisson, **normal**




```
mean: 3.0048
variance: 3.012776960000062
sigma: 1.735735279355715
skewness: 0.5846608959728767
kurtosis: 3.371672483690904
```

-> binomial, **Poisson**, normal

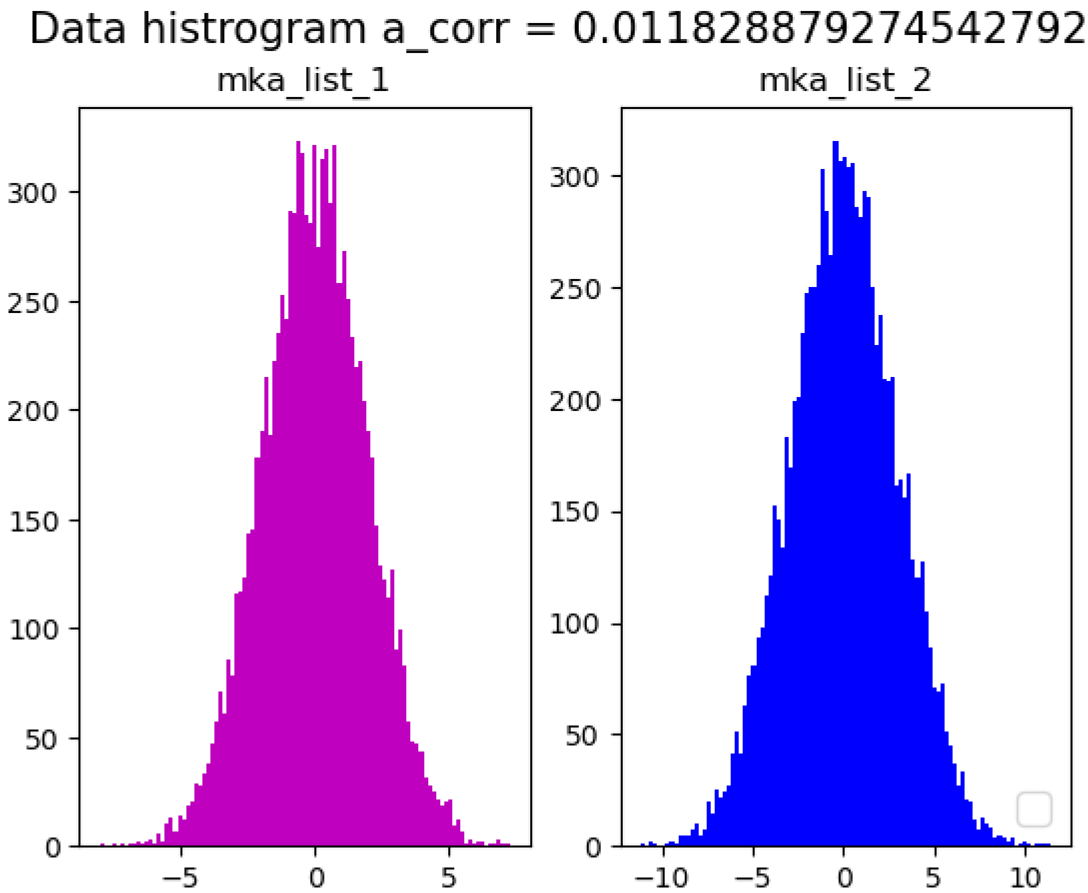


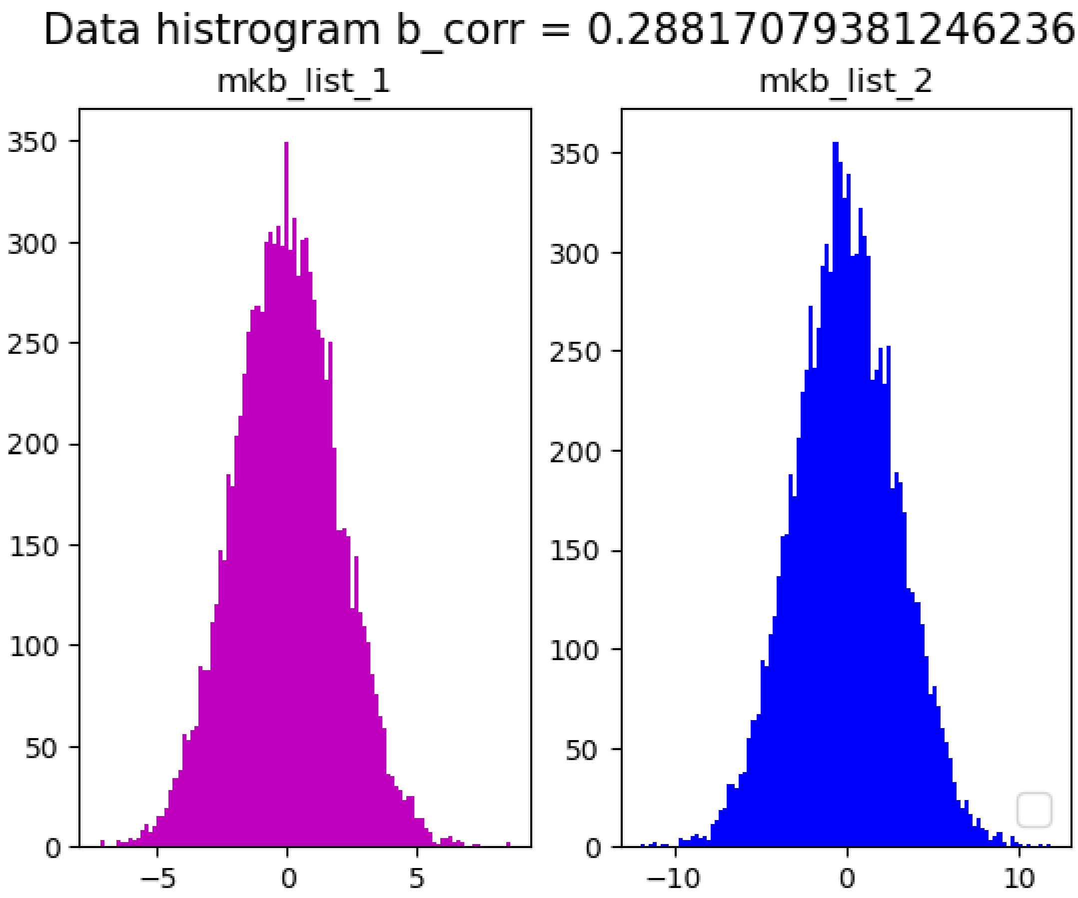

```
mean: 10.0111
variance: 9.849576789999995
sigma: 3.138403541611562
skewness: 0.31703948723652403
kurtosis: 3.1423461861501227
```

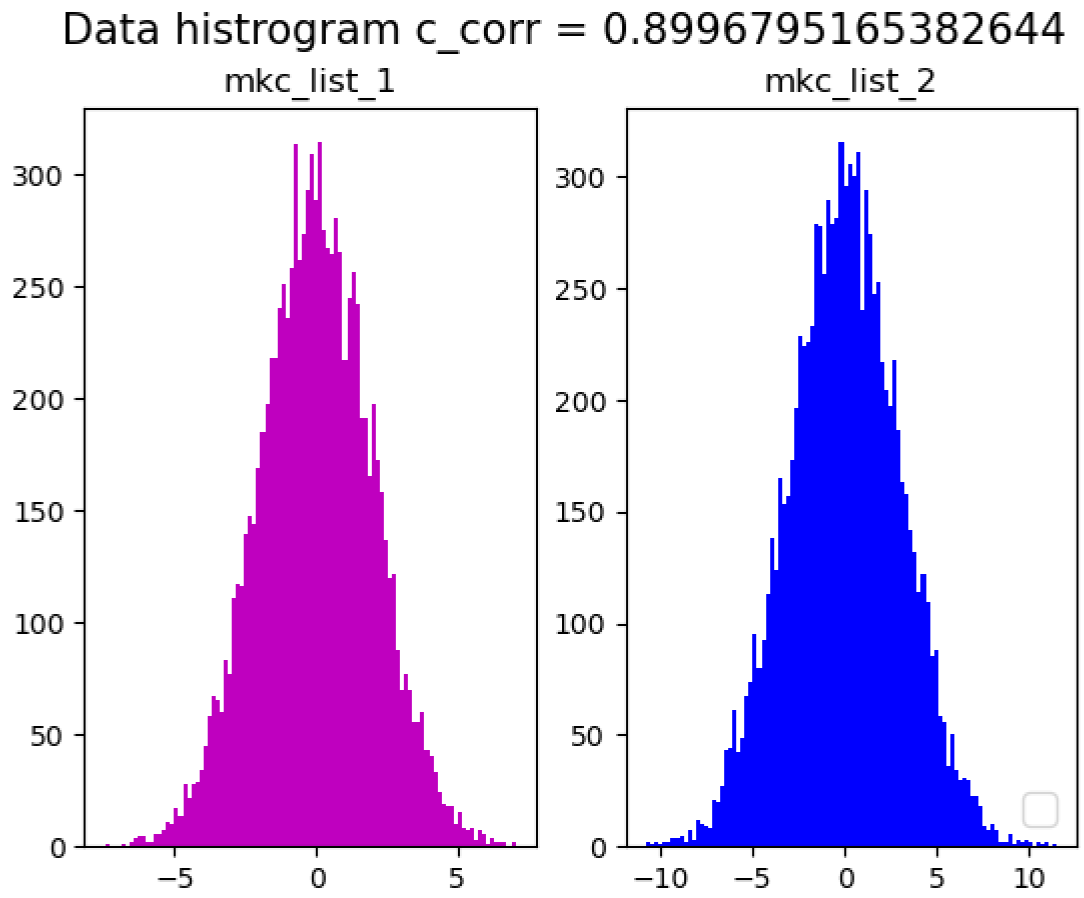
-> **binomial**, Poisson, normal

3.2 Problem 2

For the mock data A, mock data B and mock data C, plot the data and calculate the correlation coefficient.




Data histrogram b_corr = 0.28817079381246236
mkb_list_1
mkb_list_2
350
300
250
200
150
100
50
0
−5
0
5
−10
0
10


Data histrogram c_corr = 0.8996795165382644
mkc_list_1
mkc_list_2
300
250
200
150
100
50
0
−5
0
5
−10
10

3.3 Problem 3

We want to track cosmic rays with some detectors which are 90% efficient. We need to detect cosmic rays with at least 3 detectors to define their tracks. Calculate how often we can detect a track (having 3 detection) using a stack of 3 detectors? How things will change if we use 4 or 5 or 6 detectors?

In this case, we can use the binomial distribution.

P(X = k) = (n choose k) * p^k * (1 - p)^(n - k)

n is the number of trials (in this case, the number of detectors).
k is the number of successes (in this case, 3 or more detections).
p is the probability of success (in this case, the efficiency of a single detector, which is 0.90 or 90%).

If we get 3 detectors: n = 3 k = 3 p = 0.90
P(X = 3) = (3 choose 3) * (0.90)^3 * (1 - 0.90)^(3 - 3)
P(X = 3) = (1) * (0.90^3) * (0.10^0) = **0.729**

or 4 detectors: n = 4 k = 3 p = 0.90
P(X ≥ 3) = P(X = 3) + P(X = 4)
P(X = 3) = (3 choose 3) * (0.90^3) * (0.10^0) = 0.729
P(X = 4) = (4 choose 4) * (0.90^4) * (0.10^0) = 0.6561
P(X ≥ 3) = P(X = 3) + P(X = 4) = 0.729 + 0.6561 = **1.3851**

or 5 detectors: n = 5 k = 3 p = 0.90
P(X ≥ 3) = P(X = 3) + P(X = 4) + P(X = 5)
P(X = 3) = (3 choose 3) * (0.90^3) * (0.10^0) = 0.729
P(X = 4) = (4 choose 3) * (0.90^3) * (0.10^1) = 0.243
P(X = 5) = (5 choose 3) * (0.90^3) * (0.10^2) = 0.027
P(X ≥ 3) = P(X = 3) + P(X = 4) + P(X = 5) = **0.999**

or 6 detectors: n = 6 k = 3 p = 0.90
P(X ≥ 3) = P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6)
P(X = 3) = (3 choose 3) * (0.90^3) * (0.10^0) = 0.729
P(X = 4) = (4 choose 3) * (0.90^3) * (0.10^1) = 0.243
P(X = 5) = (5 choose 3) * (0.90^3) * (0.10^2) = 0.027
P(X = 6) = (6 choose 3) * (0.90^3) * (0.10^3) = 0.001
P(X ≥ 3) = P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) = **1.000**

3.4 Problem 4

Show that for a Binomial distribution <r>=np, V(r)=np(1-p), Skewness=[1-2p1/[np(1-p)]^0.5 and derive its kurtosis too.

3.5 Problem 5

Show that for Poison distribution <r>=\lambda and V(r)=\lambda

3.6 Problem 6

Calculate the mean and standard deviation for the skewness and Kurtosis of the subsets you analysed in Homework 1.

1. mk1_list subsets

skewness_list

[0.09575763322813595, -0.16177283808103388, -0.10972560898072747, 0.0512612504811 1259, -0.12599935871928417, 0.14122203703015895, 0.09983087668837057, 0.002190110148564 5707, 0.12405175635471381, -0.02217250661484894]

mu = 0.0094643351535162, sigma = 0.10525332540973201

kurtosis_list

[2.933979479546409, 3.015858588822157, 2.6653909040907684, 3.091474810026859, 2.8844542102131303, 3.179038627921752, 3.1233245298773853, 3.0850576896986617, 3.210962902545103, 3.029646535931132]

mu = 3.021918827867336, sigma = 0.15276466055968413

2. mk2_list subsets

skewness_list

[0.1256775113989512, 0.10851278475070665, -0.02929698364515778, -0.08711926902096166, 0.07419974465953064, 0.2234355924130936, 0.007730463230614208, 0.02025136276425956, 0.07248166673867679, 0.1245789388595356]

mu = 0.06404518121492488, sigma = 0.08465963446849745

kurtosis_list

[2.917384736093718, 2.8731127957814326, 3.0112792470586336, 2.8960359564319726, 2.836334808216459, 3.2777085418582783, 3.11328753985225, 2.9890796757236084, 2.7288161344525212, 2.845764721071054]

mu = 2.948880415653993, sigma = 0.14896983880876086

3. mk3_list subsets

skewness_list

[0.17015433293519475, 0.2965969497879963, 0.35415205724655435, 0.3130493006853107, 0.25396947484105503, 0.2211234361783912, 0.2241042855603896, 0.20841474962740483, 0.2888345795414705, 0.27918118996106395]

mu = 0.2609580356364831, sigma = 0.05270473065419619

kurtosis_list

[2.8161261786548226, 3.067806147609972, 3.168935577381038, 3.0128084384313234, 2.9941662968791776, 2.8450217050961526, 2.875336729361191, 2.6403572911394093, 3.1146793328096556, 3.287002584913436]

mu = 2.9822240282276176, sigma = 0.18122048330483728

4. mk4_list subsets

skewness_list

[1.3603894543607413, 1.1453658145349797, 1.6742002079310436, 1.2711493942805294, 1.2335040158445711, 1.2010593388935613, 1.3610156175275385, 1.5658697600594647, 1.3498644743368338, 1.4449958076654776]

mu = 1.360741388543474, sigma = 0.1560937100024589

kurtosis_list

[5.69797433439668, 4.809593758015298, 7.689490559018485, 5.036320380968432, 4.995749759633996, 5.0476291669554465, 5.70076039108694, 6.283655341592659, 5.802680935519187, 6.225988522370861]

mu = 5.728984314955797, sigma = 0.8205112635488622

3.7 Questions

- Binomial distribution Skewness, kurtosis.
-