**Probability and Statistics**

-Project 1



|  |  |
| --- | --- |
| 학번 | 2016025469/2016025705 |
| 이름 | 서건식/이도현 |
| 수업 | 확률통계론 |
| 교수님 | 배완덕 교수님 |
| 제출일 | 2017.03.30 |

**The question we are interested in.**

We live in fine dust. So we have always wondered about concentration of the fine dust in the air. Also, we want to analyze the impact of the fine dust on life through various statistical figures. Therefore we decide to collect and analyze data about one years of the density of fine dust In Seongdong-gu.

Also, we wondered about the statistical meaning of data we collected. We want to know this data have reasonable meaning, by collecting and analyzing the sampling data, and comparing w sampling data’s statistical figures with population data’s.

**Technical Aspects.**

** Formal description of the population of interest.**

The population are 365 days in the area of Seongdong-gu, Seoul, from March 28, 2016 to March 28, 2017. We collected data from a competent institution, Climate Management Division of Seoul.

** Formal deﬁnition of the variable you will study.**

The variables we decide to study are the average concentration of fine dust (the size of fine dust is pm-10.) in a day at the area of Seongdong-gu. Because the variables are the amount of dust (㎍/㎥) in the air, these can be measured. (quantitative.)

** Formal statement of the four-step process for data analysis.**

State: Calculate the statistical value of the fine dust concentration for one year, and What do these data mean statistically?

Plan: We should examine one years of fine dust in Seongdong-gu through Graph (time-plot and histogram) and Software tools (Python, Microsoft Excel) for data analysis. Because this is a variable that changes over time, you should also look at a time plot. and using histogram for data analyzing.

Solve: (a) Provide a histogram and time plot (a graph illustrating a return value for 4-days in Seongdong-gu)

(b) Provide the mean and standard deviation of the dataset, and the five-number summary of population and each sample. ( minimum value, quartile Q1, median, quartile Q3, maximum value)

(c) Compare the sample(size 40)’s histogram with population histogram.(The changes about minimum value, quartile Q1, median, quartile Q3, maximum value, etc.)

(d) Examine the Similarity of statistical data between population group and sample group.

(e) Using The means of many sample group (sampling many times), Make the histogram and determine whether the mode value of the histogram is similar to that of the population(check center tendency).

Conclude: Provide a conclusion from analyzing the graph and data that we collect

** Your plan for selecting a simple random sample from your population.**

We collect random sample by using the random modules in Python and make function for extracting samples and not making duplication.

** Formal statement of the data analysis using mean and standard deviation of the sampled data.**

Examine the relationship between sample and population statistical figure according to number of times samples were extracted (each times is 20, 40 times.)

Examine the relationship between sample and population statistical figure according to size of sample group ( each size is 20 , 40)

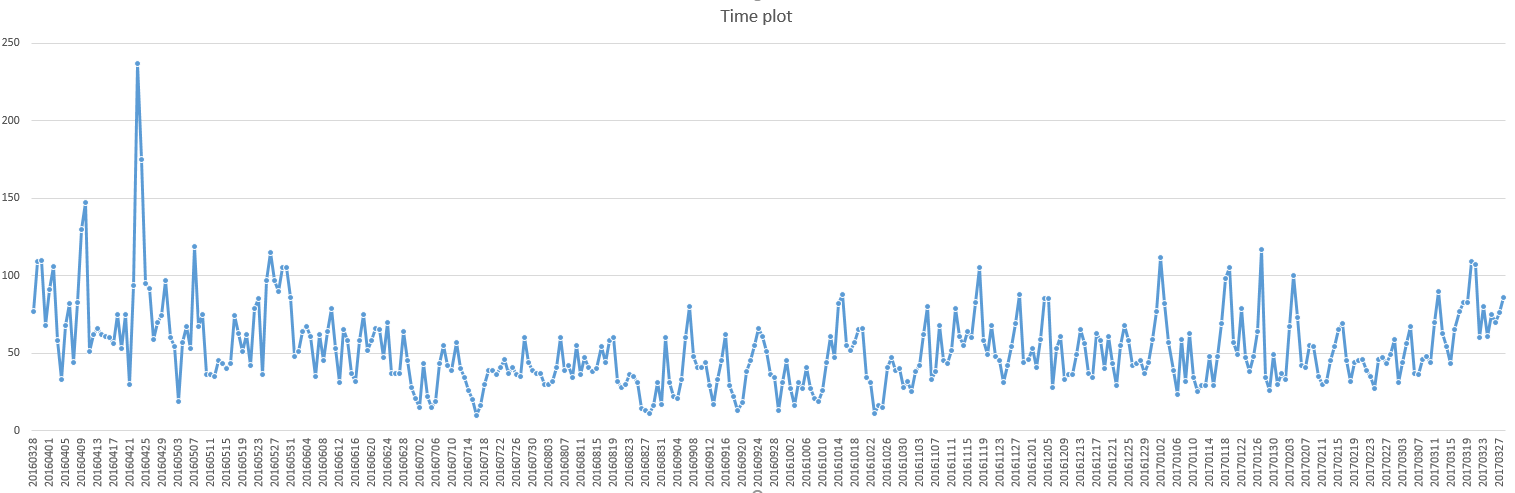
** The summary of the observatory study**

1)Data (모집단)

(a) Data is collected by competent institution, Climate Management Division of Seoul (<http://data.seoul.go.kr/openinf/sheetview.jsp?infId=OA-2218>) and we collected sample(random variables) by using random module in python. Comparing sample with population is reasonable because sample have enough random variable(40) and they have similar hisogram.(the evidence exists in later paragraph)

(b) We provide separate excel file.

(c)



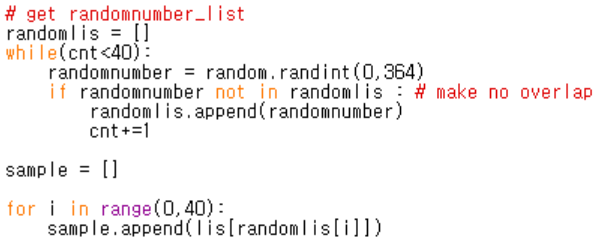
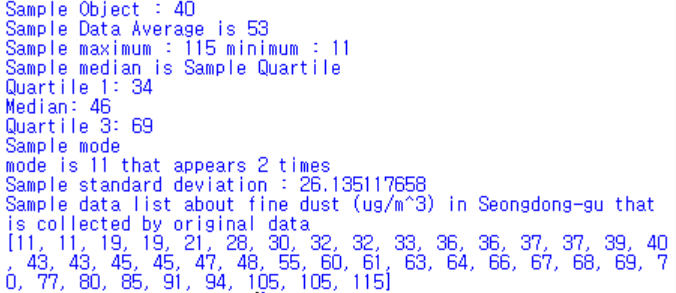
When we see the time plot(one-years) we think our variable is not related well to season except outlier (2016.04.25). so we plan to compare our variable by one-year.

2)

State : Calculate the statistical value of the fine dust concentration for one year, and What do these data mean statistically?

Plan : (a) We collected data form Climate Management Division of Seoul and we make program that collected random sample from data by using random module in python.

(b) And to analyze data of sampling , population data, We plan to use the excel program to get time-plot and histogram



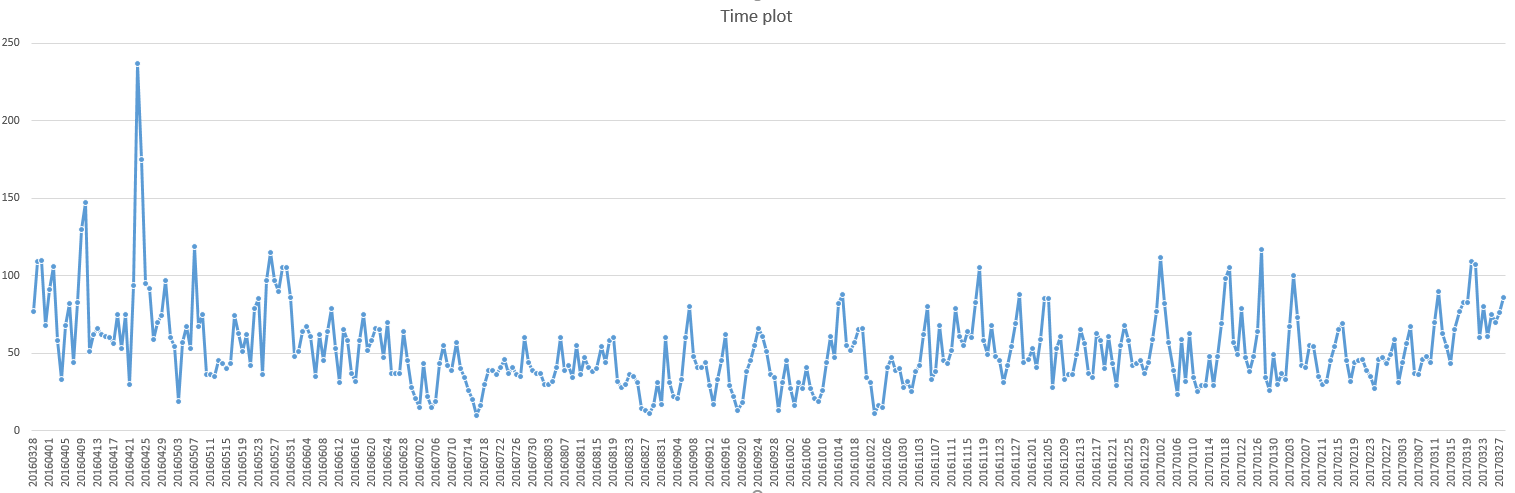
->

Solve :

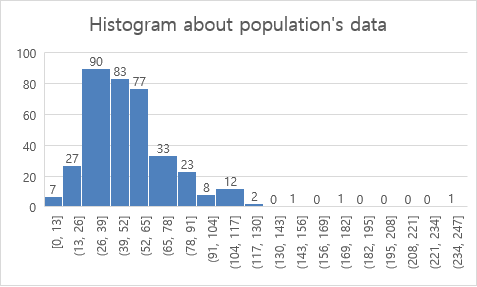
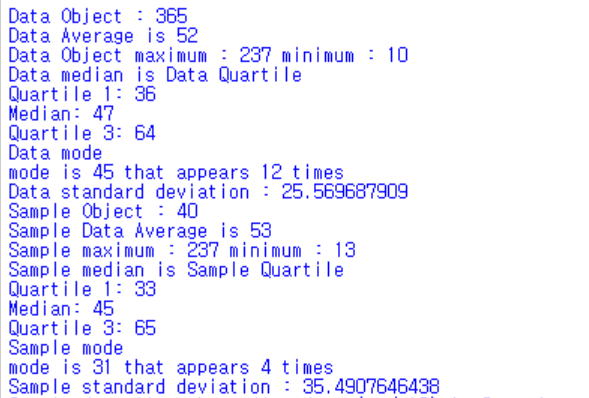
1. Find statistically information about population.

In timeplot , unit of variable is ㎍/㎥.

We can find change of variable about time in time-plot. And we find data’s overall pattern is not related to season well.



In histogram, we find asymmetric(Right skewed).



(b) We calculate our data by making python program.

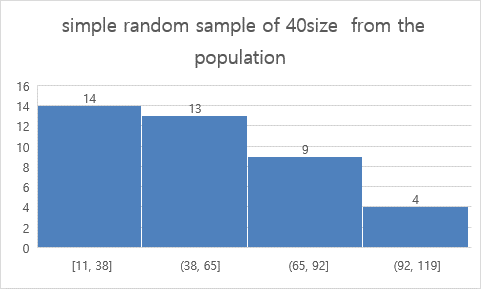
In program, we can find mode is 45, median is 47, average is 52.2857

It appears that our data is right-skewd(mode < mean < average)

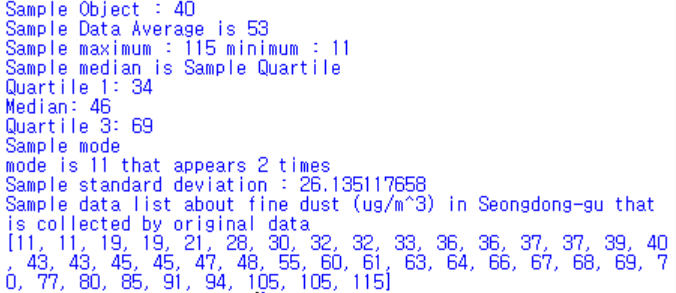
In program, data average is 52, maximum is 237, minimum is 10, quartile 1 is 36, Median(quartile2) is 47, quartile 3 is 64, mode is 45 and standard deviation is 25.5696.

(c)

Next, we make sample (40 random variables) by using python program.



In histogram, Because of sample size it is hard to define but we can find it is similar to population’s data histogram and it seems like right-skewed.



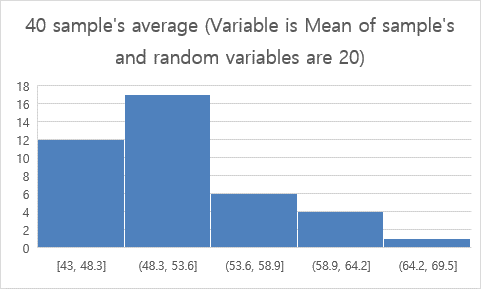
(d)

In data, we find sample size (size of random variable) is 40, average is 53, Quartile 1 is 34, Median(Quartile2) is 46, Quartile3 is 69 and Mode is 11.

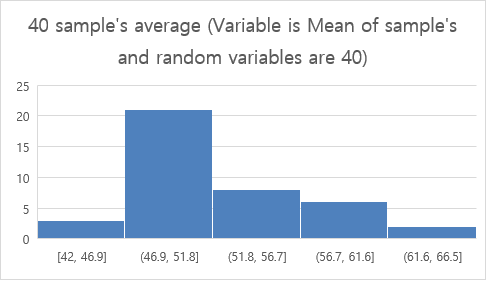
We can find information that sample’s histogram graphic is similar to population’s histogram graphic. It’s average, Quartile(1,2,3) is similar to population’s. and sample’s standard deviation is higher than population’s standard deviation.

(e)

we’re sampling the data many times, and make histogram (variable is ) )and we find change of histogram according to number of sampling and number of sample size(random variable size)

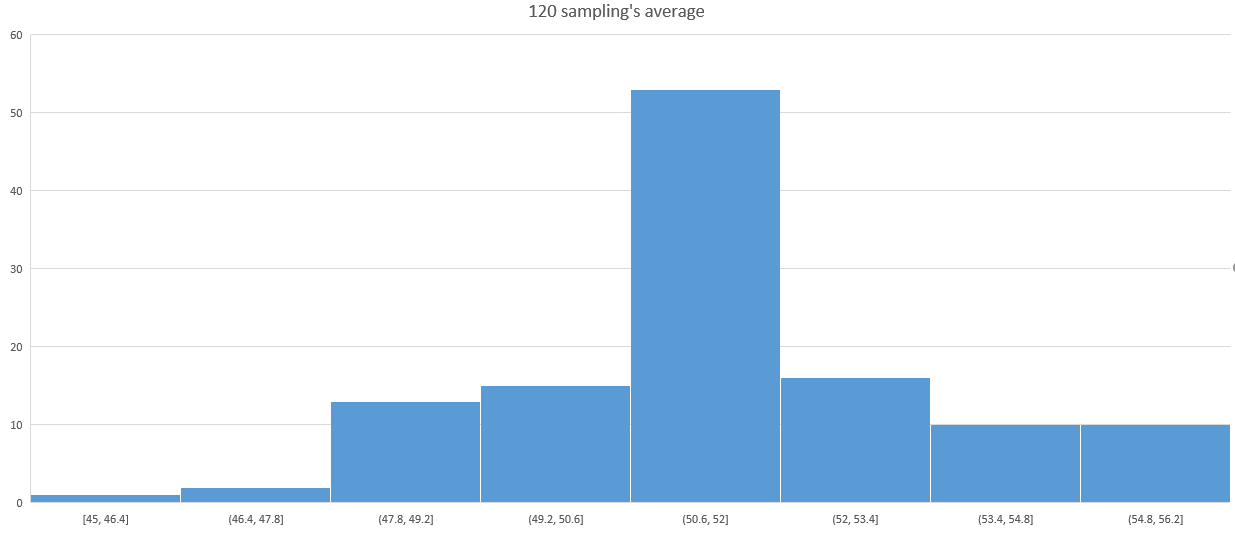


This histogram’s mean is 51.275



This histogram’s mean is 52.83

And make histogram (Variable is Mean of sample’s and random variables are 120)

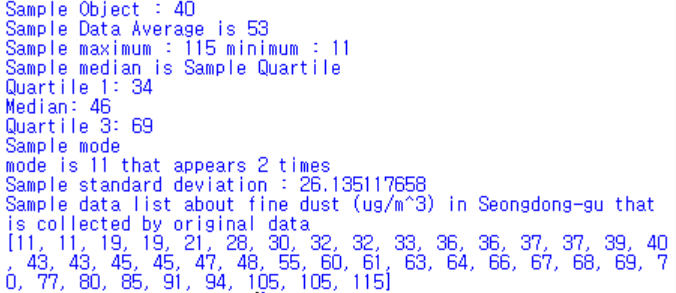


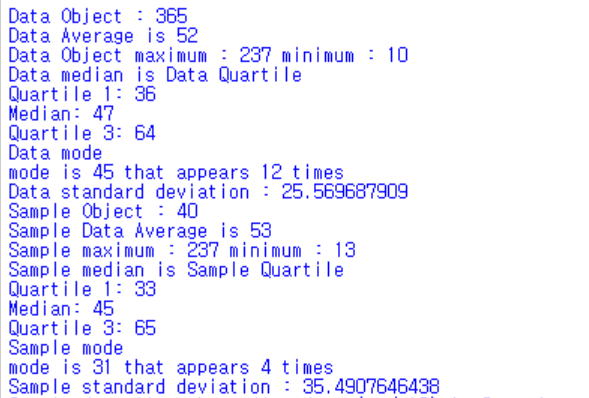
This Histogram’s mean is 51.88333.

From these histograms, we can find if we select more samples and more random variables, In statistics, we usually get information E( ) is near to population’s mean(52.2857).

Conclusion : If we extract sample of moderate size , histogram, mean of population is similar to sample of that. and when making the histogram using sample's mean, The more the number of sample group have, the more histogram's mean made by sample mean is similar to mean of population. We also conclude that the more number of sampling can make more symmetrical histogram with loss outlier.

4) sample calculate





-End.