**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 with sampling data’s statistical figures and 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 every day 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

(d) Compare the change according to size of sample with population. (The changes are minimum value, quartile Q1, median, quartile Q3, maximum value, etc.)

(e) Examine the Similarity of statistical data between population group and sample group according to size of sample and number of times samples were extracted

(f) 여러 번 샘플링하여 나온 표본집단의 평균들을 히스토그램화 시켜 그 히스토그램의 최빈값이 모집단의 평균과 유사한지를 판정한다.

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)

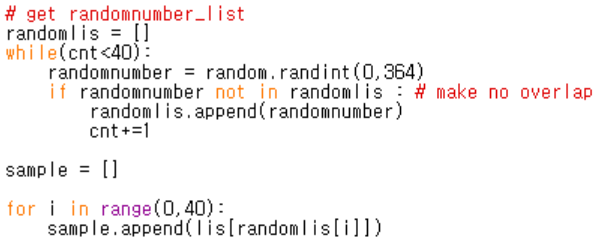
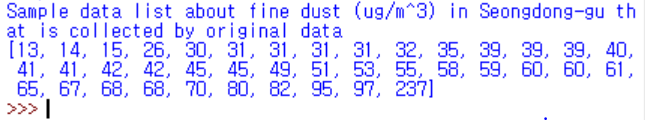
(a) Data is collected by competent institution, Climate Management Division of Seoul. And we collected sample by using random module in python.(여기서 reasonable한지를 묻는게 sample이랑 data를 비교하는게 타당한지를 물어보는 것인지 sample을 reasonable하게 뽑은걸 말하라는건지 잘 모르겠다.)

(b) We provide separate excel file.

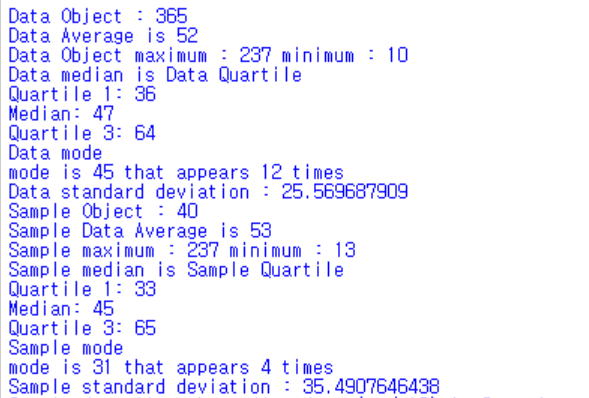
(c) 여기에 histogram 혹은 time – plot을 그리자./ time plot을 먼저그린다음, time plot에 의해 미세먼지가 계절과 크게 연관이 있나 없는가를 분석하고 histogram의 기간을 설정하자. 라고 쓰자

2) 여기는 우리가 위에쓴 plan solve conclude같은거 실행한결과 쓰는듯

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



->

Solve : (a) 여기에 그래프그리기(히스토그램 혹은 timeplot)

(b)

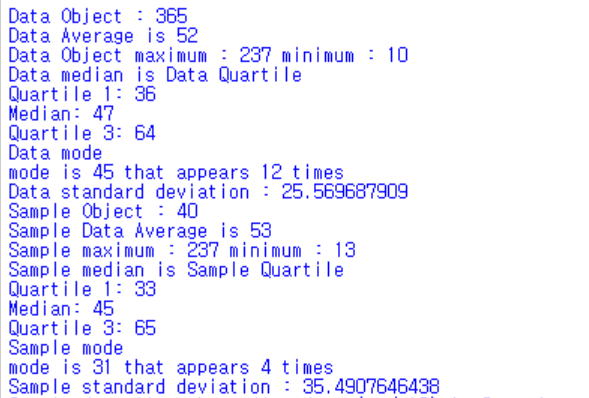
Conclusion : 여기는 우리가 그래프 그리고 sample하고 다 분석, 비교해서 우리가 조사했던 결과에대한 우리의 결론을 내리는 것인듯.

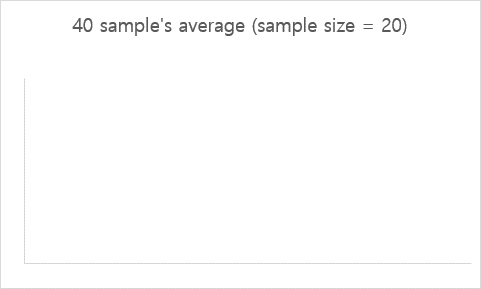
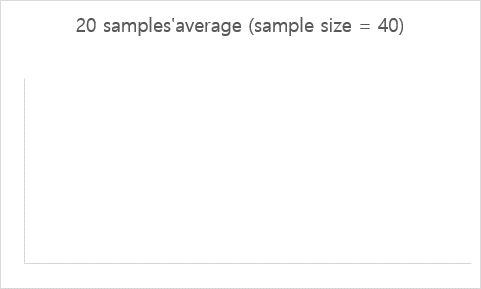
3) 우리가 그린 히스토그램 이 인구가 정규분포에 근사한것인가(?)

skewness or outlier에 대한 표시0

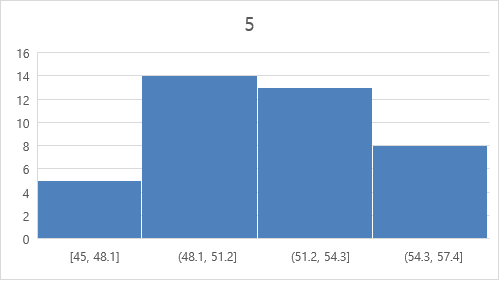
data가 대칭적인가 no outlier인가(?)

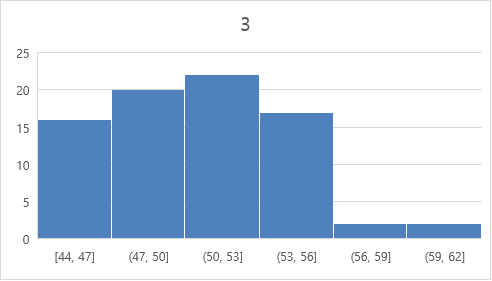
4) sample calculate한 것 그리는곳

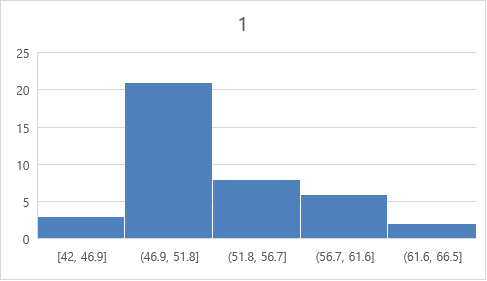












5) Report in one or two sentences whether or not you reject the null hypothesis and what this means about your population. 맞다 여기하지말래,