**Air pollution**

Author: Jose Manuel Alvarez Gonzalez

Email Id: jmanuel.ag89@gmail.com

Date: 13/3/2022

**Introduction**

For the following project, we will continue to analyse the data set of “airpollution” from the first part of the project. For this data set, we have 12 different measurements in 80 different cities around the USA.  
  
The first stage of this project is to understand the descriptive statistics of each variable, we will statistical summaries, complemented by visual tools from histograms and box plots.  
  
Then we will continue with our analysis using the central limit theorem and hypothesis testing, we want to understand the population behaviour out of the 80 different samples we have. For this stage of the process, we are going to be focusing on three different variables: TMR, PMAX and SMAX.  
  
From the last three variables, the goal is to understand their distribution, find average values in the population and clearly identify any similitudes or differences between them.

**Methodology and Motivation**

For the analysis of this project, we are going to break it apart on four different stages for the analysis and the procedures we will be using:

**- Descriptive Statistics and Exploratory Analysis**

In this stage of the process, we are working with the 12 variables. For this part, we will be identifying the measurements of centrality and dispersion.  
  
To complement we are using visual tools like histograms and boxplots, then we will use different statistical measurements. The goal is to be able to have clear measurements and descriptions of the variables.

**- Normal Distribution and Central Limit Theorem**

Here we are going to be focusing on the TMR. We will identify if the variable is normally distributed, then we can apply the Central Limit Theorem (CLT).

The reason of using the CLT is to use some probability and statistic, to find the likelihoods of the variable having different measurements around the population, this process creates percentages of different case scenarios for hypothetical values of the variable.

**- Estimations and determining the population mean**

Here we are working with the PMAX and SMAX variables, from here we will be creating five different samples of size 30.  
  
The reason of doing this samples is to work with the CLT and create some estimations about what would be the averages of the population, we will be working with different values and try to find the probability that the averages of the variables are under or over a certain parameter, creating a confidence interval for them.

**- Testing Claims and Hypothesis Testing**

In the final stage of the project, we will be formulating some final questions about the main three variables. First we will compare PMAX and SMAX with each other, also which has a higher value. For the TMR we will estimate the average value in the population and find if it falls under a certain range.  
  
The reason of doing an hypothesis testing is to have statistical and numerical reasons to support or deny our findings with a 5% or 10% margin of error.

**Results**

**- Descriptive Statistics and Exploratory Analysis**

First let’s look at the Histogram of each variable:

Chart

Description automatically generated

Figure Histogram for each of the 12 variables of the data set "airpollution"

The idea of the histogram is to see how each variable behaves and to try to find any symmetry, we can see that some of the variables have one peak and do look relatively symmetrical.

Looking at the boxplot on figure 2, we can see that all of the variables except “NONPOOR” have outliers present. This is an important stage in trying to use measures for centrality and dispersion.

**Chart, box and whisker chart

Description automatically generated**

Figure Boxplot for each of the 12 variables of the data set "airpollution"

For the next part we are going to create a descriptive statistics of each of the variables:

**Table

Description automatically generated**

Figure Descriptive Statistics for the variables of data set "airpollution"

Now we can back up our ideas, first and most important finding is the fact that except for one, all the variables have outliers. The effect of the outliers is that they skew the shape of the distribution towards one side or the other. We can back up that by looking at the difference of mean vs median, when the data has higher end outliers, it will be right skewed and the mean > median.

From that we can describe the shape of our variables on this table:

Table

Description automatically generated

Figure Table for shape description of variables in data set "airpollution"

Next we have our measures of centrality and dispersion for each variable, as we have outliers present, we will have to be really strict about the skewness. If any variable has an absolute value of Skewness higher than .5, we will use median and IQR, otherwise it will be mean and SD.



Figure Table for Centrality and Dispersion measurements for variables of data set "airpollution"

**- Normal Distribution and Central Limit Theorem**

The first stage is to apply a normality test to find if variables are normally distributed.

For that we do a Shapiro Test and we have the following findings:

Table

Description automatically generated

Figure Table for normality test

For the TMR variable we can apply the CLT from a normal distribution, from here we can see that:

P(TMR > 1200) = 2.68%

P(TMR < 783) = 15.83%

P(TMR < 691.44) = 5%

**- Estimations and determining the population mean**

For this part of the report, we are creating five different samples of size 30 from our variables PMAX and SMAX.

Looking at the descriptive statistic of the samples, we have the following:

**PMAX**

Table

Description automatically generated

Figure Table for descriptive statistics of PMAX samples

**SMAX**

**Table

Description automatically generated**

Figure Table for descriptive statistics of SMAX samples

From the CLT we can use the theorem:

Calendar

Description automatically generated with medium confidence

As we already have the measurements of the parent population, we can use them with the samples to make our parameters:

The parameters we have for a sample size 30 of PMAX: 275.54 29.05

The parameters we have for a sample size 30 of SMAX: 219.88 21.92

Then from the first sample, we can create a Confidence Interval at 95%, we follow the equation:

CI at 95% for PMAX = [170.94, 284.8]  
CI at 95% for PMAX = [183.84, 269.76]

At 99%

CI at 99% for PMAX = [152.93, 302.81]  
CI at 99% for PMAX = [170.26, 283.34]

If we ask ourselves what will be the likelihood that from a sample, the average size of PMAX is less than 150, we have . 0007737417%

**- Testing Claims and Hypothesis Testing**

In the final stage of the analysis, we want to see if is no difference between the variables PMAX and SMAX. First we can see that visually with a boxplot of the two variables:

**Chart

Description automatically generated**

Figure Boxplot for PMAX and SMAX highlighting their mean values

Then we are going to do an independent two sample T-Test. For our null hypothesis we stipulate that is no difference between the averages of PMAX and SMAX

H0 : μPMAX = μSMAX   
H1 : μPMAX μSMAX

We want to do our test at a 5% significance level. After running our test we get the following P-Value:

P-Value = .015 < α = .05

We can conclude that at a 5% significance level is not enough statistical evidence to support the null hypothesis, saying that is no difference between the averages for the variables PMAX and SMAX

The new alternative hypothesis we have says: the average of PMAX is greater than the one of SMAX

Ha : μPMAX > μSMAX

The significance level we have is 10% with the value of α = 0.1

We create a difference interval from PMAX – SMAX and we get:  
The interval is (18.57, 92.74)  
  
We can conclude that, we are 90% confident that PMAX is greater than SMAX between 18.57 and 92.74 on average.

Next we want to know if TMR is lower than the value of 1,000 on average. For that we have our hypothesis:

H0 : μTMR = 1,000   
H1 : μTMR 1,000

The significance level we have is 10% with the value of α = 0.1. For that we can find our critical T-Value equal to a confidence interval of 90%. We use the formula and we get:

CI at 90% for TMR = [898.99, 951.99]  
  
We can conclude to say that at a 10% significance level, is enough statistical evidence to support the alternative hypothesis that on average the TMR is less than 1,000.

**Conclusions**

After doing different tests to our variables, we can see that we are likely to find outliers across the population. To communicate the centrality and dispersion of the variables, we are better using the median and the IQR, this will be a better interpretation as a whole and less affected from the outliers.  
  
We were able to find that the TMR behaves normally distributed around the total population, this is a really helpful finding to be able to make predictions and the probability of having an exact measurement. We identified that the lowest 5% of values are under the value of 691.44 and only less than 2.68% will have a value higher than 1,200. Then we tested that we are 95% confident, that on average the TMR values are between 898.99 and 951.99  
  
For the variables PMAX and SMAX we were able to find their average parameters on the population with a confidence rate of up to 99%. At the same time we identified with a 95% confidence that on average they have different values and we are 90% confident that PMAX is on average 18.57 and 92.74 higher than SMAX.

**QUESTIONS**

**Use R to produce a histogram of all the variables**.

Text

Description automatically generated with medium confidence

Chart

Description automatically generated

**Use R to produce Descriptive Statistics for all the variables.**

**Graphical user interface, text

Description automatically generated**

**Table

Description automatically generated**

**Use R to produce boxplots describing the variables side by side. This should be one picture.**

**Graphical user interface, text

Description automatically generated**

**Chart, box and whisker chart

Description automatically generated**

**Using your output from (a) to (c), comment on the shape of the distribution for each variable. In particular, briefly describe in a table form:**

Table

Description automatically generated

**Which measures of central tendency and dispersion are the most appropriate to numerically summarise the data? For full marks, justify your choice of measures, interpret the corresponding values and display them in a table.**



**Use R to test the variables for Normality. Briefly describe whether the data follows a Normal Distribution.**

**Graphical user interface, text, application

Description automatically generated**

Table

Description automatically generated

**Based on part (f), consider the variable TMR that is approximately Normally distributed. Use R to calculate the probability that the TMR will be more than 1200.**

**Text

Description automatically generated**

****

P(TMR > 1200) = 2.68%

**Based on part (f), use R to calculate the probability that the TMR will be less than 783.**

**Text

Description automatically generated with low confidence**

****

P(TMR < 783) = 15.83%

**If a city measures in the lowest 5% of the cities, what is the required TMR level? For full marks, show all your working out by hand, provide a correct probability statement and include the R output to verify your answer.**

We need to find the X value that gives: P(TMR < X) = .05

We have the TMR values of

To get the X value we can use the equation:

We know that P(z < -1.6449) = .05, we can use Z = -1.6449 for the equation to get the lower 5%

(-1.6449 \*142.2672) + 925.45 = 691.44

TMR required is < 691.44

Text

Description automatically generated

**By using R, produce five random samples of size 30 by randomly selecting 30 values from the PMAX variable. Repeat for SMAX variable. For full marks, provide a screenshot of your samples in a table format.**

Table

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Table

Description automatically generated

**Use R to produce the descriptive statistics for each sample and store the information in another table, please ensure you state the mean and standard deviation of each sample. Determine the sampling distribution of means for PMAX and SMAX and state the parameters based off your first sample? Justify your answer, quoting any theorems you used.**

**Text

Description automatically generated**

**Graphical user interface, text, application, table, Excel

Description automatically generated**

**Graphical user interface, application, table

Description automatically generated**

For our sample we have n > 30, we can happily apply the central limit theorem. For that we know:

Calendar

Description automatically generated with medium confidence

From the CLT we can see that each sample will slowly create a normal distribution that recreates the mean of the parent population. The size of the samples will increase or decrease the parameter of it.

Base on the fact that the samples were drawn out of the original variables, then the original variables will be consider the parent population. For this we know the mean and standard deviation of the variables from the parent population.

The parameters we have for PMAX:

Sample Distribution Mean = 275.54  
Standard Error = 29.05

Text

Description automatically generated with medium confidence

The parameters we have for SMAX:

Sample Distribution Mean = 219.88  
Standard Error = 21.92

A picture containing text

Description automatically generated

**Calculate the probability that the average PMAX is less than 150 based on your sampling distribution of the means from part (k). For full marks, show all your working out by hand, provide a correct probability statement and include the R output to verify your answer.**

We need to find P(PMAX < 150) = X

We can start by using our Z score and find its value, for that we know:

*275.54  
z = (150 – 275.54)/ 29.05  
z = -4.32*

*P(z<-4.32) = 0.000007737417*

Text

Description automatically generated

Text

Description automatically generated with low confidence

**Construct and interpret the 95% confidence interval for the population mean for PMAX and SMAX based on the sample data. All calculations should be done *manually* without using R, however, use R to visualise and verify the results.**

For a 95% CI we use Z = 1.96

From the sample we have the values:

**PMAX**

**SMAX**  
 21.92

The equation for the confidence interval

**PMAX**

227.87 – 1.96 \* 29.05, X + 1.96 \* 21.05  
227.87 – 56.93, X + 56.93  
  
CI at 95% for PMAX = [170.94, 284.8]

**SMAX**

226.8 – 1.96 \* 21.92, X + 1.96 \* 21.92  
226.8 – 42.96, X + 42.96  
  
CI at 95% for PMAX = [183.84, 269.76]

Text

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

**Repeat the previous question for a 99% confidence interval for the population mean based on the sample data. Compare and contrast the two confidence intervals and comment whether the means of the original dataset *airpollution* for PMAX and SMAX are included in these interval estimates. Justify your answer by providing calculations for full marks.**

For a 99% CI we use Z = 2.58

From the sample we have the values:

**PMAX**

**SMAX**  
 21.92

The equation for the confidence interval

**PMAX**

– 2.58 \* 29.05, + 2.58 \* 21.05  
 – 74.94, + 74.94  
  
CI at 99% for PMAX = [152.93, 302.81]

**SMAX**

– 2.58 \* 21.92, + 2.58 \* 21.92  
 – 56.54, + 56.54  
  
CI at 99% for PMAX = [170.26, 283.34]

Text

Description automatically generated

Graphical user interface, application

Description automatically generated

**A health organisation is debating that there is no difference between the variables PMAX and SMAX. Statistically test at a 5% level of significance if there is a difference in the average of PMAX and SMAX. Give a verdict and conclusion to your analysis.**

Let’s plot the two variables to understand their averages and measurements better next to each other.

Text

Description automatically generated

**Chart

Description automatically generated**

To prove that PMAX and SMAX are equal, we are going to follow a two sample T-Test. For that we are going to formulate our hypothesis first. The null hypothesis we want to formulate says that is no difference between the variables PMAX and SMAX, we have the following:

H0 : μPMAX = μSMAX   
H1 : μPMAX μSMAX

For the t-test we are going to apply, we have two independent variables, each one measures a different property. The two independents sample T-Test will have the following procedure:

A picture containing text, clock

Description automatically generated

Text

Description automatically generatedText

Description automatically generated





T-Statistic = 2.5

The significance level we have is 5% with the value of α = 0.05. For that we need to have a P-Value equivalent or higher than 0.05 to prove our null hypothesis. Now we need to prove from our T-Statistic





P-Value = .015 < α = .05

We can conclude that at a 5% significance level is not enough statistical evidence to support the null hypothesis, saying that is no difference between the averages for the variables PMAX and SMAX

**They now believe that PMAX should be greater than SMAX in most cities on average. Statistically test at the 10% level of significance whether this claim is true based on the data collected.**

For this question we need to formulate an alternative hypothesis, we now understand that PMAX and SMAX averages are different. Now we need to prove that PMAX is greater than SMAX. The new alternative hypothesis we have says: the average of PMAX is greater than the one of SMAX

Ha : μPMAX > μSMAX

The significance level we have is 10% with the value of α = 0.1. For that we can find our critical T-Value equal to a confidence interval of 90%.   
For the degrees of freedom we use the smallest out of the two.  
N1 is (80 – 1) and N2 is (80 – 1).





T-Value(.05, 79) = 1.664

From here we can build a confidence interval of 90% from the difference of the means PMAX – SMAX

A picture containing text

Description automatically generated





The interval is (18.57, 92.74)

We can conclude that, we are 90% confident that PMAX is greater than SMAX between 18.57 and 92.74 on average.

**Another claim the health organisation is making is that ideally the average TMR should be less than 1000. Statistically test at a 10% level of significance whether the average TMR is less than 1000. Include the set-up of the null and alternative hypotheses, a diagram and conclusion in your answer. All calculations should be done without using R.**

The claim we have here is that TMR < 1000, for our hypothesis we formulate:

H0 : μTMR = 1,000   
H1 : μTMR 1,000

From the TMR data we have:

Mean = 925.45  
Standard Deviation = 142.27  
N = 80  
Standard Error = 58.18

We want to know the average intervals for the values of TMR. The significance level we have is 10% with the value of α = 0.1. For that we can find our critical T-Value equal to a confidence interval of 90%.





T-Value(.95, 79) = 1.664  
T-Value(.05, 79) = -1.664

Now we calculate our interval for the values of TMR

The equation for the confidence interval

925.45 – 1.664 \* 15.90, 925.45 + 1.664\* 15.90  
925.45 – 26.46, 925.45 + 26.46  
  
CI at 90% for TMR = [898.99, 951.99]

We have enough statistical evidence with a 90% confidence that our null hypothesis of TMR average mean is not equal to 1,000.

Now we can apply the one sample T-Test at a 10% significance level for our alternative hypothesis, this will be a one sided test comparison with our T-Value(.1, 79) = -1.292

Text

Description automatically generated with medium confidence

T-Statistic = (925.45 – 1000)/58.18

T-Statistic = -4.69

T-Statistic < T-Value(.1, 79)

P(t < -4.69) = almost 0, which is less than 10% for our one sided test.  
  
We can conclude to say that at a 10% significance level, is enough statistical evidence to support the alternative hypothesis that on average the TMR is less than 1,000.

**What is the meaning of a Type I Error? Does this mean there is an probability that a Type I Error was made?**

Type I error means that we reject a null hypothesis that is actually real on the population. The level we set will increase or decrease the chance of making this mistake by chance, it refers to the probability number of making this error, the lower we set it, the lower the chance exist of making this error.