How Simpson's Paradox effect can be reduced in Covid-19

Abstract.-Simpson's Paradox can cause a lot of confusion in decision making hence it is necessary to reduce the effect of Simpson's Paradox. In this paper Covid-19 dataset of USA is used where we are comparing case fatality rate of different race/ethnicity and are trying to detect confounding variables which are causing Simpson's paradox and try to reduce their effects. There are 2 stages in which Simpson's effect can be detected and reduced one during the design phase and one during the analysis phase. During the design phase, randomization block design and restriction methods will be used for reduction while during the analysis phase stratification will be used. By reducing Simpson's effect we would be able to remove confusion in decision making and we would be able to understand the main reason because of which Simpson's paradox occurs.

Keywords: Simpson's Paradox, case fatality rate, race, design phase and analysis phase, decision making

1 Introduction

Coronavirus (Covid-19) is a contagious disease which was caused by severe acute respiratory syndrome coronavirus (SARS-COV2). Coronavirus was first identified in China in December 2019 and from there on it started spreading all over the world causing lots of deaths. Later on, the virus becomes so severe that World Health Organization declared it as Pandemic. (Ren et al., 2020)

The severity of Covid-19 disease can be indicated by the case fatality rate which is defined as the proportion of people who died from Covid-19 out of infected people within a certain period of time. In this report, we are going to use Covid-19 dataset of the USA which was taken from the USA Centre of Disease Control and Prevention (Data.cdc.gov, 2021). This dataset consists of Covid-19 cases and deaths related to different races/ethnicity. Our dataset consists of 2 lakh confirmed cases separated into different age groups from 0-9 till 80+ for different race/ethnicity. We are going to calculate and compare the case fatality rate for White Non-Hispanic and Black Non-Hispanic race and find out which race/Ethnicity has a higher case fatality rate.

After calculating the case fatality rate for different race/Ethnicity we get Table 1 and from this table, we can clearly see that the case fatality rate of Black Non-Hispanic race is having a higher case fatality rate for all age groups except 10-19. But if we look at the total case fatality rate percentage it is showing the opposite result as White Non-Hispanic got 13.7 % which is higher than case fatality rate of Black Non –Hispanic.

Table 1. Case Fatality rate of Black and White Non-Hispanic race by age group and in aggregated form

Age Groups	Case Fatality Rate of Black Non-Hispanic race	Case Fatality Rate of White Non-Hispanic race
0-9	0.2 %	0.255 %
10-19	0.171 %	0.055 %
20-29	0.431 %	0.157 %
30-39	0.847 %	0.42 %
40-49	2.568 %	1.14 %
50-59	6.007 %	3.37 %
60-69	15.4 %	10.37 %
70-79	27 %	26.7 %
80+	48.3 %	45.51 %
Total	10.186 %	13.7%

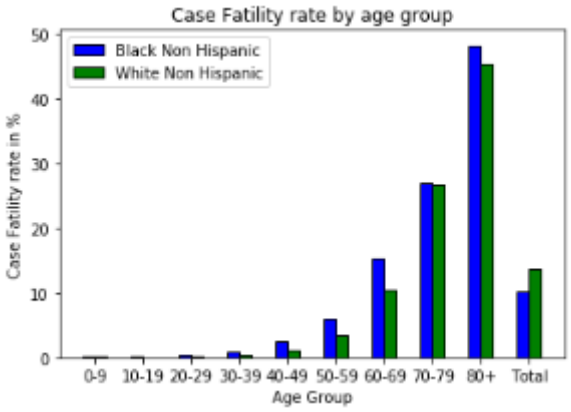

Fig. 1. Screenshot of Covid-19 Case Fatality rate in Black and White Non-Hispanic race by age group and in aggregated form ("Total")

We have even plotted the Table 1 data so that it is easy for comparing and the results are cleared that the White Non-Hispanic race got higher case fatality rate as compared to Black Non-Hispanic. For most of the people, this graph might look wrong as how it is possible that Black Non-Hispanic Race got higher case fatality rate for majority of age group but still overall case fatality rate is higher for White Non-Hispanic race. The problem which is occurring in our dataset is called Simpson's Paradox.

Simpson's Paradox can be defined as a problem in which the aggregated data show a particular trend, but the trend is reversed when the same data is divided into subgroups. Now we are going to explain this problem with an easy example. Let us take a famous Simpson's paradox example of UC Berkeley in 1973. In this example, UC Berkeley was sued for gender discrimination against women as lesser number of women were getting admitted as compared to men. We can clearly see from Table 2 that the total number of men admitted is 44 % while total number of women admitted is 35 % which means that men were getting more admitted as compared to women. However, if we look at each department we can see that for departments A, B, D, and F men were getting less admitted. If we look at individual departments it looks like men were getting discriminated as more women were getting admitted to majority department except for C and E. Since the group data and aggregated data were showing different trends hence it was considered Simpson's paradox. Simpson's Paradox can cause a lot of confusion as choosing between aggregate or group data is difficult as both lead to different results and therefore wrong decisions can be made.

Table 2. Data of UC Berkeley Discrimination Case 1973 for largest 6 departments

Departments	Men		Women	
	Applicants applied	admitted	Applicants applied	Admitted
A	825	62 %	108	82 %
B	560	63 %	25	68 %
C	325	37%	593	34 %
D	417	33%	375	35 %
E	191	28%	393	24 %
F	272	6%	341	7 %
Total		44 %		35 %

The main purpose of this report is to detect Simpson's paradox and try to reduce its effects on decision-making. In the Covid-19 dataset, our main objective is to answer the question why the White Non-Hispanic race is having higher case fatality rate in aggregated data but subgroup data shows a different result.

1.1 Background:-

From Simpson's paradox we understand that the trend of data changes when divided into subgroups but we are still not sure why did this happen? The main two reasons because of which Simpson's paradox occurs are:-

1 **Undetected Confounding Variable**:-Confounding variables are those variables that can affect the real relationship between dependent and independent variables. It can also be defined as extra variables that weren't considered in the experiment which can ruin the experiment and give inaccurate results. For example, let us suppose we collect data related to ice cream consumption and sunburn, and after analyzing data we found that higher ice cream consumption is related to sunburn. But does that mean ice cream consumption causes sunburn? No the true relationship between the dependent variable sunburn and the independent variable ice cream consumption is affected by confounding variable temperature which we didn't take into account. As high temperature causes people to eat more ice cream and due to hot weather they go outside and get sunburn.

Similarly in the UC Berkeley example applicants admitted and gender relation was getting affected as from the data it looks like men were getting more admitted as compared to women. However, it was the department variable which was causing problem in their relation as more men were applying for departments A and B as can be seen from Fig. 3 and since that departments were having higher acceptance rate as can be seen from Fig. 2 that's why it was looking that men were getting more admitted. While females were applying for departments D, E, and F which has low acceptance rate hence were getting less admitted.

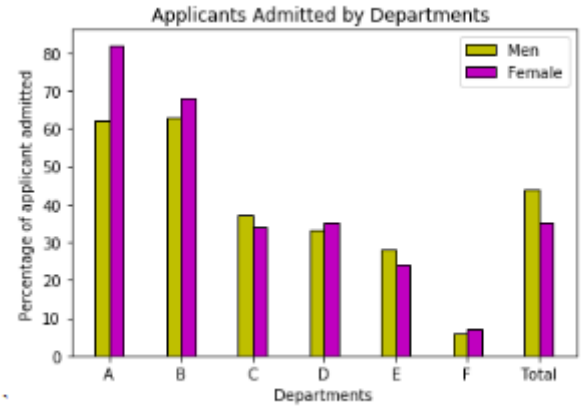

Fig. 2. Snapshot of percentage of Applicants Admitted in different departments as male and female

2 Second main reason behind Simpson's paradox was the disproportionate distribution of confounding variables in the group. This means that confounding variable is distributed unequally within groups In the UC Berkeley example department was unequally distributed as for some departments men were having high in number as compared to women and vice versa. As we can see from Fig. 3 that a large number of men applied for departments A and B as compared to other departments as those departments were less competitive and has high

acceptance rate. While large number of women applied for departments E and F which was more competitive departments and has a low acceptance rate as compared to A and B. Hence disproportionate distribution of confounding variables causes Simpson's Paradox.

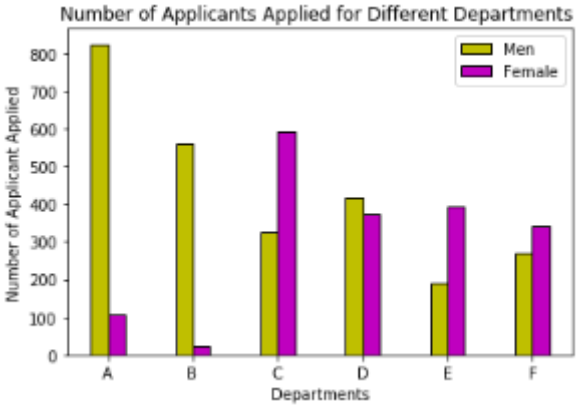

Fig. 3. Snapshot of Number of Applicants Admitted in different departments.

Once we have understood what causes Simpson's paradox it is necessary to understand methods that can be used to reduce its effect

There are many methods that are introduced to solve to issue of Simpson's paradox such as:

1 **Randomization Block Design**: - It is a method in which experimental subjects are divided into blocks and then treatment are assigned to these random blocks. But in the blocks, it is necessary to put subjects which are similar. By using this method we can reduce the variability of confounding variables For example in the UC Berkeley example suppose we have 1200 applicants we are going to divide gender variable which contains males and females into two different blocks each having 600 applicants as we can see from Fig. 4 and then put males into 6 groups randomly with each group having 100 applicants as there are 6 departments one group for each department and similarly for female and the group count should be equal for all men and for all women. Now our problem of men applying in departments that has high acceptance rate and female applying into departments that has low acceptance rate will be solved as there are equal number of subjects for all departments. Hence the variation in the confounding variable department would be solved.

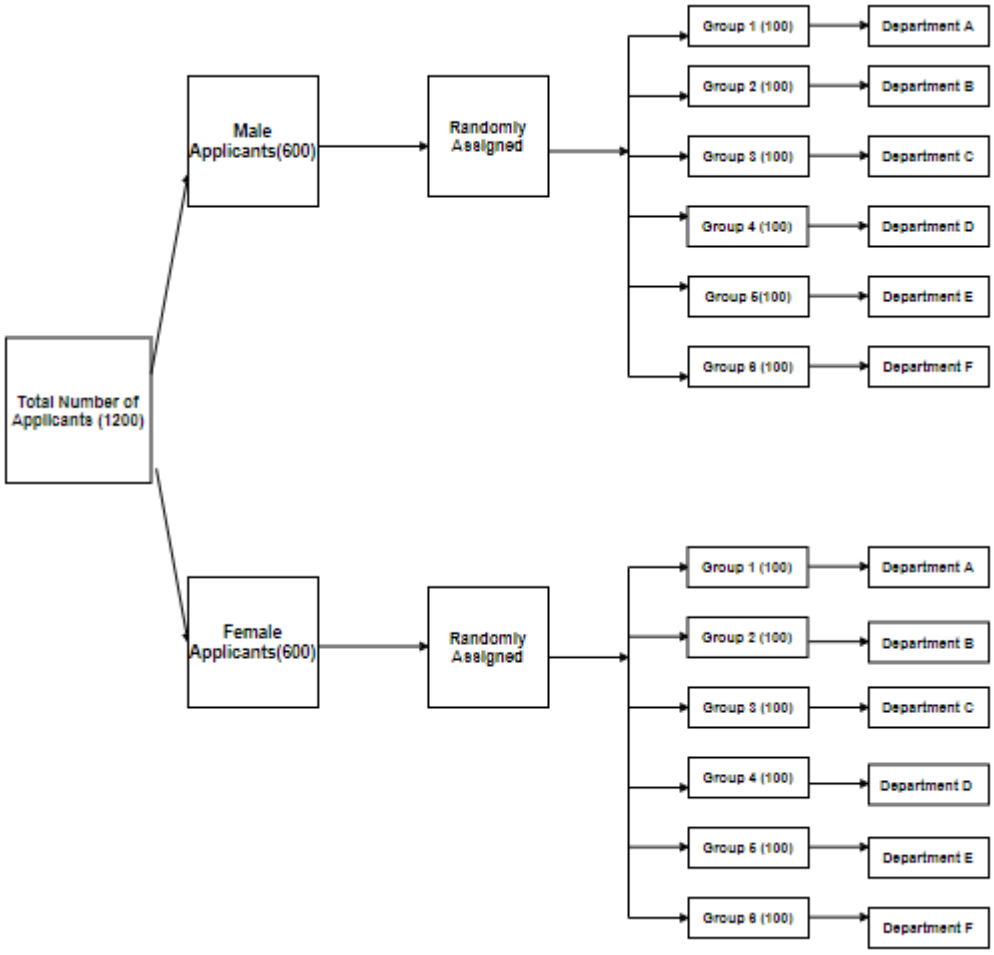

Fig.4. Snapshot of the process of Randomized block design for UC Berkeley example

2 **Restriction**:-Restriction is another method that can be used during the design phase. One of the main reasons of occurring Simpson's paradox is an unequal distribution of confounding variables and that's what the restriction method will try to eliminate and hence reduce Simpson's paradox. Restriction method will restrict the sample by only including those subjects who have the same value of confounding variables. For example in our UC Berkeley example, we have already identified that department was confounding variable as it was unequally distributed as for some departments like A, B men were more applying as their acceptance rate was high while for department E and F women were applying though their acceptance rate is less. By using restriction we can eliminate this problem by restricting confounding variable department. If we restrict department and consider only department A and B in our dataset and then calculate aggregated data then in both aggregated an subgroup it will show women were getting discriminated then our problem will be solved while if we consider department E and F then we would be clear that men were getting discriminated hence our confusion would be solved by choosing one of the two. As we can also see from Fig. 5 that that men were getting discriminated in total and also in subgrouped data if we choose departments A and B only.

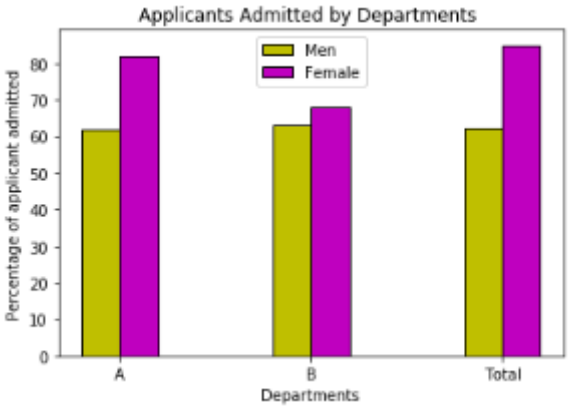

Fig.5. Snapshot of Applicants Admitted for department A and B and in aggregated form

3 **Stratification**:-Stratification is very similar to random sampling except that in stratification the data is divided into subgroups called strata instead of blocks in randomization block design and then we have to decide and choose sample size and calculate stratified sample calculation which is done by this formula:- Sample Size of strata=(size of sample/population size) *layer size. In this formula, layer size represent count in strata and then random sampling is done based on the ratio of a group to the total population. Hence it will produce groups in which confounding variables do not vary. This method is used in the analysis phase and is specially used when we have less number of confounding variables. For example, let us suppose we have 1000 population of males and female and we divide the population into two strata of 700 male and 300 female applicants and there are 6 departments they can apply and then we choose sample size let us assume we choose sample size 100 and now we are going to calculate the sample size of strata.

For males we will have (100/1000)*700 = 70 sample size of strata while for females we will have (100/1000)*300=30. After that we just need to use random sampling and produce groups for each department in that way variation of confounding variable is reduced. The whole process has been shown in Fig. 6 along with the group samples for each department.

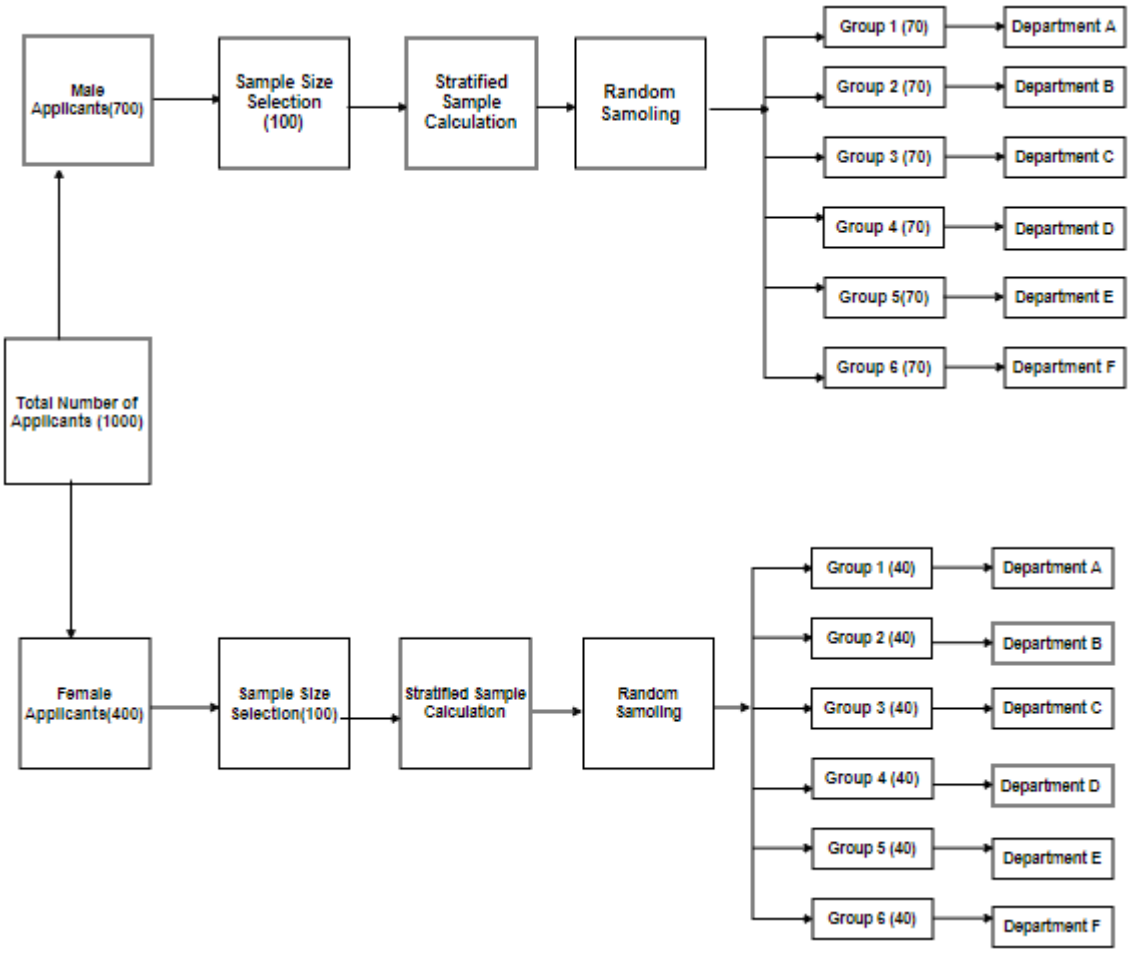

Fig.6. Snapshot of the process of Stratification for UC Berkeley example

1.2 Research Problem:

Since there are a lot of variables in the data set of Covid-19 such as sex, age, etc., it will be very difficult to identify which variables are causing Simpson's Paradox or which variable is confounding variables. Also, sometimes even after detecting some confounding variables, there will be one or two variable left which are not detected. So there should be a method which can help to identify confounding variables and once they are identified we need a method to find out if there are any undetected confounding variables left.

Since Simpson's paradox occurs due to the disproportionate distribution of confounding variables in the group, so we try to find a way by which we can reduce its effect on decision making which can help us in making the right decision and it does not cause confusion.

For controlling the effect of confounding variables there are many methods like randomization block design, stratification, restriction. But the main problem would be to identify which statistical measure would be best in the context of Covid-19 for handling these confounding variables.

1.3 Expected Outcomes:

1. Once all the steps of the methodology are performed, confounding variables causing Simpson's paradox in the Covid-19 dataset are detected.

2. Proposing methods like stratification, randomization block design, and restriction would try to reduce the effect of disproportionate distribution problem of confounding variables and hence controlling Simpson's paradox effects.

3. Confusion of which race/ethnicity is having the higher case fatality rate would be solved.

1.4 Significance and Benefits:

Wrong decisions can be avoided by reducing the effect of Simpson's paradox. For example, in the Covid-19 data set if aggregated data is preferred then the White Non-Hispanic race would have been selected as the race with higher case fatality rate. However, if we prefer sub grouped data results would have been different.

It also indicates how important is to understand the data as there might be a lot of hidden bias present in the dataset. We can't just trust numbers or figures we need to understand data and how it is generated especially for the medical industry as one wrong decision can cause a lot of problems. Let us take a famous example of kidney stone where a doctor has two treatments A and B for dealing with kidney stones. The doctor performs treatment A and B on 350 patients each and found out that treatment B has a success rate of 82 % as compared to treatment A which has 79 % success. However later on doctor analyze this treatment on small and large stones found out that for both small and large stone treatment A was better as compared to B as we can see from Table 3. Hence if the doctor would have gone with figures or numbers doctor would have performed treatment B on every patient with kidney stones. Hence we can't just trust data we need to understand how data is generated as in this case we need to understand how patients were chosen for different treatments. In treatment A surgery was performed while in treatment B pills were given hence majority of the hard cases like the large stone cases were given to treatment A which made this treatment less successful as compared to B in aggregated data as with harder case treatment A has less success rate and for B majority of cases of small stone were given which make the treatment overall better. But actually for small and for large treatment A is better as it can solve hard case with 73 % and small stone case with 96 % as compared to B which has lower success rate for larger and smaller stones.

Table 3. Data of Kidney Stone Treatment for treatment A and B for small and large stone and for combine stones.

Stone Size	Treatment A	Treatment B
Small	96 % (84/87)	87 % (234/270)
Large	73 % (192/263)	68 % (55/80)
Both(Small & Large)	79 % (276/350)	82 % (289/350)

2 Literature Review:-

In order to detect and reduce Simpson's paradox, different approaches have been used. One such approach was given by (Austin, 2011) where the author has used propensity methods like stratification, propensity score matching, and covariate adjustment to reduce the effect of confounding variables. Propensity score is the

conditional probability of being assigned to a treatment group based on some characteristics of the unit. By the use of propensity score differences between groups can be balanced hence eliminating bias in the dataset. However, this method does not work best for smaller dataset and for continuous independent variables or independent variables which change with time (Braga et al., 2012).

Another approach was given by (Jeon et al., 1987) where the author has used probability assignment to reduce Simpson's paradox. The two special cases considered in this paper are the balanced and unbalanced case. In balance case, every patient either male or female were given equal probability of 1/2 which is independent of treatment while in unbalance case male were assign probability p and for female 1-p probability were assigned. In balance case randomization has been used to avoid paradox while in the unbalanced case the author found out that by increasing the probability chances of Simpson's paradox decreases. One of the approach was given by (Fitzmaurice, 2006) where regression adjustment method is used by the author to adjust for confounding variables. This approach is very useful for many confounding variables as it avoids sparse data problem. However, these models have some strong assumptions like linearity which can cause problems.

If we look at the above three approaches being followed they all have considered only one method to reduce confounding variables and reaching to a conclusion. But for better and accurate results more than one method should have been used as there are less chances of detecting all confounding variables using a single method. Also, not necessarily one method can be the best approach for all the datasets as every method have different advantages and disadvantages. In the (Kievit et al., 2013) authors has introduced a new package for the detection of Simpson paradox which is written in R language. By using this package author can directly check or detect clusters within two continuous variables. If the correlation sign of the group is different from the correlation sign within cluster then it is considered as Simpson's paradox. However, even after running this package there might be few undetected confounding variables still left. Regarding undetected confounding variables, no method has been discussed or implemented.

3 Research Methodology:-

There are several steps that need to be performed to reduce the impact of Simpson's paradox which are given as follows:-

Firstly we are going to identify confounding variables which can be done in two phase: - design phase and analysis phase. For the identification of confounding variables during the design phase, it is necessary for us to understand the dataset properly and think about which variable might affect our dependent and independent variable. Since in this case, we are comparing case fatality rate for different races age group and sex are the main variables which might affect our case fatality rate as older people tend to be infected and die more as compared to young people and men are dying more as compared to women due to Covid-19. Hence it can be considered confounding variables. To confirm if it is a confounding variable or not we will use the data visualization technique which is used in the analysis phase.

So here we are going to analyze the proportion of confirmed case by age group for different races/ethnicity

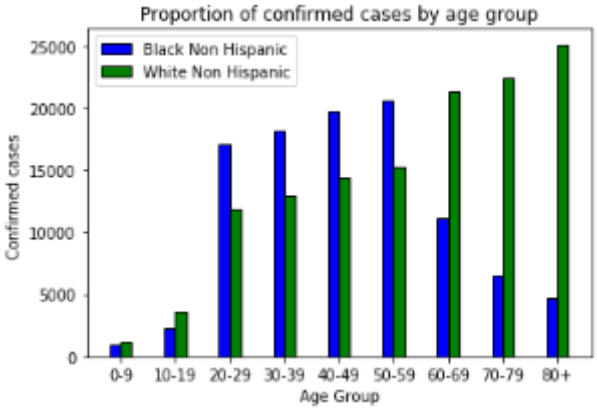

Fig. 7. Confirmed cases within each age group for Black and White Non-Hispanic race

From Fig. 7 we can clearly see that the White Non-Hispanic race has very high confirmed cases for older people as compared to the Black Non-Hispanic race which is a clear sign of disproportionate distribution of confounding variable. Hence age group can be considered as a confounding variable. By comparing for different variables such as sex no confounding variable was detected hence age group is our only confounding variable in the dataset which is affecting the case fatality rate.

Now we are going to use different methods such as randomization block design, restriction which is used in the design phase, and stratification method which is used in the analysis phase to find out which method is best for our dataset and reduce the effect of confounding variable age group so that confusion of decision making can be removed.

3.1 Randomization Block Design:-

In Randomization block design we are going to divide experimental subjects into blocks such that variability between blocks should be less than variability within blocks. In this method, we are going to randomly assign an equal number of subjects for different age groups such that the confounding variable effect is reduced. After performing randomization block design we get 5000 cases in every age group and after that, we can easily compare case fatality rate overall vs. sub grouped data.

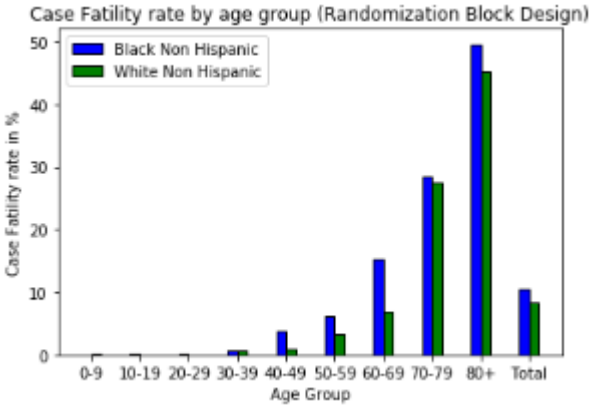

Fig. 8. Screenshot of Covid-19 Case Fatality rate in Black and White Non-Hispanic race by age group and in aggregated form ("total) by using Randomization block design method.

Table 4. Case Fatality rate in Black and White Non-Hispanic race by age group and in aggregated form ("total) by using Randomization block design method.

Age Groups	Case Fatality Rate of Black Non-Hispanic (Randomization Block Design)	Case Fatality Rate White Non-Hispanic (Randomization Block Design
0-9	0 %	0.2%
10-19	0.2 %	0%
20-29	0.2 %	0 %
30-39	0.6 %	0.6 %
40-49	3.8 %	0.8 %
50-59	6.2 %	3.2%
60-69	15.2 %	7 %
70-79	28.4 %	27.4 %
80+	49.6 %	45.2 %
Total	10.42 %	8.4%

After implementing Randomization block design method we again compare case fatality rate and we get Fig. 8 and Table 4 and we can clearly see that now the Black Non-Hispanic race got the higher case fatality rate in both aggregated data and for majority of age group except 0-9 group. One of the advantages of randomized block design is that it eliminates unknown confounding variables or confounding variables which we are not able to detect.

3.2 Restriction Method:-

Since we have already found out that especially for age above 70 there are more White Non-Hispanic race cases as compared to Black Non-Hispanic race then for that we are going to restrict our dataset and chose only case which are less than 70 age. This way the effect of confounding variable will be reduced. After implementing restriction we get Fig. 9 and from this graph, we can clearly see that the Black Non-Hispanic race got the higher case fatality rate for majority of age group and for aggregated total but however in our graph there is no age group of 70-79 and 80+ as we have restricted our age group.

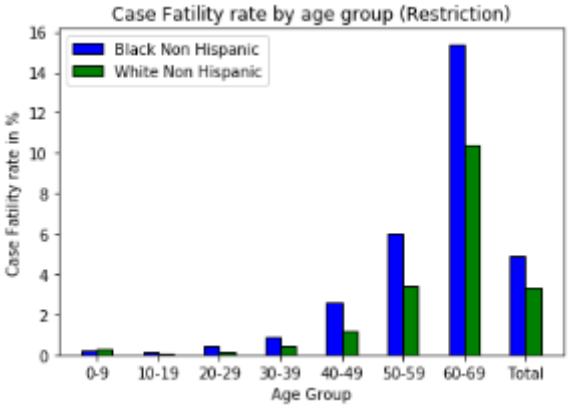

Fig. 9. Screenshot of Covid-19 Case Fatality rate in Black and White Non-Hispanic race by age group and in aggregated form ("total) by using Restriction method.

3.3 Stratification:

Stratification is pretty similar to randomization block design except for the fact that data is divided into subgroups called strata in which random sampling is done based on the ratio of a subgroup to the total population while in randomization block design equal groups are produced. Here we are going to produce age group strata based on the ratio of a subgroup to the total population and then use random sampling to reduce the effect of confounding variable.

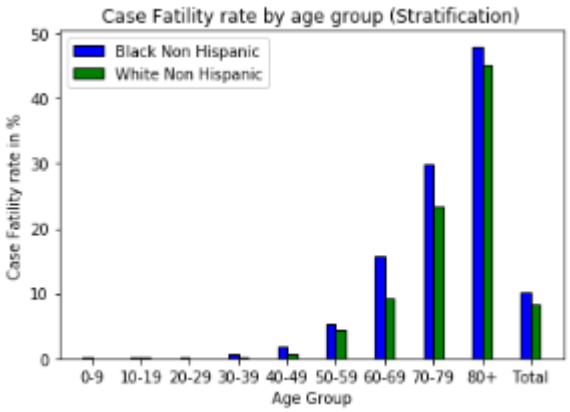

Fig. 10. Screenshot of Covid-19 Case Fatality rate in Black and White Non-Hispanic race by age group and in aggregated form ("total") by using Stratified Random Sampling.

Table 5. Case Fatality rate in Black and White Non-Hispanic race by age group and in aggregated form ("total) by using Stratification

Age Groups	Case Fatality Rate of Black Non-Hispanic (Stratification)	Case Fatality Rate White Non-Hispanic (Stratification)
0-9	0.2%	0%
10-19	0.2 %	0.2%
20-29	0.2 %	0 %
30-39	0.6 %	0.2%
40-49	1.8 %	0.6 %
50-59	5.2 %	4.4%
60-69	15.6 %	9.2 %
70-79	29.8 %	23.4%
80+	48 %	45.2 %
Total	10.12 %	8.36%

This Fig.10 and Table 5 is pretty similar to randomization block design and the graph again tells the same conclusion that the Black Non-Hispanic race got the higher case fatality rate in both overall and subgrouped data.

3.4 Checking for Undetected Confounding Variables:

After implementing all the methods we must find any missing confounding variable which we are not able to detect or not able to found if any undetected confounding variable is found then we have to use again all of the 3 methods to remove its effect. For checking of undetected confounding variable, we are going to use VIF (Variance Inflation Factor) and we can also create a baseline model and drop variable to see the difference in magnitude for detection of confounding variables. VIF is used to detect multicollinearity between two independent variables by multicollinearity I mean when two independent variables are correlated or when two variables move in the same direction or have a linear relationship. If we get a VIF of higher than 10 it can be considered as multicollinear. If two variables have high multicollinearity it also means that there are so highly correlated that we can't able to see the effect on the dependent variable. So by calculating VIF for every variable we try to detect undetected confounding but no VIF was higher than 10 for our dataset.

While in the second method we build a baseline model and select all variables and remove one variable and see what effect it has on other variables. For example let us assume we have age, gender, and weight variables if we drop weight from the model and if there is a change of 10 % or 20% in gender or male then it can be considered as a confounding variable. One of the disadvantages of this method is it is time consuming as we have to drop every variable and see the difference in their magnitude value. For our USA dataset, I have used multiple regressions as a baseline model and try to find out the difference in percentage. No variable were having a high percentage of change in magnitude. Hence no undetected confounding variable was found.

4 Outcome and Analysis:-

4.1 Implication of Decision Making:-

The main problem in our research project is to decide which data we should consult aggregated data which show White Non-Hispanic race got the higher case fatality rate or the subgrouped data based on age groups which show for the majority of age group Black Non-Hispanic race got the higher case fatality rate. As we have already found that age group was confounding variable as the size of groups were ignored as according to US census data around 9 % of White Non-Hispanic race is over 75 while for Black Non –Hispanic race it is only 4 %over 75. Since there are more older people in White Non-Hispanic race as compared to Black Non-Hispanic race there are more chance of people dying in White Non-Hispanic White due to Covid-19 as older people tend to die more. Therefore White Non –Hispanic race got the higher case fatality for aggregated data however for the majority of subgroup data based on age group we got the Black-Non Hispanic race as the race with higher case fatality rate.

Simpson's paradox decision making is totally conditional as it is not fixed whether we need to choose aggregated data or subgrouped data. However in our case according to the result we get I would choose subgroup data which means the Black Non-Hispanic race got the higher case fatality rate. I will explain it with a simple example let us suppose some person died due to Covid-19 and let us suppose his or her age was 25 and we need to guess what could be his or her race/ethnicity. According to the data which we have it should be Black Non-Hispanic race as they got the higher case fatality rate compared to white Non -Hispanic in majority of age group. Now let us assume that another person died but we don't know the age of person should we consider Black or White Non-Hispanic race? As for majority of age group we get Black Non-Hispanic race even we don't know the age of person but since for majority of age group Black Non-Hispanic race got the higher case fatality hence we would choose Black Non-Hispanic race. After using the methods stratification, randomization block design, and restriction we get the same result as when confounding variable effect is reduced we get the correct answer which is Black Non-Hispanic race got the higher case fatality rate. Hence our confusion of decision making is resolved.

4.2 Comparing Methods:-

After implementing all the 3 methods we get the same result however there are few advantages and disadvantages of the way these methods implement on the dataset. Randomization block design most important advantage is that it can even handle undetected confounding variables which is good for our dataset but however one of the major disadvantage of randomization block design is that it is not that effective for small trials and works only for clinical trials and after implementing Randomization block design from Table 4 we can see that for age group 0-9 White Non-Hispanic race has still larger case fatality rate. While Restriction method main advantage is that it is very quick and easy to implement however it has many disadvantages such as it can limit the sample size as by restricting the subject and it is not good with multiple confounding variables and we can't even evaluate the effect of restrictive variable as it doesn't vary it is just same of our dataset except the fact that it has removed the last few age group. For our dataset which is not that big and has only one confounding variable stratification method is best as it provides better precision as compared to randomization block design as we can see from Table 5 that no age group has higher case fatality for White Non-Hispanic while for randomization block design age group 0-9 does have higher case fatality rate for White Non-Hispanic and hence show that it has much better precision and accuracy and it is effective even with small sample and works best when we have less number of confounding variables.

4.3 Outcomes:-

1. After the research methodology section, we can clearly tell that the age group was the confounding variable which was causing confusion in decision making.

2. After implementing stratification, randomization block design, and restriction we get the same result that the Black Non-Hispanic race is having higher case fatality rate as compared to the White Non-Hispanic race and the stratification method is best for our dataset.

3. Our confusion of decision making is solved.

4.4 Future Work:-

In the future we can add more methods and form an algorithm that implements all the methods as depending on one method for solving Simpson's paradox is not good practice. For undetected confounding variables, we can use Residual confounding which is much faster and complex as compared to VIF and baseline model method.

5 Conclusion:-

Simpson's paradox can cause a lot of confusion as the aggregated data show a particular trend but the trend is reversed when data is sub-grouped. In this report, Covid-19 dataset of the USA is used where we try to compare case fatality rate for different race/ethnicity. After comparing we came to know that Simpson's paradox exists in the dataset as the White Non-Hispanic race got the highest case fatality rate for aggregated data but when we subgroup the data based on age group result was reversed. To solve this problem confounding variables need to be detected and its effect needs to be reduced as it causes Simpson's paradox. This will be done in two phases: - design phase and the analysis phase. In the design phase, we try to find confounding variables by understanding the dataset while in the analysis phase we try to use the data visualization technique. After implementing both phases we came to know that age group was considered as a confounding variable which was causing Simpson's Paradox. To reduce the effect of confounding variable we have used Randomization Block Design, Restriction, and Stratification methods. After implementing all these methods we came to know that the Black Non-Hispanic race was the race with the highest case fatality rate. By understanding the data and by using these methods we were able to remove the confusion in decision making. At the end we also compare all three methods and found out that the stratification method was best for our dataset.

For future work, we can add more methods as every method will provide some advantages and disadvantages and we can use residual confounding to find undetected confounding variables as it is a much faster and complex method.

6 References:-

1. Austin, P.C., 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, *46*(3), pp.399-424.
2. Braga, L.H.P., Farrokhyar, F. and Bhandari, M., 2012. Practical tips for surgical research: confounding: what is it and how do we deal with it?. *Canadian Journal of Surgery*, *55*(2), p.132.
3. Fenton, N., Neil, M. and Constantinou, A., 2019. Simpson's Paradox and the implications for medical trials. *arXiv preprint arXiv:1912.01422*.

4. Fitzmaurice, G., 2006. Confounding: regression adjustment. *Nutrition*, *22*(5), p.581.
5. Froelich, W., 2013. Mining association rules from database tables with the instances of Simpson's paradox. In *Advances in Databases and Information Systems* (pp. 79-90). Springer, Berlin, Heidelberg.
6. Jager, K.J., Zoccali, C., Macleod, A. and Dekker, F.W., 2008. Confounding: what it is and how to deal with it. *Kidney international*, *73*(3), pp.256-260.
7. Jeon, J.W., Chung, H.Y. and Bae, J.S., 1987. Chances of Simpson's paradox. *Journal of the Korean Statistical Society*, *16*(2), pp.117-127.
8. Kievit, R., Frankenhuis, W.E., Waldorp, L. and Borsboom, D., 2013. Simpson's paradox in psychological science: a practical guide. *Frontiers in psychology*, *4*, p.513.
9. Pathak, I., Choi, Y., Jiao, D., Yeung, D. and Liu, L., 2020. Racial-ethnic disparities in case fatality ratio narrowed after age standardization: A call for race-ethnicity-specific age distributions in State COVID-19 data. *Medrxiv*.
10. Pourhoseingholi, M.A., Baghestani, A.R. and Vahedi, M., 2012. How to control confounding effects by statistical analysis. *Gastroenterology and hepatology from bed to bench*, *5*(2), p.79.
11. Skelly, A.C., Dettori, J.R. and Brodt, E.D., 2012. Assessing bias: the importance of considering confounding. *Evidence-based spine-care journal*, *3*(1), p.9.
12. Skrivanek, S., 2010. Simpson's Paradox (and How to Avoid Its Effects). *published on Jun*, *21*.
13. Tripepi, G., Jager, K.J., Dekker, F.W. and Zoccali, C., 2010. Stratification for confounding–part 1: the Mantel-Haenszel formula. *Nephron Clinical Practice*, *116*(4), pp.c317-c321.
14. von Kügelgen, J., Gresele, L. and Schölkopf, B., 2020. Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects. *arXiv preprint arXiv:2005.07180*.
15. Data.cdc.gov. 2021. [online] Available at: <https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf> [Accessed 6 May 2021].