**Applied Business Statistics Team Project – Phase 1**
**Descriptive Statistics**

Analyzing the Deaths of the Counties of Arkansas on 1/01/2022 due to COVID-19

**Team Members:** Kaisa Wyly, Morgan O'Hearn, Emma Wolf, Natalie Bendlin (Stat Queens)

```
     Deaths            Date
 Min.   :  9.00   Min.   :2022-01-01
 1st Qu.: 48.25   1st Qu.:2022-01-01
 Median : 64.50   Median :2022-01-01
 Mean   :107.94   Mean   :2022-01-01
 3rd Qu.:139.25   3rd Qu.:2022-01-01
 Max.   :691.00   Max.   :2022-01-01
```

## 1.) Executive Summary

Dear Dr. Michael Osterholm, after analyzing data from the COVID 19 death count in the 50 counties within the state of Arkansas on January 1, 2022, we would like to address some recommendations, limitations, and concerns that we have regarding the data. We retrieved our data from https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties-2022.csv. This website summarizes the amount of deaths that took place in the counties of Arkansas on one specific date. COVID-19 has had a major impact on the amount of lives that were taken, in such a short amount of time. We feel that there are some crucial recommendations that need to be made in order to prevent this disease from spreading any further. One recommendation would be to make vaccines mandatory for the whole population of Arkansas. Another recommendation we have is to mandate wearing masks. The data proves how dangerous this virus was to the state of Arkansas, and it should highlight why our recommendations should be taken into consideration. We discovered that the maximum number of deaths was 691 and the minimum number of deaths for this date was 9. This shows that there was a wide range of deaths throughout the counties of Arkansas on this specific date, and that there are definitely some outliers. The 1st quartile is 48.25 and the 3rd quartile is 139.25 which is a good statistic to highlight because in between these two numbers is the interquartile range. The interquartile range tells us the spread of the middle half of the deaths in the counties of Arkansas, which gives us a good idea of where the number of deaths on 01/01/22 lies around. The mean is 107.94 which means that the average amount of deaths throughout the counties was around 108. Between these 50 counties on this day, there were a total of 5397 deaths. Our results are only showing 50/75 of the counties that are in Arkansas, so we can conclude that there were most likely 8000+ deaths throughout the state, which is a very concerning number. As we mentioned earlier, our 2 recommendations we have is to mandate receiving a vaccine, and mandate wearing a mask. When an individual receives a vaccine, it helps protect against the virus by creating an antibody response without having a potentially severe case of COVID-19. Since the vaccine reduces the risk of an individual catching a deathly variant, mandating vaccination could reduce the amount of deaths in the state.

Wearing masks is important so that if you're sick, then it helps keep your germs from infecting the people around you. When a group of people are together in close contact, COVID-19 has a greater chance of spreading, so the wearing of a mask may keep others safe. One limitation to our recommendation for the mask mandate is that you can't force someone to wear a mask, so not everyone follows the policy. When not everyone is on board with the mandate, it is much harder to stop the virus from spreading. Another limitation is that masks don't 100% keep others protected from those who have the virus. One limitation to our recommendation for our vaccine mandate is a very similar limitation for the mask mandate, in that mandating vaccines is also hard to ensure everyone follows the policy. We hope that the future holds lower deaths in the counties of Arkansas and all around the United States. We believe that this goal could be achieved if our recommendations are taken into consideration by Dr. Osterholm. Although Dr. Osterholm cannot do this on his own, we hope that he relays this information onto Arkansas personnel to promote vaccination and mask mandates.

1.) **Five Measures of Location:**

- **Mean: 107.947766**
- **Median: 64.50**
- **Mode: 55**
- **Trimmed Mean: 85.75**
- **Midrange: 350**

**Summary:** By analyzing the measures of location we got a very accurate summation of the data from the deaths on 1/1/22 in the Arkansas counties. We analyzed and compared the measures of location to obtain an adequate understanding of what the data can tell us about the deaths from COVID-19.

**Mean:** The mean is the most important and most widely used measure of location because it gives us an idea of where the center value is located within a dataset. It carries a piece of information from every observation that is given in each of the counties. It holds the most accuracy in showing the average amount of deaths, which is 108. We would use the mean when trying to understand the average amount of deaths in our dataset, which is typically the most useful. One disadvantage of using the mean is that it is sensitive to extreme values. For example, our maximum number of deaths in a county was 691 and our lowest was 9. These numbers are very far off of the majority of the numbers in the dataset which affects our mean to be less accurate.

**Median:** The median is the second most accurate measure of location because it is the middle-most observation in an ordered array of the dataset. In our case, the median is about 65 deaths. This is substantial in analyzing our data because of how far away it is from the mean, which is 108. When the median and mean are far away from each other that means that there is likely skewed data, with large variability. This means that the data differs greatly throughout the counties. The median is a good measure of location with skewed data because the median is not sensitive to extreme values, which means that numbers that are far off from the majority numbers

in the dataset do not throw off the median. One disadvantage of using the median is that it ignores information, which doesn't give us an accurate number that includes all of the dataset numbers. One advantage of using the mean is that all values do not have to be known to calculate the median.

**Mode:** The mode is the 3rd most important measure of location. This is because it allows us to measure the most frequently occurring observation when examining data. It is the value that is most likely to be sampled due to the fact that it involves taking the number most often shown. In our data, the mode was 55, showing that this amount of deaths happened more than once. An advantage of using the mode is that it will always be an observable value, because you simply look for the observation that occurs the most amount of times. We would use the mode when looking for a quick idea of the most common value that was observed.

**Trimmed Mean:** The trimmed mean is the 4th most important measure of location. It is computed by excluding the 10% largest and 10% smallest values of the data set, so the remaining 80% of the sample is given, representing the mean of the middle percent of the data. This helps us give more accurate data by excluding the outliers and focusing more on the central part of the data. Also, it eliminates any data points that severely skew the data giving us a better idea of the arithmetic mean. Our trimmed mean of 92.35 is closer to the median which takes out the inaccuracy of the mean. It is a limitation because it excludes some counties but ultimately it is a good measure of location to use when looking for a more accurate idea of the average number of deaths in Arkansas counties.

**Midrange:** The midrange is the 5th most important measure of location due to the fact that it doesn't take into account many of the numbers besides the highest and lowest points of the data, making it highly sensitive to extreme values. Our midrange is 350, which is the average between the highest and lowest data points and it does not take the outliers into consideration. It is a quick midpoint of the data sets, but overall is not a very accurate measure of location to determine the average amount of deaths in the counties. We would use the midrange when looking for a more accurate range of the data. One positive of the midrange is that it is quick and easy to calculate, because it only depends on 2 values, and all other values do not have to be known to calculate it. A negative is that it is more susceptible to errors in the data, and is not always an observable value.

**Mode**

```
> mfv(arkansas)
[1] 55
```

**Midrange**

| | | |
|---|---|---|
| midrange | 350 | `midrange <-(min(arkansas)+max(arkansas))/2` |

**Trimmed Mean**

| | | |
|---|---|---|
| trmean | 85.75 | `trmean<-mean(arkansas,trim=0.10)` |

2.) **Five Measures of Variability:**

- **Range: 682**
- **Standard Deviation: 115.0979**
- **Variance: 12982.58**
- **IQR: 91**
- **Coefficient of Variation**: **143.688993885492**

**Summary: write summaries of these 5 things below**

**Range:** The range represents the difference between the county in Arkansas with the least amount of covid related deaths and the county in Arkansas with the most amount of covid related deaths. The spread of covid related deaths among counties in Arkansas is 682. The range has some disadvantages because it excludes a majority of the sample space. Also, the range is extremely sensitive to outliers because the minimum and maximum are likely the outliers especially within this dataset. Since our range is a very high value, it tells us that our data varies greatly. This helps us conclude that the policies within Arkansas counties are not very strict due to the fluctuating number of deaths throughout the state. Some advantages of calculating the range would include that it is quick and easy to calculate, and all values do not have to be known to calculate it.

**Standard Deviation:** The standard deviation represents the average distance each county in Arkansas varies from then average covid related deaths in Arkansas counties. Each county on average varies 115.0979 covid related deaths from the average covid related deaths in all counties in Arkansas. We can determine that the data is more spread out because the standard deviation is a higher number. By comparing the standard deviation to our recommendations, we can say that there are counties that don't have strong enough regulations in place in order to avoid the outbreaks. The variability in the countries drive the standard deviation away from the mean which shows some inaccuracies in our data. One advantage the standard deviation holds is that it contains all of the information, so all values have to be known to calculate it. It is also associated with the mean, and since the mean is the most important measure of location it is helpful that it is included in the standard deviation. One disadvantage of standard deviation is that it can be hard to calculate it by hand because you have to take the square root of the sample and population variance.

**Variance:** Variance represents the spread of all the data regardless of outliers. The variance for the data of covid deaths in Arkansas is very high, meaning the data is super spread. This is because of the two very big outliers in 2 particular counties that bring the spread of the data up because they are so much larger than the majority of the data. The variance tells us how far on average each value is away from the mean, and this is an advantage because it treats all deviations from the mean as the same regardless of the direction. On the other hand, it can be a disadvantage because the variance gives added weight to the outliers.

**IQR:** The IQR is the upper (3rd) quartile - the lower (lst) quartile, otherwise known as the 75th percentile - the 25th percentile. The IQR of the data set is 91, which indicates that most of the values lie around 91. Additionally, larger IQR values indicate that the central portion of the data

is spread out further. Therefore, smaller values show that the middle values tend to cluster closer together. The IQR value is 91, which means that the central portion of the data is likely spread out. This means that the amount of covid deaths in Arkansas are spread out rather than clustered near each other. The IQR can be helpful because it can be used as a measure of variability if the extreme values are not being recorded exactly. It also is a better measure of variability for highly skewed data, and since our data is skewed to the right, it allows us to analyze the data better. Some disadvantages of the IQR is that it ignores some information and can be misleading if your data is not highly skewed.

**Coefficient of Variation:** The coefficient of variation represents how spread the dispersion of data is around the mean. The coefficient of variation for the covid death toll in Arkansas is 143.688993885492, which is relatively high, which means that the data around the mean is spread out. This would make sense because the amount of deaths from covid in each county can vary a lot based on many different factors.

**Range**

```
range <- max(arkansas)-min(arkansas)
```

| range | 682 |
|---|---|

**Variance**

```
> var(arkansas)*(n-1)/n
[1] 12982.58
```

**Standard Deviation**

```
> sd(arkansas)
[1] 115.0979
> variance
```

```
arkansas <- c(65,55,200,691,141,42,9,81,48,66,77,101,33,77,48
n<-50
variance <- sum((arkansas-sumdividedbyn)^2)/(n-1)
```

**IQR**

```
> IQR(arkansas)
[1] 91
```

**Coefficient of Variation**

| CV | 143.688993885492 |
|---|---|

```
CV=(155.0979/107.94)*100
```

**3.) Interpretation of Standard Deviation and Coefficient of Variation:**

- Standard Deviation: 115.0979

- Coefficient of Variation: 143.688993885492

- Summary: The standard deviation shows the average distance each specific county in Arkansas varies from the average covid related deaths in all Arkansas counties. The values in the distribution vary an average of 115.0979 units away from the mean of 107.94 units. The standard deviation allows us to use all of the information from our dataset, and all values have to be known in order to calculate it, which allows us to utilize every statistic. We use standard deviation to summarize continuous data when it is not significantly skewed. In our case, the standard deviation could be inaccurate because our dataset has extreme values. The coefficient of variation represents the ratio of the standard deviation to the mean. It shows the extent of variability, which is 143.688993885492, in relation to the mean of the population. This number is high, which memes that the level of dispersion around the mean is great. We use the coefficient of variation to compare the variability between groups that have means of different units or magnitudes.

4.) **Z-Score: 5.06577444071525**

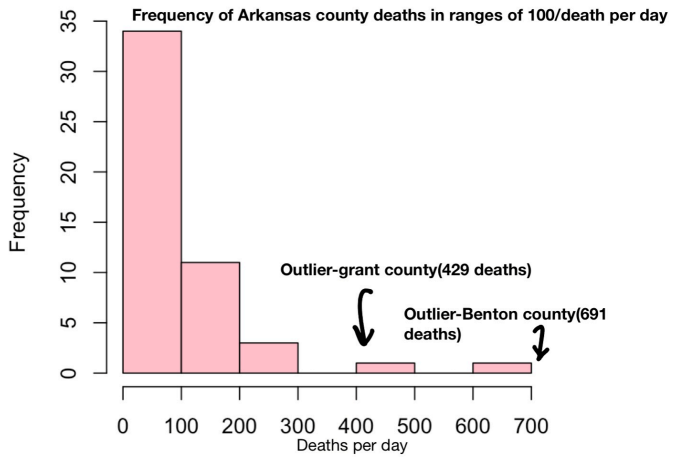
| | |
|---|---|
| z691 | 5.06577444071525 |

```
z691<-(691-sumdividedbyn)/115.0979
```

**Interpret the Z-score:** Z-scores are standardized scores that compare the distance between the data point and the mean with the standard deviation. The value of a z-score tells you how many standard deviations you are away from the mean. If a z-score is equal to 0, it is on the mean. A positive z-score indicates that the raw score is higher than the mean average. A negative z-score indicates the data value is smaller than the mean. The value of the z-score is 5.06577444071525, which tells us that the z-score is 5 standard deviations above the mean. Z-scores allow for the calculation of the probability of a score occurring within a normal distribution.

5.) **Identify and describe outliers:** There are two major outliers in the data of covid deaths in Arkansas counties. One of the outliers lie between 400 and 500 deaths, and the other outlier lies between 600 and 700 outliers. These two counties' number of deaths are much higher than the rest of the counties which makes them an outlier, and tells us that these amounts of deaths per county are not very common within our dataset.

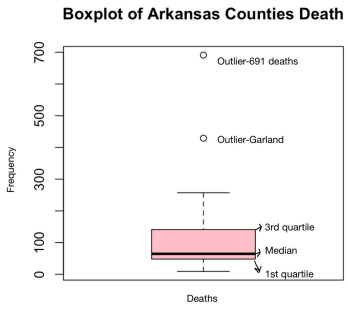
6.) **Histogram**




```
hist(arkansas, col = "pink")
```

Summarize the histogram: The histogram of Arkansas death toll is skewed to the right, meaning the death toll is clustered towards the lower number of deaths, similar to the boxplot. We can also see that the interval that is the most common for Arkansas death toll is the interval that has between 0 and 100 deaths. The median is 64.50 which falls in the most common interval. The intervals with the least amount of data are the intervals containing 400 and 500 deaths and the interval with 600 to 700 deaths. It's clear that only one county from Arkansas fell in each of these intervals.

### 7.) Boxplot




```
boxplot(arkansas, col = "pink",main="Boxplot of Arkansas Counties Death")
```

**Boxplot Summary:** The data of Arkansas counties deaths is skewed to the right, meaning the data is clustered more towards lower numbers of deaths. We can see that there are two outliers, one of them is between 400 and 500, and the other outlier is around 700. This was found by looking at the dots on the graph. These outliers represent countries where the death toll did not follow the usual pattern of Arkansas deaths. We can also see that the median for this distribution is between 50 and 150 which means that the

center of the data for all of the Arkansas counties can be concluded to be between 50 and 150. On the other hand, we can conclude that the average number of deaths is between 50-150 due to the location of the box. The box covers the interquartile interval, where 50% of the data can be found. This is a helpful measure of variability because it excludes the outliers, giving us a more accurate picture of the majority of the dataset.

**9.)Shape of the Data:** As stated earlier, our data has a skewed distribution, and it is skewed to the right. We analyzed the skewness from the histogram. We noticed that the majority of the data was clumped to the left side, meaning that the most common number of deaths in each county of Arkansas on January 1st, 2022 was less than 100 deaths. This is because the interval from 0 to 100 had the most frequency. Also, affecting the shape of the data was our outliers. We could analyze from the histogram that the interval between 400 and 500 and the interval from 600 to 700 had very small frequencies, telling us that they could potentially be outliers. This was confirmed from the boxplot, because there are dots outside of the whiskers. Our data is skewed to the right because the mean of 107.94 is greater than the median of 64.50 and mean > median. Since the mean is greater than the median, it helped us conclude that throughout the Arkansas counties, most of the death counts were relatively around the same number.