30.05.2023

Spring 2023

**INDR 252 Applied Statistics Case Final Report**

**A case study on “Effectiveness of Efforts to Reduce American Crime Rates”**

**Group Number: 8**

Enis Ata Göğüş - 79433 - egogus21

İdil İşsever - 80447 - iissever21

Ege Erdem Özlü - 80481 - eozlu21

Ceyda Özdeş - 80437 - cozdes21

Tan Karahasanoğlu - 79136- tkarahasanoglu21

Mansour Alazim - 78105 - malazim21

Eren Yavuz - 79327 - eyavuz21

INDR 252 Applied Statistics Case Final Report

The US Government wants to see the change of assault rates in 5 years after implementing precautions, to check if they were effective. The data provided is the aggravated assault rates per 100,000 population for each state, in years 1996 and 2001, respectively. In order to evaluate, they will check for three goals that are the reduction of the mean and the standard deviation of the assault rates, and the reduction of the proportion of the states that have more than 400 cases to at most 15%. In order to check if the above goals are achieved, we will compare the means and standard deviations in years 1996 and 2001 via CI and hypothesis testing, respectively, to see if there is a decrement, and formulate another hypothesis test to check if the proportion of states that have a high crime rate (i.e. more than 400 cases) is at most 15%.

To check if the population forms a random sample we check the time series plot, Figure 4, which shows fluctuations with no apparent trends and therefore the sample can be assumed random. (It should be noted that the time-series plots in Figure 4 are not time dependent but rather ordered alphabetically. Therefore the randomness is evaluated with respect to the alphabetical order.) To see if the data provided can be assumed normal, we check the normal probability plot, Figure 5, where it is apparent that data for year 1996 is closer to normal than year 2001. Both years do not form an exact normal distribution, however, both are close enough to normal to produce confidence intervals accordingly. Checking the boxplots (Figure 3), we see that there is one outlier in 1996 and one in 2001, which are the values 767 and 695, respectively and both the values are for South Carolina. This suggests that South Carolina generally has a higher crime rate in comparison to other states in the dataset, which is notable information and therefore the outliers should not be removed from the datasets. The datasets are correlated and not independent from each other. This is apparent in the time series, histograms, as well as the boxplots, which all show similar shapes and skewness for both datasets, suggesting they are dependent. This is an expected occurrence, as some states generally have lower or higher crime rates depending on the populations’ education, socio-economic status and income inequality (A state producing a lower crime rate in 1996 will be expected to produce a lower crime rate in 2001.). The descriptive statistics given in Table 1 show that the crime rates dropped noticeably in 2001 when comparing the means of assault rates in 1996 and 2001. The Assault Rates of 1996 have a mean of 320.06, and the Assault Rates of 2001 have a mean of 286.88. Both means are in an acceptable range and do not show high crime rates (i.e. higher than 400). Checking the boxplots (Figure 3), histograms (Figure 2), and time-series plots (Figure 4), we can observe that the variabilities of the data sets are similar. In terms of skewness, years 1996 and 2001 have values of 0.49 and 0.55 respectively, which are pretty close to each other and can be assumed to have low skewness.

The unbiased estimators for the means of years 1996 and 2001 are 320.06 and 286.88 respectively. Table 2 presents CIs around the mean difference between years 1996 and 2001. Since the sample size is large (n>40) the mean can be assumed to be a normally distributed random variable according to the central limit theorem. For that reason, we chose to construct confidence intervals to compare mean differences. Since the two data sets are not independent, instead they are correlated due to being constructed in the same states, the CIs around mean differences should be used, instead of CIs around means when observing the crime rate changes during years 1996 - 2001. By checking the values in Table 2, we see that the lower confidence interval limits have significantly higher negative values while upper confidence intervals have comparably lower positive values. This shows that the crime rate means are significantly lower in 2001 compared to 1996. Thus, we can confidently say that the precautions were able to reduce the assault rates. This finding also confirms our previous assumption of decreased crime rates.

The unbiased estimators for the proportions of states with a high crime rate is calculated from the formula (Values Above 400)/(Number of values), which produced values of 0.3 and 0.2, respectively, as shown in Table 3. The appropriate hypothesis test to check if the proportion is limited at 15%, we need to make 3 assumptions. Our first assumption is randomness; therefore, our data should be collected randomly or represent a random sample of the population. Our second assumption is independence, the assault rate of an individual state should be independent of the rest of the states. Our third assumption is large sample size, the conditions np ≥ 10 and n(1-p) ≥ 10 should be satisfied. For both 1996 and 2001, these conditions are satisfied. A table with an additional confidence interval table is also provided with explanations (Table 4).

The appropriate hypothesis test is given below, where H0 is the null hypothesis and Ha is the alternative hypothesis:

**H0:** p1 = 0.15 (proportion of states with assault rates > 400 in 1996 is at most 15%)

**Ha:** p1 > 0.15 (proportion of states with assault rates > 400 in 1996 is greater than 15%)

**H0:** p2 = 0.15 (proportion of states with assault rates > 400 in 2001 is at most 15%)

**Ha:** p2 > 0.15 (proportion of states with assault rates > 400 in 2001 is greater than 15%)

To perform the hypothesis test, we calculate the p-values for each year and compare them to the significance level.

**For 1996:**

The p-value that we get for this hypothesis test is 0.00149. This means that for the conventional significance levels alpha = 1%, 5%, 10%; we can reject the null hypothesis to accept our alternate hypothesis which is that the proportion of states that have a crime rate of greater than 400 is higher than 15%. Our significance value would have to be very small (alpha = 0.01%) to not reject the null hypothesis, so it is very likely that the alternative hypothesis is true.

**For 2001:**

The calculated p-value is 0.161, which is greater than significance levels (10%,5%,1%). Therefore, we fail to reject the null hypothesis (H0) for 2001. This means that we do not have strong evidence to conclude that the proportion of states with assault rates greater than 400 in 2001 is greater than 15%.

This suggests that the goal of limiting the proportion of high-crime states at 15% is satisfied.

Unbiased estimators for standard deviation are 171.82 and 152.11 for 1996 and 2001, respectively.

The hypothesis test is as follows:

**H0:** (var of 96)/(var of 01) = 1

**Ha:** (var of 96)/(var of 01) > 1

0.198 -> p value for F-test

We can reject the null hypothesis for alpha = 20% but for anything smaller than that we cannot reject the null hypothesis so most likely there has not been a significant reduction in the standard deviation.

Our null hypothesis above is that both sigma values are equal (note that to apply the F-test we need to assume that both samples are normally distributed, this can be justified by using the central limit theorem. Also, our F-test is sensitive to changes in normality. In other words, if our samples have outliers this can result in an inaccurate F-value. Also, the F-test assumes randomness in our two samples). When we apply F-Test for both samples, the P-value and F-value is (0.198, 1.292). With 80 percent confidence we can say that the null hypothesis is incorrect. Thus, it has a confidence level of %80 that the standard deviations of two populations are not equal.

The results have shown that while the government has not been able to achieve their goal of reducing the proportion of states that have a crime rate of more than 400, they have considerably reduced the crime rate, as seen by our CI’s regarding the difference in means and the hypothesis test regarding the said proportion. We cannot say that the standard deviation of the crime rates has not reduced at a level that is statistically significant due to our f-test. This is an expected result since the goal was not to make the crime rates equally distributed across all states. Overall, there is a trend that shows considerable improvement, but not to the extent that the USA government has hoped to achieve.

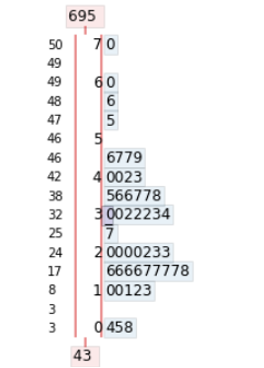
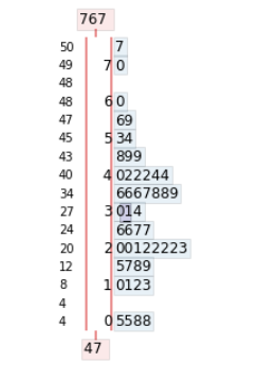
**APPENDIX**

**Table 1:** Descriptive Statistics

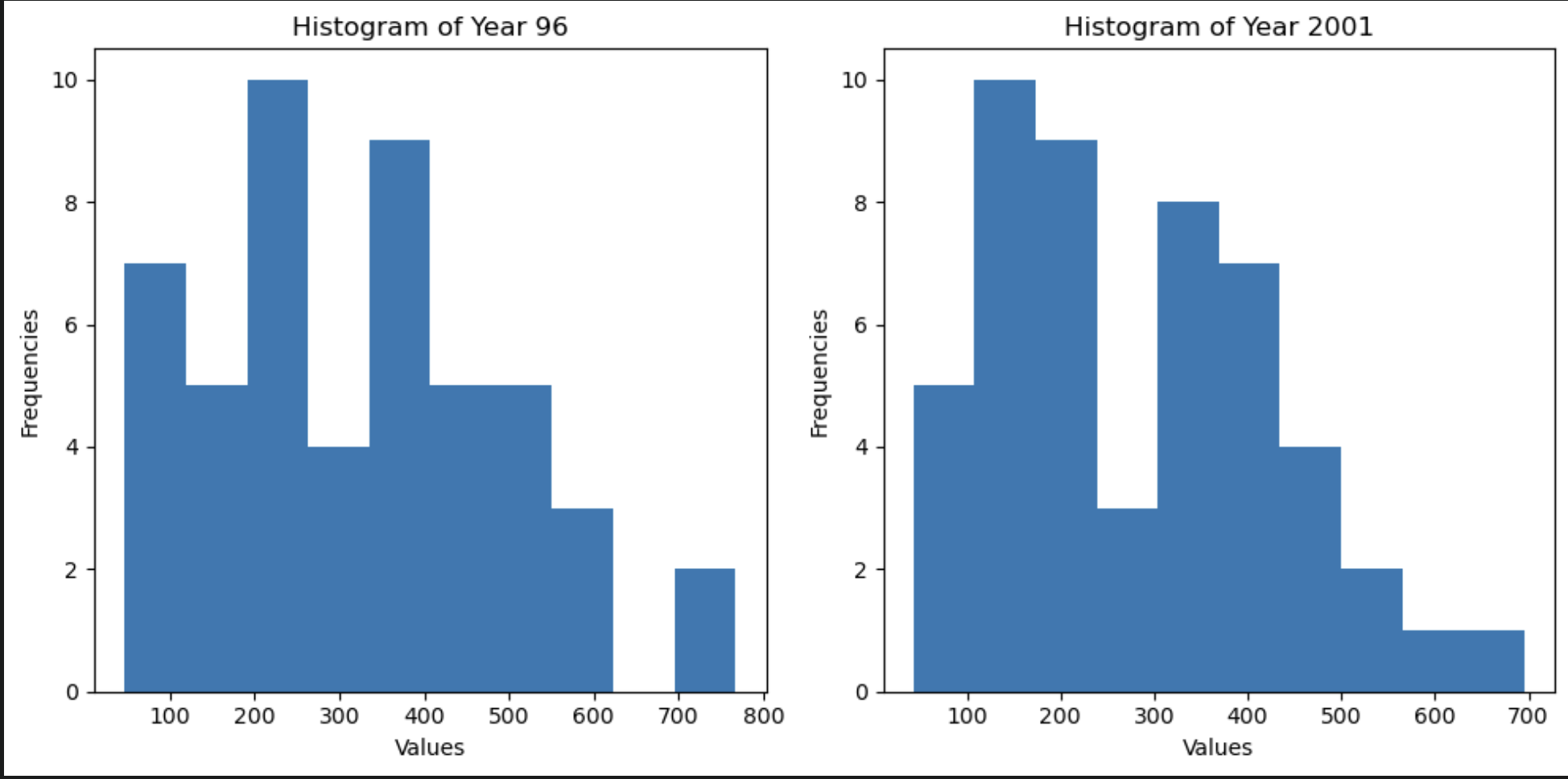
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **N** | **Mean** | **Std Dev** | **Min** | **Max** | **1st Quartile** | **Median** | **3rd Quartile** | **Skewness** |
| **Assault 96** | 50 | 320.06 | 171.82 | 47 | 767 | 197.75 | 302 | 423.5 | 0.4925 |
| **Assault 01** | 50 | 286.88 | 152.11 | 43 | 695 | 166.75 | 280.5 | 376.75 | 0.5500 |

**Figure 1:** Stem and leaf diagrams of years 1996 and 2001

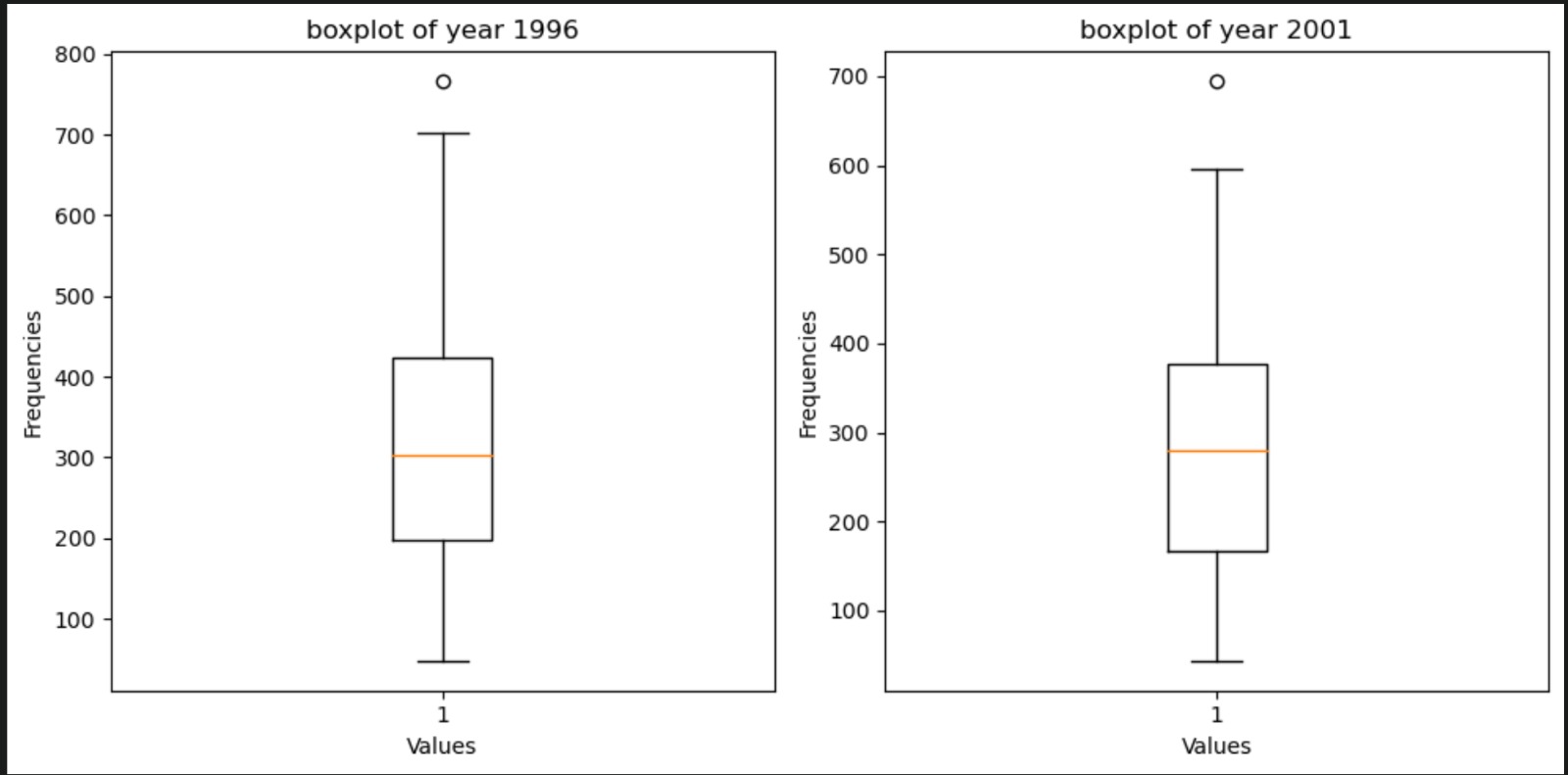
Assault rates in 1996 Assault rates in 2001



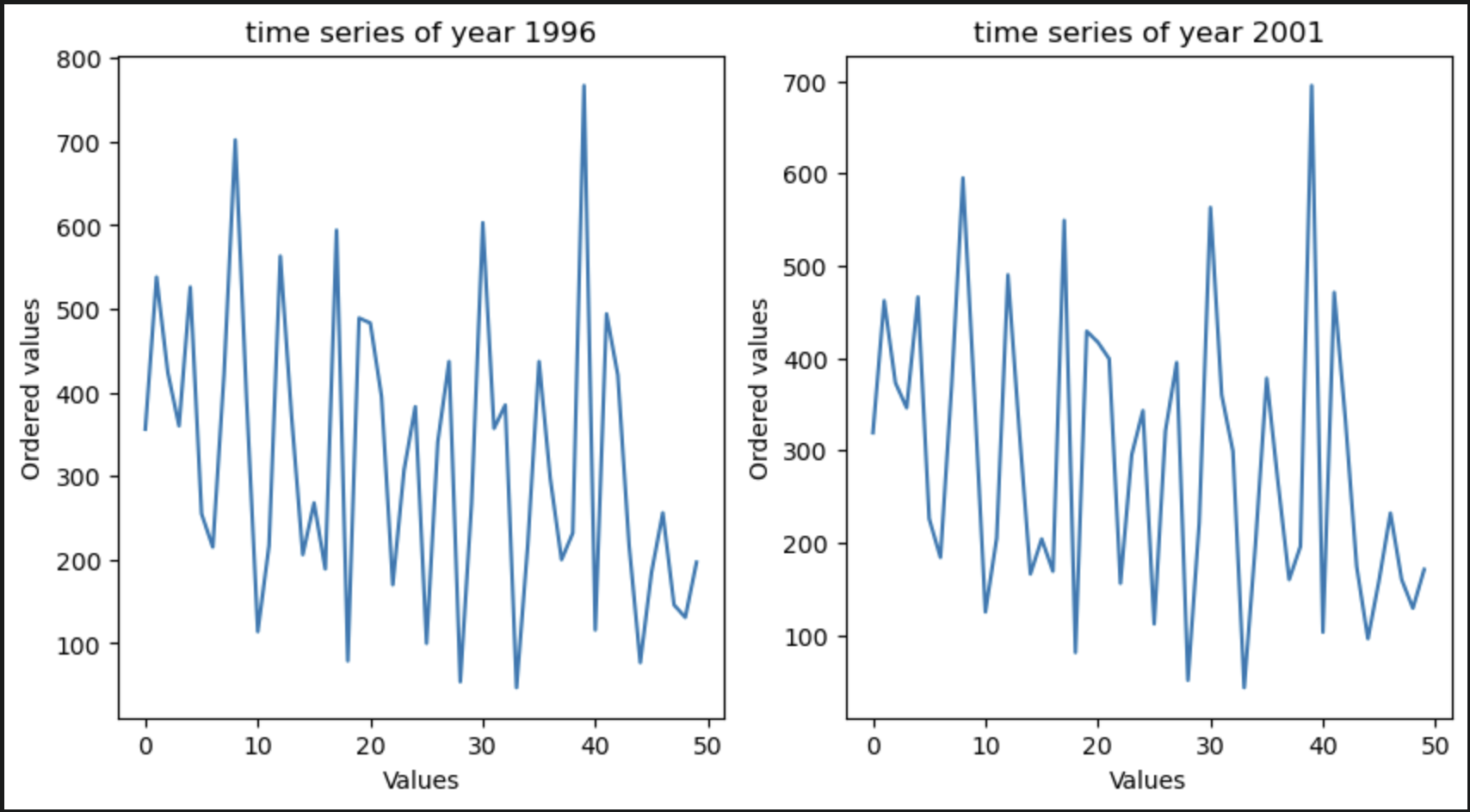
**Figure 2:** Histograms of year 1996 and 2001



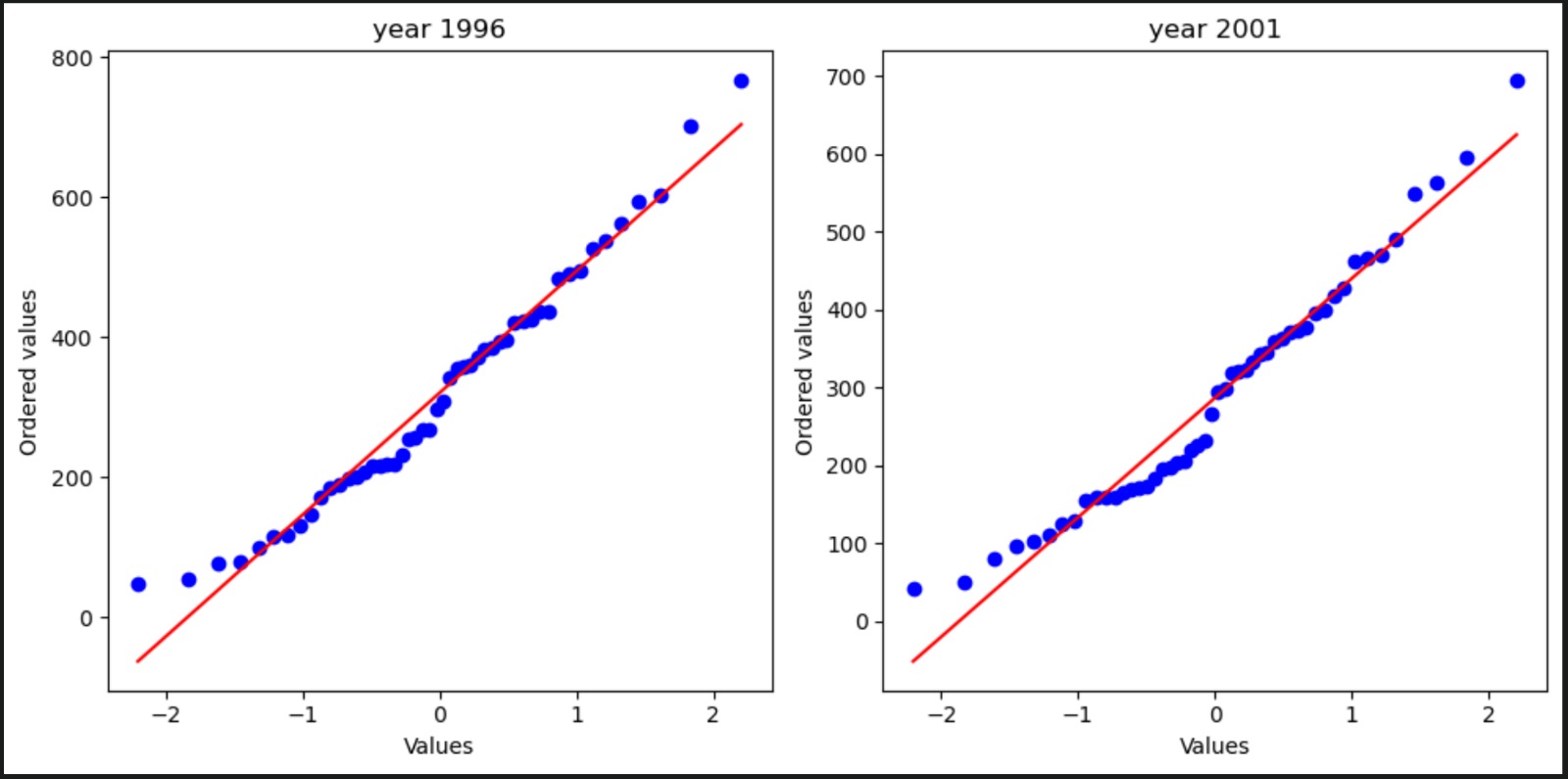
**Figure 3:** Boxplots of years 1996 and 2001



**Figure 4:** Time series plots of years 1996 and 2001



**Figure 5:** Normal probability plots of years 1996 and 2001



**Table 2:** Confidence intervals around the mean differences in the assault rates for years 1996 - 2001

|  |  |  |
| --- | --- | --- |
|  |  | Years 1996 - 2001 |
| 90% | Lower Confidence Limit | -86.561 |
| Upper Confidence Limit | 20.201 |
| 95% | Lower Confidence Limit | -96.787 |
| Upper Confidence Limit | 30.427 |
| 99% | Lower Confidence Limit | -116.774 |
| Upper Confidence Limit | 50.414 |

**Table 3:** High Crime Rate Proportions (above 400)

|  |  |  |
| --- | --- | --- |
| Point Estimates | Assault 96 | Assault 01 |
| p(Regular) (Values Above 400)/N | 15/50=0.3 | 10/50=0.2 |
| std | 22.91 | 20 |
| Standard Error for proportions | 0.0632 | 0.0565 |
| Standard Error of Difference of Proportions | 0.0848 | |

**Table 4: Confidence intervals of proportions of each year and the explanation**

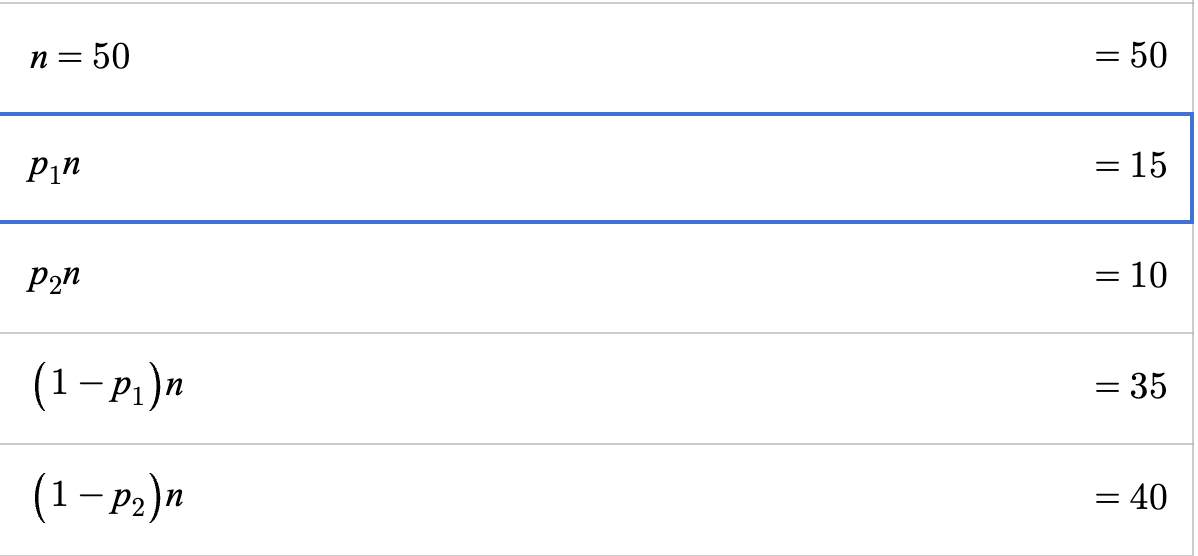
In Table 4, confidence intervals were constructed using the calculated proportions. These intervals were designed to ideally contain a crime rate of 0.15. In order to visually represent this, the cells in the table were color-coded: green for intervals containing 0.15 and red for intervals not containing it.

Upon analyzing the data for the year 1996 at a confidence level of 99%, it was observed that the crime rate of 15% did not fall within the constructed confidence interval. As a result, we can state with 99% confidence that the crime rate in 1996 is higher than 15%.

Conversely, when examining the data for the year 2001 with different confidence levels (90%, 95%, and 99%), it was determined that the crime rate could potentially be below 15%. These findings allow us to conclude that there is a possibility that the crime rate in 2001 is lower than 15%.

|  |  |  |
| --- | --- | --- |
| Confidence / Years | 1996 (Lower Bound-Upper Bound) | 2001(Lower Bound-Upper Bound) |
| %85 | 0.2238 - 0.4057 | 0.2238-0.4057 |
| %90 | 0.2108 - 0.4187 | 0.1291-0.3152 |
| %95 | 0.1909 - 0.4386 | 0.1113-0.3331 |
| %99 | 0.1520 - 0.4776 | 0.0764-0.3679 |

**Table 5:** Large Sample Size Assumptions Proof (0.3,0.2)

****