Drug Overdose Death Rates, by Drug Type, Sex, Age, Race, and Hispanic Origin: United States

STAT 3703 Statistical Computation Project

Jason Nguyen

April 26, 2024

Abstract

The world continues to face a public health crisis of opioid overdoses. This problem seems to worsen every year and plagues our society. This project will explore a dataset on drug overdose death rates, specifically examining how various demographic factors influence these rates. We will employ statistical methods to uncover significant relationships and trends of these demographic factors. This knowledge can then be used to inform prevention strategies, treatment approaches, and resource allocation, ultimately working towards saving lives and mitigating this ongoing tragedy.

Department of Computer Science and Engineering

University of Arkansas – Fort Smith

Contents

1 Dataset

This research uses drug overdose death rate data from the Centers for Disease Control and Prevention and National Center for Health Statistics. The data is separated by five features, drug type, sex, age, race, and Hispanic origin. It uses the United States population with data ranging from 1999 to 2018. A total of 6228 drug overdose death rates at the measurement of deaths per 100,000 resident population were presented in the dataset. For this study, it will be assumed that the data was collected randomly, follows a normal distribution, and was collected properly and reliably. [1]

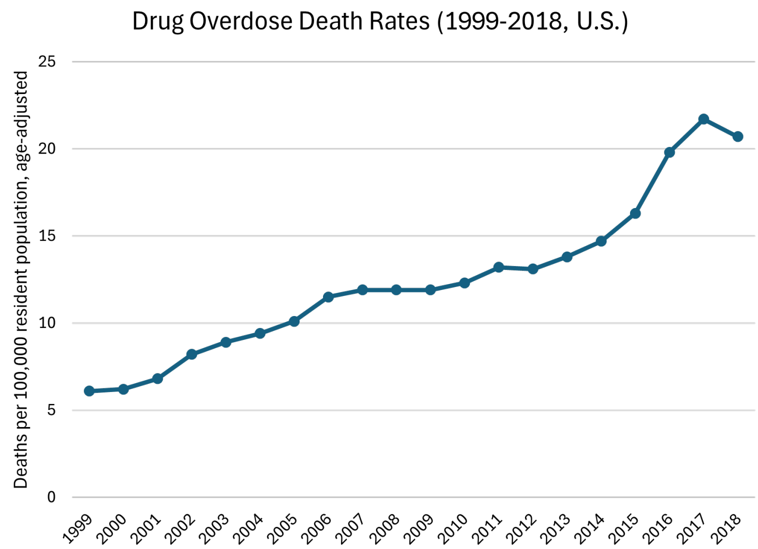


Figure 1: Yearly death rates in the U.S.

2 Hypothesis Tests

2.1 Hypothesis 1: Male vs. Female

The following hypotheses are formulated for this test:

$$\begin{aligned} H_0 &: \mu_m = \mu_f, \mu_m - \mu_f = 0 \\ H_1 &: \mu_m \neq \mu_f, \mu_m - \mu_f > 0 \end{aligned} \tag{1}$$

We will be looking at the explanatory variable of sex (male, female) and response variable of the drug overdose death rates for this test. The null hypothesis, H_0, states that there is no association between sex and drug overdose death rates ($\mu_m = \mu_f, \mu_m - \mu_f = 0$). In other words, the male and female means are equal or the difference of their means is equal to zero. The alternate hypothesis, H_1, states that there is an association between sex and drug overdose death rates ($\mu_m \neq \mu_f, \mu_m - \mu_f > 0$). In other words, the male and female means are not equal or the difference of their means is greater than zero. A two-sample t-test assuming unequal variances using a 95% confidence interval will be used for this test to compare male and female.

Male		Female	
Mean	15.905	Mean	8.945
Standard Error	1.35641701	Standard Error	0.692154341
Median	14.9	Median	9
Mode	14.8	Mode	10.2
Standard Deviation	6.066081281	Standard Deviation	3.095408314
Sample Variance	36.79734211	Sample Variance	9.581552632
Kurtosis	0.384775403	Kurtosis	-0.730294714
Skewness	0.964332565	Skewness	0.016729234
Range	20.9	Range	10.5
Minimum	8.2	Minimum	3.9
Maximum	29.1	Maximum	14.4
Sum	318.1	Sum	178.9
Count	20	Count	20

Figure 2: Descriptive statistics

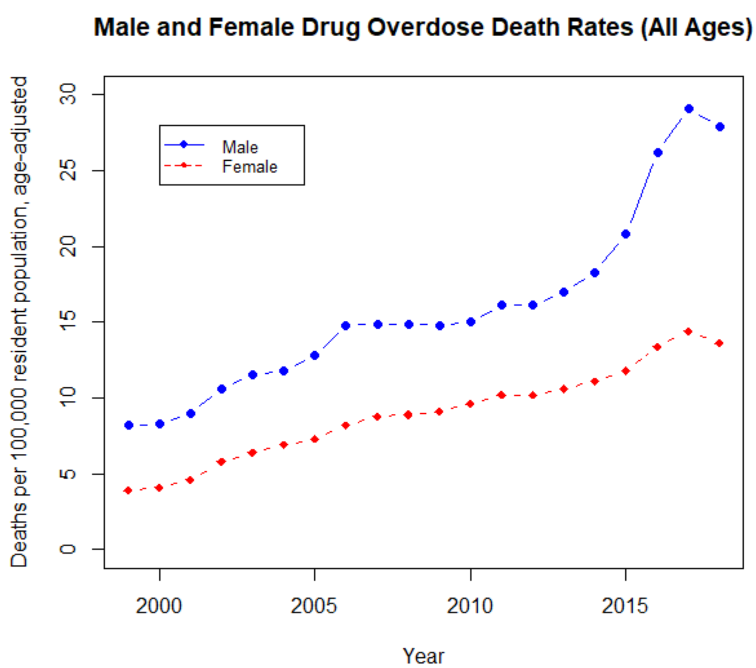


Figure 3: Plot of male and female drug overdose death rates (all ages)

t-Test: Two-Sample Assuming Unequal Variances		
	Male	*Female*
Mean	15.905	8.945
Variance	36.79734211	9.581552632
Observations	20	20
Hypothesized Mean Difference	0	
df	28	
t Stat	**4.570503439**	
P(T<=t) one-tail	4.48317E-05	
t Critical one-tail	1.701130934	
P(T<=t) two-tail	**8.96634E-05**	
t Critical two-tail	2.048407142	

Figure 4: Two-sample t-test results

With the p-value of $0.000090 < 0.05$, we reject the null hypothesis and state that there is sufficient evidence at the 0.05 level to conclude an association between sex and drug overdose death rates ($\mu_m \neq \mu_f$).

2.2 Hypothesis 2: Type of Drug

The following hypotheses are formulated for this test:

$$\begin{aligned} &H_0 : \mu_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ &H_1 : \text{At least one } \mu_i \text{ is different from the others} \end{aligned} \tag{2}$$

We will be looking at the explanatory variable of drug type (any opioid, heroin, methadone, natural/semisynthetic opioids, other synthetic opioids (other than methadone)) and response variable of the drug overdose death rates for this test. The null hypothesis, H_0, states that there is no difference in drug overdose death rates amongst the different type of drugs ($\mu_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4$). In other words, the drug type means are equal or the difference of their means is equal to zero. The alternate hypothesis, H_1, states that there is a difference in drug overdose death rates amongst the different type of drugs (At least one μ_i is different from the others). In other words, the drug type means are not equal or the difference of their means is greater than zero. An analysis of variance (ANOVA) test using a 95% confidence interval will be used for this test since we are considering more than two variables for comparison, saving us the unnecessary effort of performing multiple two-sample t-tests.

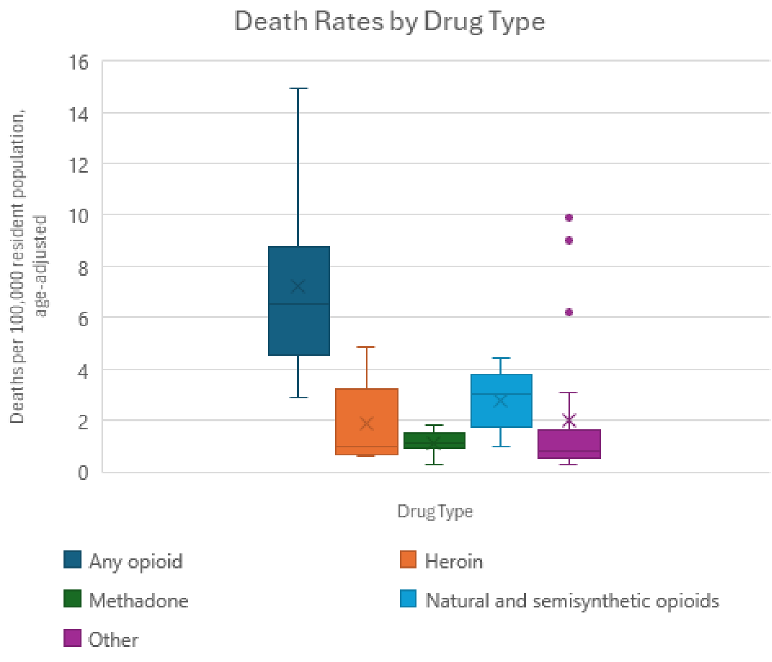


Figure 5: Box-whisker plot of death rates by drug type

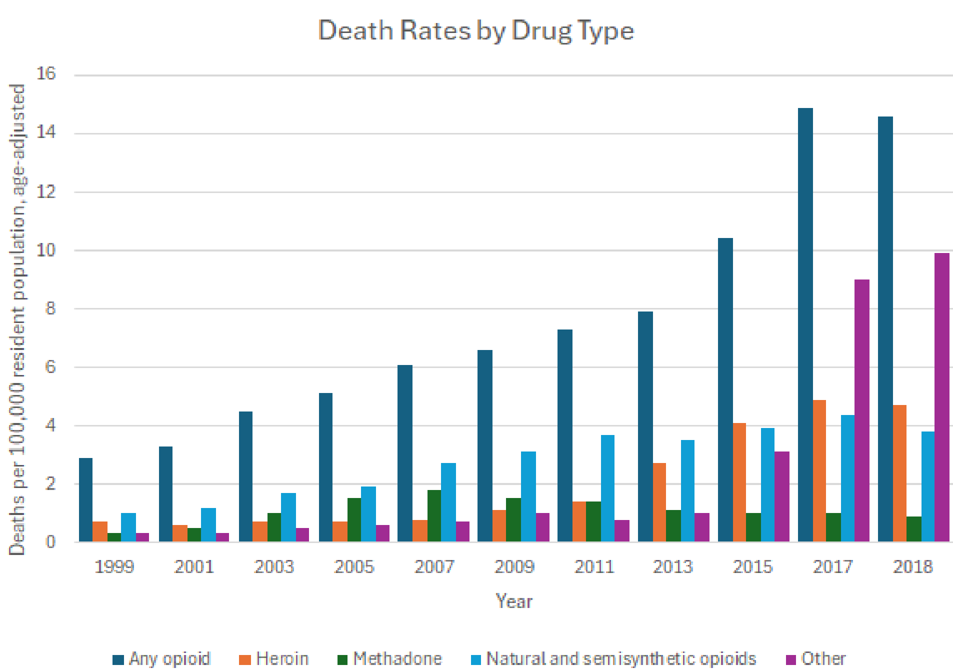
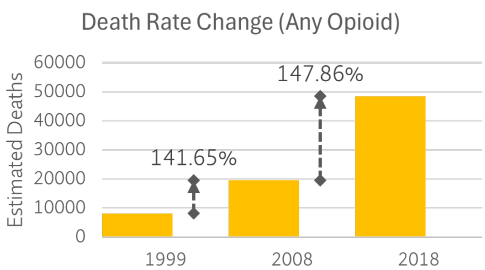
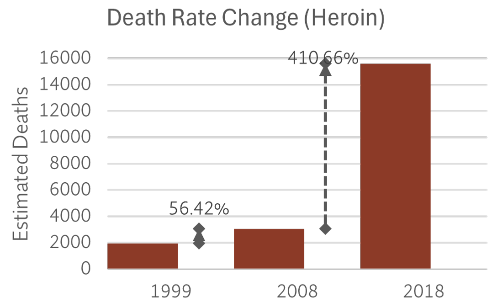
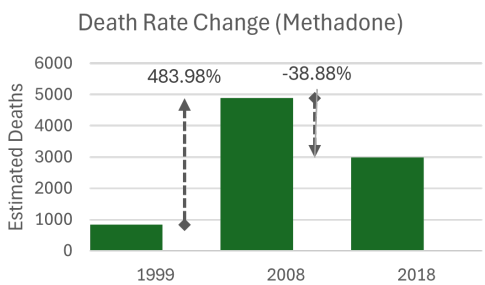
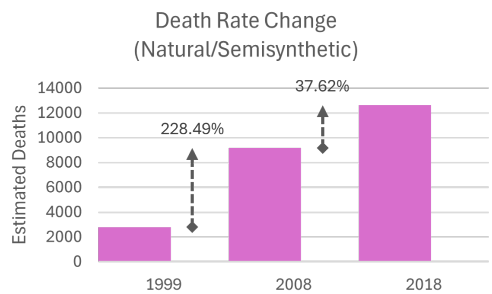


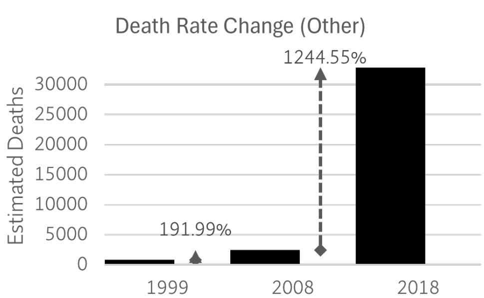
Figure 6: Clustered bar plot of death rates by drug type

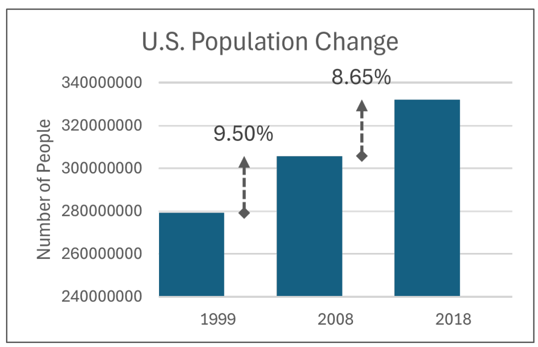












Above are visual representations of the individual change in death rates of each drug type over time. Also, the United States population over time is shown to roughly compare with the proportional relationship of the death numbers and overall population. [2] [3]

The following ANOVA results are from the Data Analysis Toolpak in Excel and R, respectively:

With the p-value of $0.0000000009971 < 0.05$, we reject the null hypothesis and state that there is sufficient evidence at the 0.05 level to conclude a statistically significant difference among the drug type means. In other words, drug type makes a difference in average death rate.

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Any opioid	20	144.2	7.21	13.06515789
Heroin	20	37.3	1.865	2.593973684
Methadone	20	22.7	1.135	0.183447368
Natural and semisynthetic opioids	20	55.7	2.785	1.302394737
Other	20	40	2	8.311578947

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	470.7354	4	117.68385	23.11464787	2.32678E-13	2.467493623
Within Groups	483.6745	95	5.091310526			
				> F crit		
Total	954.4099	99				

Figure 7: ANOVA results: Excel Data Analysis Toolpak

```
oneway.test(Estimate~Type, var.equal = FALSE):

One-way analysis of means (not assuming equal variances)

data: Estimate and Type
F = 21.712,
num df = 4.000,
denom df = 41.544,
p-value = 9.971e-10
```

Figure 8: ANOVA results: R

2.3 Hypothesis 3: Sex and Race

The following hypotheses are formulated for this test:

$$\begin{aligned} &H_0 : \mu_0 = \mu_1 = \mu_2 = ... = \mu_k \\ &H_1 : \text{At least one } \mu_i \text{ is different from the others} \end{aligned} \tag{3}$$

We will be looking at the explanatory variables of sex and race ({male, female}×{White, Black/African American, American Indian/Alaska Native, Asian/Pacific Islander}) and response variable of the drug overdose death rates for this test. The null hypothesis, H_0, states that there is no difference in drug overdose death rates for grouping of sex and race ($\mu_0 = \mu_1 = \mu_2 = ... = \mu_k$). In other words, the means are equal for groupings of sex and race or the difference of their means is equal to zero. The alternate hypothesis, H_1, states that there is a difference in drug overdose death rates for grouping of sex and race (At least one μ_i is different from the others). In other words, the means are not equal for groupings of sex and race or the difference of their means is greater than zero. A regression test will be performed for this test with the goal of assessing the relationship of how sex and race affect the drug overdose death rates. Also, an analysis of variance (ANOVA) test using a 95% confidence interval will be used for this test since we are considering more than two variables for comparison, saving us the unnecessary effort of performing multiple two-sample t-tests.

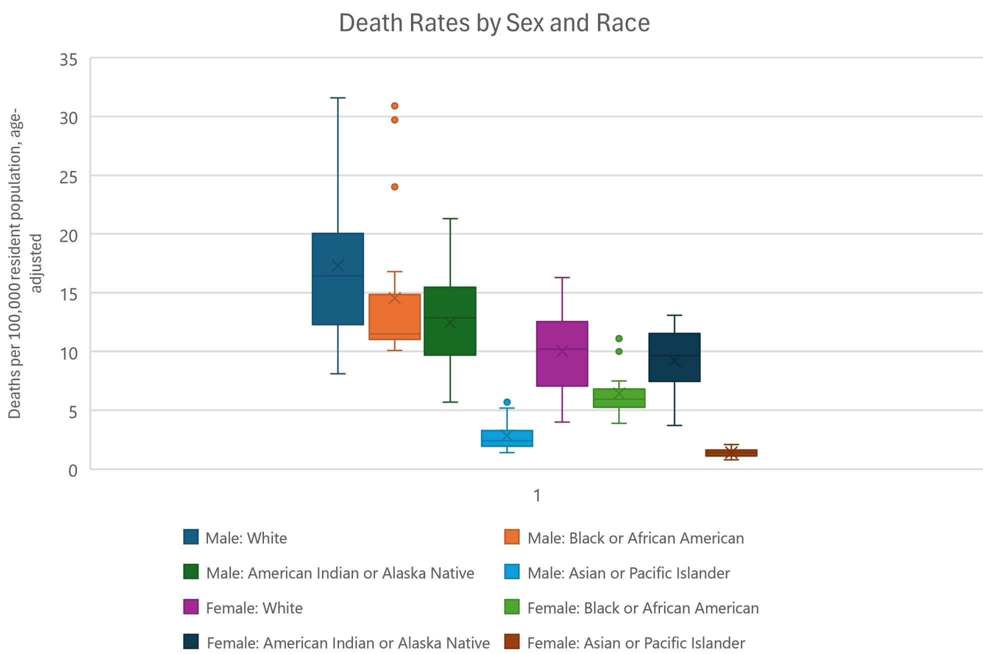


Figure 9: Box-whisker plot of death rates: sex and race

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.766461434
R Square	0.58746313
Adjusted R Square	0.570365405
Standard Error	4.247975229
Observations	160

$$y_s = 11.145 + 5.025x$$
$$y_b = 11.145 - 3.1775x$$
$$y_{aian} = 11.145 - 2.81x$$
$$y_{api} = 11.145 - 11.5375x$$

Sex	White	Black	American Indian	Asian	Estimate
1	1	0	0	0	8.1
1	1	0	0	0	8.4
1	1	0	0	0	9.2

ANOVA

	df	SS	MS	F	Significance F
Regression	5	3983.02925	796.60585	55.18099829	1.26396E-32
Residual	155	2797.0205	18.04529355		
Total	160	6780.04975			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	11.145	0.750943023	14.84133904	1.44318E-31	9.661596817	12.62840318	9.661596817	12.62840318
Sex	5.025	0.671663858	7.481420858	5.13082E-12	3.698203858	6.351796142	3.698203858	6.351796142
White	0	0	65535	#NUM!	0	0	0	0
Black	-3.1775	0.949876138	-3.34517299	#NUM!	-5.053873099	-1.301126901	-5.053873099	-1.301126901
American Indian/AK	-2.81	0.949876138	-2.958280441	0.003578903	-4.686373099	-0.933626901	-4.686373099	-0.933626901
Asian	-11.5375	0.949876138	-12.14632049	2.81814E-24	-13.4138731	-9.661126901	-13.4138731	-9.661126901

Figure 10: Regression with Excel: sex and race

```
lm(formula = Estimate ~ Sex + White + Black + AmericanIndian +
    Asian, data = sexraceregress)

Residuals:
   Min     1Q Median     3Q    Max
-8.070 -2.441 -0.460  1.893 17.907

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.3925     0.7509  -0.523    0.602
Sex              5.0250     0.6717   7.481 5.13e-12 ***
White           11.5375     0.9499  12.146  < 2e-16 ***
Black            8.3600     0.9499   8.801 2.50e-15 ***
AmericanIndian   8.7275     0.9499   9.188 2.47e-16 ***
Asian                NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.248 on 155 degrees of freedom
Multiple R-squared:  0.5875,    Adjusted R-squared:  0.5768
F-statistic: 55.18 on 4 and 155 DF,  p-value: < 2.2e-16
```

Figure 11: Regression with R: sex and race

Anova: Two-Factor With Replication

SUMMARY *White*	Male	Female	Total
Count	20	20	40
Sum	346	200.3	546.3
Average	17.3	10.015	13.6575
Variance	46.26526	13.40134	42.67635

American Indian or Alaska Native	Male	Female	Total
Count	20	20	40
Sum	249.5	184.4	433.9
Average	12.475	9.22	10.8475
Variance	18.55776	8.137474	15.72204

Black or African American	Male	Female	Total
Count	20	20	40
Sum	290.8	128.4	419.2
Average	14.54	6.42	10.48
Variance	38.72147	4.322737	37.87651

Asian or Pacific Islander	Male	Female	Total
Count	20	20	40
Sum	56.8	28	84.8
Average	2.84	1.4	2.12
Variance	1.539368	0.123158	1.341641

Total	Male	Female
Count	80	80
Sum	943.1	541.1
Average	11.78875	6.76375
Variance	55.2719	17.76639

Figure 12: ANOVA summary statistics with Excel: sex and race

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Sample	2973.004	3	991.0014	60.48751	8.59E-26	2.664107
Columns	1010.025	1	1010.025	61.64864	6.86E-13	3.903366
Interaction	306.7175	3	102.2392	6.240346	0.000504	2.664107
Within	2490.303	152	16.38357			
Total	6780.05	159				

Figure 13: ANOVA results with Excel: sex and race

```
             Df Sum Sq Mean Sq F value    Pr(>F)    
SEX           1 1010.0  1010.0   61.65 6.86e-13 ***
RACE          3 2973.0   991.0   60.49  < 2e-16 ***
SEX:RACE      3  306.7   102.2    6.24 0.000504 ***
Residuals   152 2490.3    16.4
```

```
> anova(fit)
Analysis of Variance Table

Response: Estimate
                Df  Sum Sq Mean Sq F value    Pr(>F)    
Sex              1 1010.02 1010.02  55.972 5.131e-12 ***
White            1 1023.75 1023.75  56.732 3.864e-12 ***
Black            1  425.87  425.87  23.600 2.877e-06 ***
AmericanIndian   1 1523.39 1523.39  84.420 2.473e-16 ***
Residuals      155 2797.02   18.05
```

Figure 14: ANOVA results with R: sex and race

The regression results tell us with the multiple R value of 0.7665 that the linear relationship of the death rates and sex and race is 76.65%. The adjusted R square value will be referenced since we have multiple independent variables. Adjusted R square is 0.5704, so we interpret this as 57.04% of variability of death rates around μ is explained by the corresponding behavior of sex and race. Conversely, 43% of variability is due to random chance or lurking variables. The significance F value is very low which will help support rejecting the null hypothesis.

The ANOVA results support our conclusion to reject the null hypothesis. Thus, there is sufficient evidence at the 0.05 level to conclude a statistically significant difference in drug overdose death rates for grouping sex and race. In other words, sex and race make a difference in average drug overdose death rate.

3 Conclusion

This study brought valuable insight into the topic of drug overdose death rates in the United States. The findings from this study indicate that there must exist a tangible influence on drug overdose deaths in the United States in terms of sex, drug type, and race.

The main takeaways from the statistical tests are:

- There is an association between sex and drug overdose death rates.
- There is a difference in drug overdose death rates amongst the different type of drugs.
- There is a difference in drug overdose death rates for grouping of sex and race.

For the tests on drug type and sex and race, the results may be invalid due to not fully satisfying the required assumptions for using the ANOVA test.

Some key difficulties to highlight from this study are:

- Creating useful visual representations of the data/tests
- Manipulating the data to work for the tests and for seeing any relationships
- Ensuring homogeneity of variances for the ANOVA tests was confusing

In the future, improvements and additions to be made include using a larger dataset with more features and data points, testing for correlations, testing for any biases in the data and results, and using a different test in place of ANOVA if required assumptions cannot be satisfied.

4 References

[1] National Center for Health Statistics, "Drug overdose death rates, by drug type, sex, age, race, and Hispanic origin: United States." `https://data.cdc.gov/NCHS/Drug-overdose-death-rates-by-drug-type-sex-age-rac/95ax-ymtc/about_data`.

[2] U.S. Census Bureau, "RACE. Decennial Census, DEC Redistricting Data (PL 94-171), Table P1." `https://data.census.gov/table/DECENNIALPL2020.P1?g=010XX00US`.

[3] Macrotrends LLC, "U.S. Population 1950-2024." `https://www.macrotrends.net/global-metrics/countries/USA/united-states/population`.