Investigating the Use of Alcohol and Marijuana in the United States

Adriana Wiggins and Helen Moses

Department of Mathematics and Statistics, Carleton College
STAT 260: Introduction to Sampling Techniques
Dr. Katie St. Clair
March 11, 2024

Introduction

Drug use, particularly of those that are deemed to be relatively safe in comparison to hard drugs, is widely accepted in social environments across the world. Though alcohol and marijuana are among the more popular drugs of choice in the United States, there exists a difference in social stigma between the two. We are curious how this stigma affects the use of each of these drugs in the United States.

Led by Richard Nixon, the war on drugs in the United States had an incredible impact on the legacies of incarceration and drug use since the 1970s. This initiative criminalized the use of marijuana in order to incarcerate African Americans on a large scale, one that was disproportionate to that of white Americans, and further systemic income inequality. This initiative used racial stereotypes, particularly the stereotype that Black people are dangerous, to reinforce such racial stereotypes, garner support for this abuse of the justice system, and further the generalization that drugs should be feared (Provine, 2011). This, of course, also had an impact on how American voters viewed opposing political parties and candidates. The war on drugs has had a lasting impact on Black households and culture in the United States, which has not occurred by accident.

Following the legalization of marijuana, starting in Colorado and Washington in 2012 (Matthews and Hickey, 2023), the stigma of marijuana began to shift as marijuana became a profitable economic industry. Though the stigma around marijuana has become more positive and socially acceptable, the negative stereotypes and history of marijuana in the context of the war on drugs are still present. By contrast, there is much less of a negative stigma associated with alcohol. The consumption of alcohol is widely acceptable in social, work, and family settings. Despite the risk of death that occurs from alcohol and its misuse, its stigma remains untarnished.

This report investigates the rates of alcohol and marijuana use, in an attempt to investigate how the stigmas surrounding these substances affect their popularity. What proportion of individuals living in the United States, aged twelve and older, have ever once used marijuana? Additionally, what proportion of Americans, aged twelve and older, have ever once used alcohol? We examine these questions using data from the 2021 National Survey on Drug Use and Health (NSDUH), a survey that aims to measure the rates of substance abuse and mental illness in the United States.

Methodology

The Substance Abuse and Mental Health Services Administration (SAMHSA) strives to improve the behavioral health of individuals living in the United States by contributing to and motivating public health efforts (Center for Behavioral Health Statistics and Quality, 2022). A major way that the SAMHSA contributes to public health efforts is via the NSDUH. This survey provides nationally representative data on the use of various drugs as well as a variety of mental health issues. Estimates made based on this data can be used to inform “researchers, clinicians, policymakers, and the general public to better understand and improve the nation’s behavioral

health" (Center for Behavioral Health Statistics and Quality, 2022). The target population for the NSDUH includes individuals living in the United States who are at least twelve years old. More specifically, the population is limited to people in the United States who are not institutionalized. The term institutionalized refers to individuals in both jails, hospitals, and any form of rehabilitation centers. Furthermore, the survey does not attempt to capture the population of homeless individuals who do not use shelters or military personnel on active duty. SAMHSA estimates that the target population makes up roughly 98.3% of the total United States population of individuals at least twelve years old. However, SAMHSA recognizes that the other roughly 1.7% of the population likely have very different rates of substance use and mental health problems compared to the individuals in the target population.

The data in this analysis is from the 2021 sample collection, but the same sampling design was used from 2014-2022. The sampling design allows for repeated sampling of dwelling units (DUs) every two years. Thus, if a DU is selected in one year it is not eligible for selection the following year. However, there is still a chance an individual might be sampled in two consecutive years if they move locations and their new address is selected following the prior year when their old address was selected. While it is important to acknowledge this potential overlap, it does not affect this analysis because the estimates were made using only data from 2021 and no temporal component was added to the analysis. In 2021, data collection was completed four times throughout the year, once per quarter.

To implement this analysis, SAMHSA utilized a complex sampling design that includes stratification and multistage clustering. Figure 1. depicts a visual representation of the implemented sampling design.

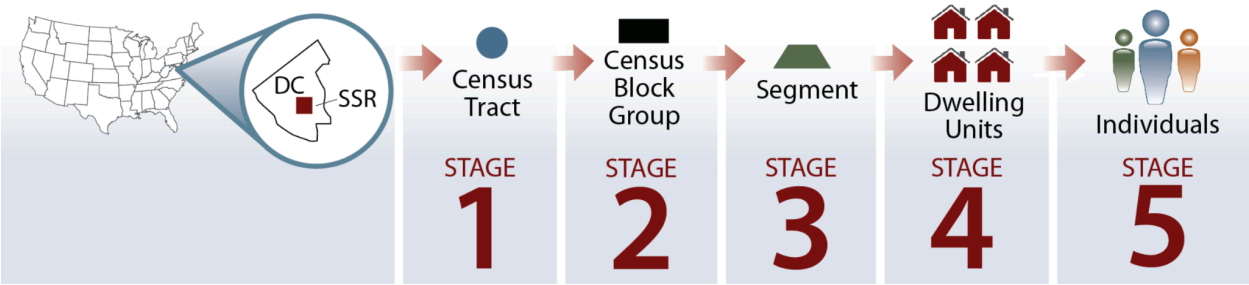

Figure 1. Visual depiction of the stratification and stages of selection implemented by SAMHSA. *This graphic was taken from the 2021 Methodological Summaries and Definitions report.

The following details an overview of the sampling design based on a close reading of Section 2.1 in the 2021 Methodological Summaries and Definitions report produced by SAMHSA. The first level of stratification breaks up the country by each state. The next level of stratification breaks up each state into roughly equally populated regions known as state sampling regions (SSRs). There were a total of 750 SSRs. After these two levels of stratification, the design switched gears to a five-stage cluster sample design. The first stage of cluster sampling involved selecting census tracts within each SSR. The census tracts were selected with probability proportional to size. Within the selected census tracts, the second stage of cluster

sampling involved selecting census block groups. The third stage of the cluster sampling involved selecting area segments, which are a collection of census blocks within the selected census block groups. Every year, there are 48 area segments created within each SSR. Of the 48 area segments, 8 are selected each year with an equal probability and are allocated equally into the four samples used each quarter. In 2021, 4 of the 8 segments were selected for the 2020 survey but were used again. The other 4 selected segments in 2021 will be used in the 2022 sample. Within the selected area segments, the fourth stage of cluster sampling involved selecting DUs. The combination of all area segments selected contained a total of 1,138,827 DUs. Of the total number of DUs, the sampling frame included the 1,021,716 DUs that were determined to be eligible sample units. From the sampling frame, a total of 220,743 DUs were selected in the 2021 sample. Within the selected DUs, the fifth and final stage of cluster sampling involved selecting up to two residents who were at least twelve years old. Table 1. shows the target percentage compared to the achieved percentage of the sample allocation based on age.

Sample	**12 to 17**	**18 to 25**	**26 or Older**	**26 to 34**	**35 to 49**	**50 or Older**
Target	16,877 (25%)	16,877 (25%)	33,753 (50%)	10,126 (15%)	13,501 (20%)	10,126 (155)
Achieved	13,239 (19%)	16,460 (24%)	40,151 (57%)	11,421 (16%)	15,186 (22%)	13,544 (19%)

Table 1. Target versus achieved sample allocation by age group. * This table was taken from the 2021 Methodological Summaries and Definitions report.

Due to the COVID-19 pandemic, the sampling procedures for 2021 were slightly different from previous years in that they included two versions of the survey. One version was the traditional interview procedure. The other version of the survey was a new web-based procedure. In both formats, the individual could complete the survey in English or Spanish. Additionally, consent was obtained from individuals in both formats. Selected individuals under the age of 18 required parental permission along with respondent assent. Both formats started with demographic questions, moved to sensitive questions, and then ended with other demographic questions. The interview procedure involved a field interviewer (FI) visiting the DU of the selected individual and conducting the interview in a private space. The web-based interview procedure involved the individual working in an online format. Thus, the web-based individuals were personally responsible for keeping the interview moving to the next step. Individuals who participated in the web-based format were still reminded to take the interview in a private space. Additionally, a four-number PIN was created by the individual for an added level of security. Each individual who completed the whole interview, regardless of whether it was in-person or web-based, received a $30 cash incentive once the procedure had concluded. After the interviews had concluded, a process ensued that verified the validity of the screening and interview data. If the verification process identified inconsistencies in the responses, modifications were made via the use of a computer program that involved changing a response to allow for consistency.

After the editing of the data was complete, SAMHSA utilized imputation to replace missing values. It is important to note that the variables related to mental health service utilization, suicidal thoughts and behavior among youths, and major depressive episodes among youths were not imputed. The two imputation methods that were used are predictive mean neighborhood and modified predictive mean neighborhood. For more details related to these statistical imputation methods, refer to Section 2.3.3.1 of the 2021 Methodological Summaries and Definitions report produced by SAMHSA.

To allow for an easier estimation process utilizing statistical software, SAMHSA created an ID variable, a stratification variable, and a weighting variable. We used these variables in conjunction with the Survey package in R to obtain our estimates. Details of this procedure can be found in the Technical R Appendix. SAMHSA created the weighting variable to reflect the probability of selection based on the five stages of cluster sampling described earlier in this section. Additionally, weight adjustment factors were used to reduce nonresponse bias, to poststratify based on auxiliary variables, and to control for extreme weights. The weight adjustments were calculated using a generalization of Deville and Särndal's (1992) logit model. More details describing the weight adjustment calculations based on the logit model can be found in Section 2.3.4.1 of the 2021 Methodological Summaries and Definitions report produced by SAMHSA.

Beyond the general weighting methods described above, there were several additional weighting adjustments specific to the 2021 sample. Because the web-based interviews did not allow the eligibility to be known if there was no response in the DU selected, imputation was used to assign an eligibility rate if it was unknown based on the historical eligibility rate in that state. Additionally, educational attainment was added as a variable to be used for stratification adjustment because adults in the 2021 NSDUH were more likely than adults in the 2019 American Community Survey to be college graduates. Finally, break-off analysis weights were created to account for individuals completing the web-based survey who successfully completed the substance use section of the survey but did not finish the other sections.

Note: for any other clarifications or details pertaining to the methodology used for this analysis, consult the 2021 Methodological Summaries and Definitions report produced by SAMHSA.

Results

Figure 2. displays two bar plots: the proportion of people who have ever consumed alcohol and the proportion of people who have ever consumed marijuana. A strong majority of respondents have consumed alcohol at least once, but a minority of respondents have consumed marijuana at least once.

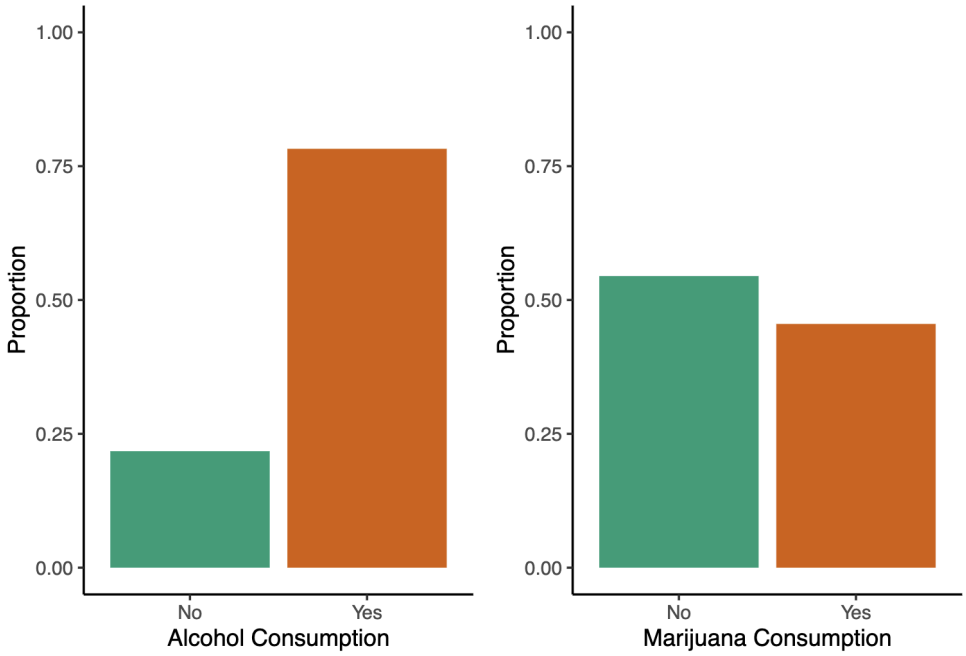

Figure 2. Left side: bar chart of the proportion of individuals who have consumed alcohol at least once based on the NSDUH data. Right side: bar chart of the proportion of individuals who have consumed marijuana at least once based on the NSDUH data.

Based on the survey data collected, we estimate that 78.24% of people over the age of twelve in the United States have consumed alcohol at least once $\hat{p}_{alc} = 0.7824$ with a $SE(\hat{p}_{alc}) = 0.0036$. We are 95% confident that the true proportion of people over the age of twelve in the United States who have consumed alcohol at least once is anywhere between 0.7751 and 0.7897. We also estimate that 45.53% of people over the age of twelve in the United States have consumed marijuana at least once $\hat{p}_{mar} = 0.4553$ with a $SE(\hat{p}_{mar}) = 0.0047$. We are 95% confident that the true proportion of people over the age of twelve in the United States who have consumed marijuana at least once is anywhere between 0.4458 and 0.4647. From these results, it appears that there is a higher proportion of people who have consumed alcohol at least once compared to marijuana. However, these statistics are not independent and were constructed from the same sample rendering any rigorous comparison between estimated proportions improper.

Summary

The results of our investigation match our intuition, that the proportion of people in the US aged twelve or older who have tried alcohol at least once is higher than the proportion of people in the US aged twelve or older who have tried marijuana at least once. Though we cannot identify the source of this difference, it is possible that the difference in stigmas between the two

substances contributed to this difference in proportions. It is also possible that as a result of this stigma, the use of alcohol is normalized from a young age and young people are more inclined to try alcohol than marijuana, even if they're underage.

Our estimates were computed using ratio estimation because we do not know the total number of people aged twelve or older in the US at the time of this data collection. Thus, these estimates are inherently biased. Though the scheme of the sampling weights in this sampling design is not publicly available, adults were more likely to respond because their participation was not subject to parental consent. There is also nonresponse in this data, inevitably due to the nature of a survey, but as described in the *Methodology* section, weights were created to adjust for both nonresponse bias and the difference in the rate of response of individuals in the sample compared to the rate of individuals in the population. However, the weight adjustments used in this analysis will not fully remove the nonresponse bias or the effect of a difference in the rate of response.

References

Center for Behavioral Health Statistics and Quality. (2023). *2021 National Survey on Drug Use and Health Public Use File Codebook*. Substance Abuse and Mental Health Services Administration, Rockville, MD. https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/NSDUH-2021/NSDUH-2021-datasets/NSDUH-2021-DS0001/NSDUH-2021-DS0001-info/NSDUH-2021-DS0001-info-codebook.pdf

Center for Behavioral Health Statistics and Quality. (2022). *2021 National Survey on Drug Use and Health (NSDUH): Methodological summary and definitions.* Substance Abuse and Mental Health Services Administration. https://www.samhsa.gov/data/sites/default/files/reports/rpt39442/2021NSDUHMethodSummDefs100422.pdf

Deville, J.-C., & Särndal, C.-E. (1992). *Calibration estimators in survey sampling*. Journal of the American Statistical Association, 87, 376-382. https://doi.org/10.1080/01621459.1992.10475217

Matthews, A. L. (2023, November 7). *More US states are regulating marijuana. See where it's legal across the country.* CNN. https://www.cnn.com/us/us-states-where-marijuana-is-legal-dg/index.html#:~:text=The%20movement%20to%20legalize%20has,to%20approve%%20recreational%20use.

Provine, D. M. (2011, August 4). *Race and Inequality in the War on Drugs.* Annual Review of Law and Science. https://www.annualreviews.org/doi/abs/10.1146/annurev-lawsocsci-102510-10544520legal

Technical R Appendix

Adriana Wiggins and Helen Moses

2024-03-11

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.0
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.2.1     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(survey)
```

```
## Loading required package: grid
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Loading required package: survival
##
## Attaching package: 'survey'
##
## The following object is masked from 'package:graphics':
##
##     dotchart
```

```
library(patchwork)
library(RColorBrewer)
```

```
# Obtaining the data set using the code from the 'Analyze Survey Data for Free'
# website
zip_tf <- tempfile()

zip_url <-
    paste0(
        "https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/" ,
        "studies/NSDUH-2021/NSDUH-2021-datasets/NSDUH-2021-DS0001/" ,
        "NSDUH-2021-DS0001-bundles-with-study-info/NSDUH-2021-DS0001-bndl-data-r_v3.zip"
    )
```

```
download.file( zip_url , zip_tf , mode = 'wb' )

nsduh_rdata <- unzip( zip_tf , exdir = tempdir() )

nsduh_rdata_contents <- load( nsduh_rdata )

nsduh_df_name <- grep( 'PUF' , nsduh_rdata_contents , value = TRUE )

nsduh_df <- get( nsduh_df_name )

names( nsduh_df ) <- tolower( names( nsduh_df ) )

nsduh_df[ , 'one' ] <- 1
```

```
# Setting up our survey design as done in the 'Analyze Survey Data for Free'
# website
nsduh_design <-
    svydesign(
        id = ~ verep ,
        strata = ~ vestr_c ,
        data = nsduh_df ,
        weights = ~ analwt_c ,
        nest = TRUE
    )

# Re-coding some variables for ease of analysis
nsduh_df$alcever <- ifelse(nsduh_df$alcever==1, 1, 0)
nsduh_df$mjever <- ifelse(nsduh_df$mjever==1, 1, 0)

# Updating our design
nsduh_design <- update (nsduh_design,
                        alcever = nsduh_df$alcever,
                        mjever=nsduh_df$mjever)

# Estimating the proportion of individuals who have ever consumed alcohol
alc <- svymean(~alcever, nsduh_design, df=degf(nsduh_design))
alc
```

```
##            mean     SE
## alcever 0.78243 0.0036
```

```
confint(alc, df=degf(nsduh_design))
```

```
##            2.5 %    97.5 %
## alcever 0.775153 0.7897084
```

```
# Estimating the proportion of individual who have ever consumed marijuana
mj <- svymean(~mjever, nsduh_design, df=degf(nsduh_design))
mj
```

```
##           mean     SE
## mjever 0.45527 0.0047
```

```
confint(mj, df=degf(nsduh_design))
```

```
##            2.5 %    97.5 %
```

```
## mjever 0.4458235 0.4647174
```

```r
# Re-coding the variables again for plot readability
nsduh_df$alcever <- ifelse(nsduh_df$alcever==1, "Yes", "No")
nsduh_df$mjever <- ifelse(nsduh_df$mjever==1, "Yes", "No")

# Bar plot of weighted alcohol consumption proportions
p1 <- ggplot(nsduh_df, aes(x = alcever)) +
  geom_bar(aes(weight = analwt_c/sum(analwt_c ), fill = alcever)) +
  ylim(0,1) +
  labs(y = "Proportion", x = "Alcohol Consumption") +
  theme_classic() +
  theme(legend.position = "none") +
  scale_fill_brewer(palette="Dark2")

# Bar plot of weighted alcohol consumption proportions
p2 <- ggplot(nsduh_df, aes(x = mjever)) +
  geom_bar(aes(weight = analwt_c/sum(analwt_c ), fill = mjever)) +
  ylim(0,1) +
  labs(y = "Proportion", x = "Marijuana Consumption") +
  theme_classic() +
  theme(legend.position = "none") +
  scale_fill_brewer(palette = "Dark2")

p1+p2
```