Study of National Crime Victimization Survey 2020

Hongyan Wang
whymath@uchicago.edu

December 10, 2024

1 Introduction of the Dataset

The National Crime Victimization Survey (NCVS) is a annual data collected by the U.S. Census Bureau since 1972. It collects information on nonfatal personal crimes, personal and household property crimesto investigate the consequneces of crimes, estimate the numbers and types of crimes that are not reported to the police.

The sampling procedure is a multi-stage and stratified design. Fisrt, the United States is divided into Primary Sampling Units (PSUs), Large and therefore important PSUs are large self-representing (SR) and make up their own sampling strata. Smaller non-self-representing (NSR) PSUs grouped within state with similar NSRs to form strata. After SR PSUs and NSR PSUs are constructed, SR PSUs are automatically included for sampling, while NSR PSUs are sampled with probability proportional to population size. Within sampled PSUs, a systematic random sample of housing units and group quarters (GQs) is selected to meet reliability goals. Since it is nationally-representative and self-reported, it is a observational study.

Based on this sampling process, I think the NCVS data has tried to mitigate bias. The multi-Stage and stratified sampling design hopes to ensure representation across various geographic regions, population sizes, and demographic groups. However, thera are still potential bias. For example, not all selected households or individuals agree to participate. Those who refuse may systematically differ from those who participate, possibly resulting in response bias. Also certain crimes may be underreported. For example, incidents involving forced or unwanted sexual acts are often difficult to talk about and for the `Offender_Known`, people are often don't think of incidents committed by someone they know.

The dataset contains a total of 81 variables, which describe various aspects of household information (such as `Urbanicity`, `Liv_Type`, `Units`), personal details (like `Age`, `Marital`, `Sex`, `ED`), and crime-related data (e.g., `Broken_In`, `Num_Broken_In`, `Vehicle_Theft`). The dataset includes 8,043 records, covering 6,025 households. However, due to the selection of a subset of variables, some duplicates exist within the data. After removing duplicates, 6,493 unique records remain. The subsequent analysis will be based on this cleaned dataset.

2 Characteristics of sample

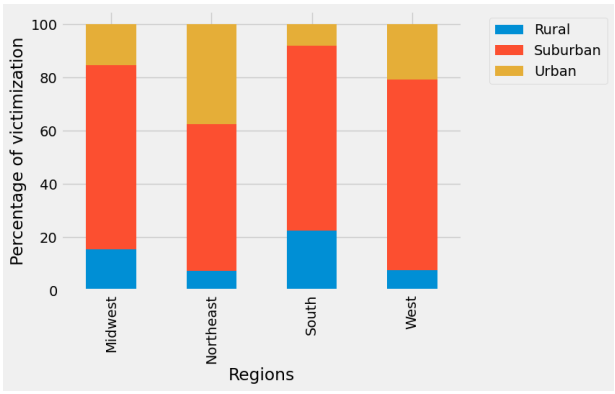
Household variables describe demographic characteristics and structural aspects, such as housing type (`Units`), `Urbanicity`, `Region`, and others. For example, the proportion of `Suburban` households in this dataset is 2.87 times higher than `Urban` households and 3.87 times higher than `Rural` households. When `Urbanicity` is further broken down by `Region` (Fig. 1a), significant regional differences in urbanicity distribution emerge. Specifically, in the `West`, `Suburban` households account for 71.43% of the responses. In the `Northeast`, `Urban` households make up a larger share compared to other regions. In the `South`, `Rural` households significantly outnumber `Urban` households, which contrasts with other regions. Additionally, we analyzed `Units` (Fig. 1b), revealing that most households or individuals live in `Single-family` homes (66.4%) or apartments with `Ten or more` units (16.5%).

In addition to demographic variables, the data also includes socioeconomic variables, such as `Income` (Fig. 2a). The income distribution initially rises and then decreases after reaching the $50,000-74,999$ range. This trend generally aligns with census data, though there are some differences in the details (Fig. 2b). We downloaded U.S. household income data from the U.S. Census Bureau [1] and reorganized the NCVS data for comparison. For households with incomes below $75,000, their representation in the NCVS data is higher than in the census data.

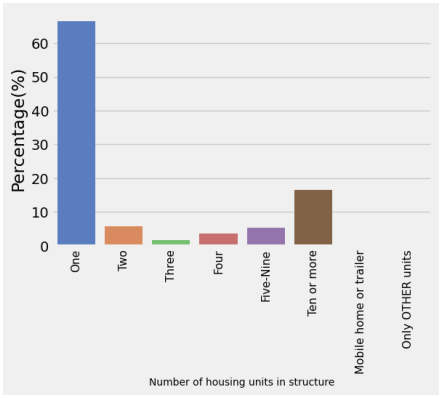
Person variables variables provide a comprehensive understanding of individuals' characteristics and their experiences with crime. Here, we summarized the distribution of `Sex`, `Marital`, `Ethnicity`(Fig. 3). The percentage of `Female` is higher than `Male`, with ratio equals 1.11:1, which is higher than the ratio in the general population(100:97[2]) according to the U.S. Census Bureau's data. This suggests that the NCVS data is somewhat skewed towards Females. The distribution

[1] https://www.statista.com/statistics/758502/percentage-distribution-of-household-income-in-the-us/

[2] https://www.census.gov/data/tables/2023/dec/2020-census-demographic-profile.html

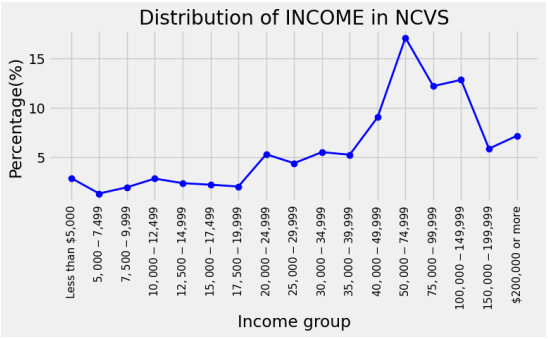


(a) Urbanicity by Regions.

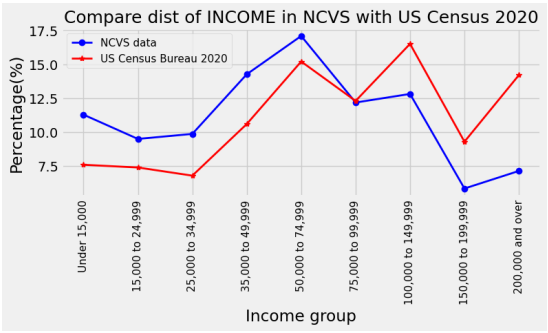


(b) Urbanicity by Regions.

Figure 1: Demographic variables of Household



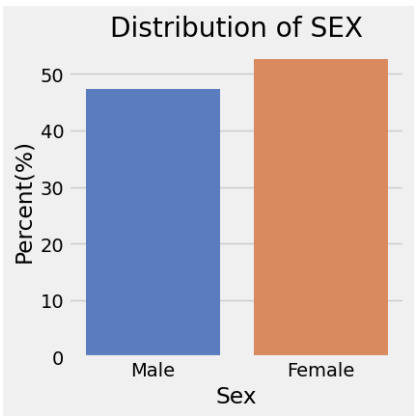
(a) Income distribution of Household.

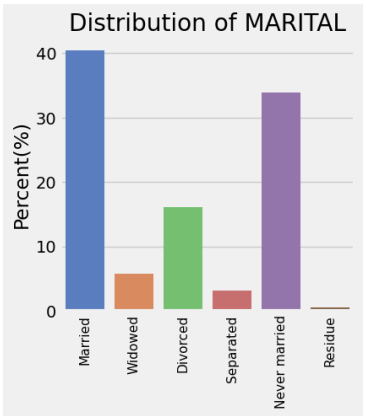


(b) Dist of INCOME in NCVS VS US Census 2020.

Figure 2: Socioeconommic variables of Household

of `Ethnicity` shows that, 13.83% responses are self-reported as `Hispanic`. This also deviates the distribution in general polution which is 18.7% [3].





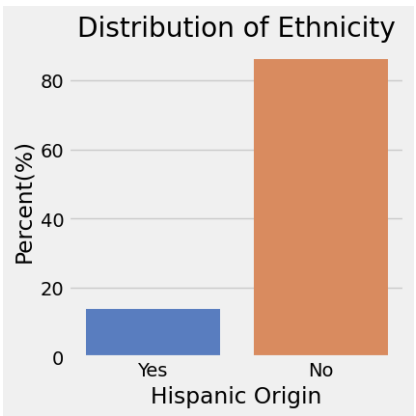
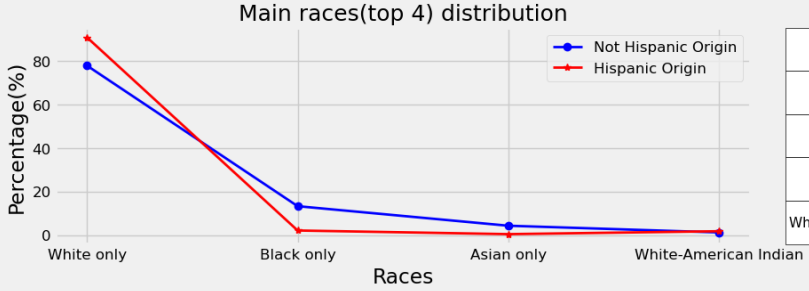


Figure 3: Demographic variables of Person(1)

We also grouped `Sex` by `Ethnicity` and summarized the distributions. The results show a notably high representation of `White-only` individuals in the NCVS data, with a percentage significantly higher than that observed in the 2020 Census data (see the .ipynb file for detailed results). The `Age` distribution doesn't provide much actionable insight on its own, so we further grouped it by `Sex` and `Ethnicity`. This analysis reveals that Hispanic victims tend to be younger(Table 1).

[3] https://www.census.gov/library/stories/2021/08/2020-united-states-population-more-racially-ethnically-diverse-than-2010.html



race	Not Hispanic	Hispanic Origin
White only	77.8	90.76
Black only	13.28	2.12
Asian only	4.31	0.45
White-American Indian	1.23	1.78

Figure 4: Demographic variables of Person(2)

	Female	**Male**	**Not Hispanic**	**Hispanic**
Average age	45.88	44.88	46.23	40.27

Table 1: Average ages across groups

3 Relationship between variables

In this research, we selected four household variables, four personal characteristics variables, and four crime-related variables to examine the relationships between household characteristics and crimes, as well as the relationships between personal characteristics and crime. Since these variables are categorical, we used the Chi-square test to assess whether pairs of variables are independent and to calculate the associated P-values. The results are presented in the following two tables.

	STOLEN	**WEAPON**	**FORCED_SEX**	**POLICE**
URBANICITY	$4.06 \times 10^{-10***}$	2.35×10^{-1}	3.22×10^{-1}	3.09×10^{-1}
REGION	5.56×10^{-1}	6.75×10^{-1}	6.73×10^{-1}	1.21×10^{-1}
UNITS	6.93×10^{-2}	1.51×10^{-1}	$3.25 \times 10^{-5***}$	2.96×10^{-1}
INCOME	1.27×10^{-1}	9.25×10^{-2}	1.43×10^{-1}	$2.12 \times 10^{-3***}$

Table 2: Dataset Summary with Scientific Notation

	STOLEN	**WEAPON**	**FORCED_SEX**	**POLICE**
SEX	7.96×10^{-1}	8.81×10^{-1}	$5.72 \times 10^{-7***}$	$1.32 \times 10^{-2*}$
RACE	4.63×10^{-1}	5.22×10^{-1}	9.04×10^{-1}	8.32×10^{-1}
HISP	9.05×10^{-1}	4.14×10^{-1}	7.19×10^{-1}	1.22×10^{-1}
ED	$1.47 \times 10^{-2*}$	1.01×10^{-1}	7.81×10^{-1}	$3.24 \times 10^{-2*}$

Table 3: Analysis Results by Variable

If we take $\alpha = 0.05$ as the significance level, which mean when P-value ≤ 0.05, the null hypotheses(independent) is rejected, we find seven pairs might be dependent(Table 2 and Table 3). Three of them show a dependency between household variables and crimes: `Urbanicity` and `Stolen`, `Units` and `Forced_Sex`, and `Income` and `Police`. Four others show a dependency between person demographic variables and crimes, which are `Sex` and `Forced_Sex`, `Sex` and `Police`, `ED` and `Stolen`, and `ED` and `Police`.

4 Contex

First, the correlation tests between household variables and crimes indicate that `Urbanicity`, the number of housing `Units` in the structure, and `Income` have significantly different crime rates across subgroups. Specifically, when examining the contingency table for `Urbanicity` and `Stolen`(Table 4), it shows that `Urban` areas exhibit the highest crime rates, while `Rural` areas report the lowest. The variable `Units` is not independent of `Forced_Sex`(Table 6), suggesting that individuals living in apartments with `Ten or more` units or in `Mobile homes or trailers` are more likely to experience higher rates of forced sex crimes. This may be because such living environments are associated with less privacy and increased exposure to risk. Furthermore, the `Forced_Sex` crime disproportionately

affects `Females`(Table 5). There is also evidence that higher levels of education correlate with increased rates of `Stolen` crimes(Table 7).

Urbanicity	STtolen_Yes	Stolen_No
Urban	0.654	0.346
Suburban	0.582	0.418
Rural	0.510	0.490

Table 4: Distribution of Stolen by Urbanicity

Gender	Forced_Sex_Yes	Forced_Sex_No
Male	0.002	0.998
Female	0.014	0.986

Table 5: Distribution of Forced Sex by Gender

Units	Forced_Sex_Yes	Forced_Sex_No
One	0.005	0.995
Two	0.005	0.995
Three	0.009	0.991
Four	0.004	0.991
Five-Nine	0.014	0.986
Ten or more	0.021	0.979
Mobile home or trailer	0.062	0.938
Only OTHER units	0.000	1.000

Table 6: Percentage distribution of Forced Sex by Units.

Education Level	Stolen_Yes	Stolen_No
Nev/kindergarten	0.625	0.375
Elementary	0.522	0.478
High school	0.568	0.432
12th grade (no diploma)	0.582	0.418
High school grad	0.577	0.423
Some college (no degree)	0.568	0.432
Associate degree	0.559	0.441
Bachelor degree	0.596	0.404
Master degree	0.647	0.353
Prof school degree	0.625	0.375
Doctorate degree	0.646	0.354

Table 7: Percentage distribution of stolen by education level.

Statistical independence tests can only tell us if two variables are correlated, not whether they have a causal relationship. Rejecting the null hypothesis does not imply that the two variables are causally related. Causal inference relies on causal assumptions, which are deeply connected to domain-specific knowledge. One of these assumptions is the existence of a sufficient set of confounders to adjust for. Take `ED` (education level) as an example. Education level is not the sole factor related to victim risk. Other socioeconomic factors, such as social status, and neighborhood environment, also play significant roles in influencing vulnerability to crime. In fact, when we test the independence of `ED` and `Stolen` within each `Urbanicity` group, we can no longer reject the null hypothesis(Please refer to the .ipynb for details.).

Many questions remain unanswered with this dataset. For instance, the dependence between `Sex` and `Forced_Sex` appears highly significant. However, is this dependence stable and unbiased? To answer this, we may need time series data, census data to evaluate sample representativeness, and other variables that could mediate the relationship between `Sex` and crime victimization.

5 Conclusion

In this study, I first explored the percentage distributions of several variables of interest. Following this, I examined the data to gain insights into the relationships between household/individual characteristics and crime. Significant dependencies were identified, providing some interesting findings. Beyond these results, this study has been a learning journey, enhancing my understanding of sampling design, high-dimensional data analysis, and interpreting analytical outcomes.

A key lesson is thoroughly understanding the data: its purpose, collection methods, and its representativeness of the population. Clearly defining the research question and ensuring that the data, under the stated assumptions, is sufficient to answer the question is equally essential. When investigating relationships between variables, caution must be exercised to avoid misinterpreting spurious correlations. Advanced techniques, such as causal inference (like stratification), can further strengthen the validity of conclusions. The last but equally important is fairness. Socioscience data involves many sensitive variables. If the analysis is used for informaing public policy, fairness must be addressed.