National Crime Victimization Survey (NCVS)
Datasheet

I. Motivation

I-A For what purpose was the dataset created?

The National Crime Victimization Survey (NCVS) series was designed to achieve four primary objectives [1]:

1) To develop detailed information about the victims and consequences of crime
2) To estimate the number and types of crimes not reported to police
3) To provide uniform measures of selected types of crime
4) To permit comparisons over time and types of areas

I-B Who created the dataset?

Is it an official law enforcement or government body? An academic research team? Other?

The survey is administered by the U.S. Census Bureau (under the U.S. Department of Commerce) on behalf of the Bureau of Justice Statistics (under the U.S. Department of Justice).

I-C Was there a specific task in mind, or gap that needed to be filled?

The NCVS began in 1972 and was developed following a survey done by the National Opinion Research Center and the President's Commission on Law Enforcement and Administration of Justice. The survey highlighted that many crimes were not reported to the police. The NCVS was created to assess the levels of and gain better understanding of criminal victimization, including from crimes that were never reported to law enforcement [1].

II. Composition

II-A What do the instances that comprise the dataset represent?

For example: crimes, offenders, court cases, police officers

Instances in NCVS correspond to a record of a criminal victimization incident.

II-B Are there multiple types of instances?

For example: offenders, victims, and the relationship between them.

Yes, there are four types of records:

1) Address ID Record
2) Household Record
3) Person Record
4) Incident Record

Each person can have multiple incident records, each household can include several persons.

II-C How many instances are there in total?

Of each type, if appropriate.

Data is collected bi-annually, from a nationally representative sample of $\sim 49,000$ households comprising $\sim 100,000$ persons on the frequency, characteristics, and consequences of criminal victimization in the United States. The number of incidents will vary from year to year.

II-D Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

For example, if it is traffic stops from a territory, is it all traffic stops conducted within that territory within a specific time? If not, is it a representative sample of all stops? Describe how representativeness was validated/verified. If it is not representative, please describe why.

The dataset contains a sample of 100,000 persons from the United States. Weights are provided at the person, household, and incident level to produce a representative sample of the US [1]. Excluded are persons who are crews of vessels, in institutions (e.g., prisons and nursing homes) or members of the armed forces living in military barracks [1]. Once in the sample, respondents are interviewed every six months for a total of seven interviews over a three-year period.

II-E What data does each instance consist of?

If there is a large number of variables, please provide a broad description of what is included.

Each instance consists of the following data [1]:

- Type of crime
- Date of crime
- Location type of crime (e.g., at home, at school.)
- Relationship between victim and offender
- Offender characteristics
- Actions taken by the victim (e.g., resisted, escaped.)
- Consequences of victimization (e.g., distress, emotional toll.)
- Type of property lost
- Crime reported
- Reasons for reporting/not reporting
- Weapons used
- Drugs involved
- Alcohol involved
- Demographic information including:
  - Age
  - Race
  - Gender
  - Income

II-F Is there a target label or associated with each instance?
Please include labels that are likely to be used as target labels, e.g. recidivism.

No variable in the dataset is designated as a label. However, whether or not a crime was reported and whether or not an arrest was made following that report may be suitable as target labels [2].

II-G Are there recommended data splits (e.g., training, development/validation, testing)?
If so, please provide a description of these splits, explaining the rationale behind them.

No.

II-H Does the dataset contain data on race and ethnicity?
If so, is it based on the individual's self-description, or based on officer's impression? Was it collected or derived in post-processing? For example, by name analysis.

Yes, race and ethnicity are reported for the victim, and sometimes for the offender as well.

For the victim, that is, the survey respondent, race and ethnicity are self-reported. The race categories are: White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, and Other. Only Hispanic ethnicity is recorded.

For the offender, race and ethnicity are perceived and reported by the respondent, i.e., the victim, if they saw the offender. The race categories are: Mostly White, Mostly Black or African American, Mostly American Indian or Alaska Native, Mostly Asian, Mostly Native Hawaiian or Other Pacific Islander, Equal number of each race, and Don't Know. The ethnicity categories are: mostly Hispanic or Latino, mostly non-Hispanic, equal number of Hispanic and non-Hispanic, and don't know.

II-I Are there any known errors, sources of noise, bias or missing data, or variables collected for only part of the datasets?
If so, please provide a description.

Weights are provided on a person, household and incident level to produce a representative sample of the US. Prisoners are excluded from the sample.

II-J Does the dataset contain data on criminal history or other data that might be considered confidential or sensitive in any way?
For example: sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; If so, please provide a description.

The experiences of criminal victimization themselves and their consequences can be considered sensitive, especially for sexual assault and rape. In addition, the dataset contains information on: age, race, gender, and income.

II-K Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?
If so, please describe how.

No, the data is sufficiently anonymized.

III. Uses

III-A What type of tasks, if any, has the dataset been used for?
If so, please provide examples and include citations.

The dataset has been used for a range of victimization studies, looking at things such as:

- Assessment of crime levels in the United States.
- Comparing Victimization across demographics.
- Comparing Victimization of specific types of crime across demographics.
- Assessing the *dark figure of crime*.[1]

III-B Is there a repository that links to any or all papers or systems that use the dataset?
If so, please provide a link or other access point.

Yes. Papers that cited this dataset can be found in: https://www.icpsr.umich.edu/web/NACJD/series/95/publications.

III-C What (other) tasks could the dataset be used for?
For example: testing predictive policing systems, predicting recidivism.

This dataset can be used for a variety of tasks which requires an understanding of the level of victimization for specific crimes, with demographic information on both the victim and offender, and information on whether the crime was reported and an arrest was made.

III-D Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

According to the authors of a book chapter on the potential sources of error in the NCVS [3], there are questions as to whether the rape and sexual assault are underestimated as they do not align with alternative surveys. "The Bureau of Justice Statistics does not provide public information on the edit process in the National Crime Victimization Survey, although processing and editing errors are an important part of any major survey data collection. The lack of transparency about these processes makes it difficult for data users to fully understand the survey's estimate" [3].

[1]The dark figure of crime is term used to illustrate the extent of committed crimes that are never reported or discovered by law enfrocment.

IV. Collection Process

IV-A How was the data associated with each instance acquired?
e.g. the data collected survey, the raw data is routinely collected by the courts.

The data was acquired from a bi-annual survey.

IV-B Was the information self-reported?
If the data was self-reported, was the data validated/verified? If so, please describe how.

Yes. The data is collected in a survey. However, the raw survey responses are not provided.

IV-C Who was involved in the data collection process?
Was this done as part of their other duties? If not, were they compensated?

The survey is administered by the U.S. Census Bureau.

IV-D Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?
If not, please describe the timeframe in which the data associated with the instances was created. If the collection was not continuous within the timeframe, please specify the intervals, for example, annually, every 4 years, irregularly.

The data is collected twice a year and released once a year. Data is available from the year 1979 and collection is still ongoing.

IV-E Were any ethical review processes conducted (e.g., by an institutional review board)?
If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown.

IV-F Were the individuals in question notified about the data collection? Did they give their consent?
If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

Yes, the survey is optional.

IV-G Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?
If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown.

V. Pre-processing, cleaning, labeling

V-A Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, removal of instances, processing of missing values)?
If so, please provide a description and reference to the documentation. If not, you may skip the remaining questions in this section.

Yes, as the survey responses were processed into the data available in the dataset. However, information on pre-processing is not supplied in the codebook.

V-B Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?
If so, please provide a link or other access point to the "raw" data.

It is not part of the publicly available dataset."

V-C Is the software that was used to preprocess/clean/label the data available?
If so, please provide a link or other access point.

No.

VI. Distribution

VI-A Is the data publicly available? How and where can it be accessed (e.g., website, GitHub)?
Does the dataset have a digital object identifier (DOI)?

The dataset is hosted at:
https://www.icpsr.umich.edu/web/NACJD/series/95.
There are multiple DOIs associated with this dataset, depending on the version and years of collection.

VI-B Is the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?
If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The license is not specified, but a citation and deposit requirement are listed:
Citation Requirement: Publications based on ICPSR data collections should acknowledge those sources by means of bibliographic citations. To ensure that such source attributions are captured for social science bibliographic utilities, citations must appear in footnotes or in the reference section of publications.

Deposit Requirement: To provide funding agencies with essential information about use of archival resources and to facilitate the exchange of information about ICPSR participants' research activities, users of ICPSR data are requested to send to ICPSR bibliographic citations for each completed manuscript or thesis abstract. Visit the ICPSR Web site for more information on submitting citations.

VII. Maintenance

VII-A Is the dataset maintained? Who is supporting/hosting/maintaining the dataset?

The dataset is hosted and supported by the Ministry of Justice and the US Census Bureau.

VII-B How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

By contacting the US Census Bureau.

VII-C Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

New versions of the dataset are released yearly.

VII-D Are older versions of the dataset continue to be supported/hosted/maintained?

Yes. Previous years of the dataset will continue to be hosted by the Ministry of Justice and are avilable to download.

VII-E If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?
If so, please provide a description.

No.

References

[1] "National Crime Victimization Survey (NCVS) Series." [Online]. Available: https://www.icpsr.umich.edu/web/ICPSR/series/95

[2] R. Fogliato, A. K. Kuchibhotla, A. Xiang, Z. Lipton, D. Nagin, and A. Chouldechova, "Estimating the Likelihood of Arrest on Police Records in Presence of Unreported Crimes," 2022.

[3] C. Kruttschnitt, W. D. Kalsbeek, C. C. House, N. R. Council *et al.*, "Potential sources of error in the NCVS: Sampling, frame, and processing," in *Estimating the Incidence of Rape and Sexual Assault*. National Academies Press (US), 2014.