BJS Bureau of Justice Statistics

Direct Estimation Methods and the National Crime Victimization Survey

Alexandra Thompson, BJS

Erika Harrell, PhD, BJS

Marcus Berzofsky, DrPH, RTI

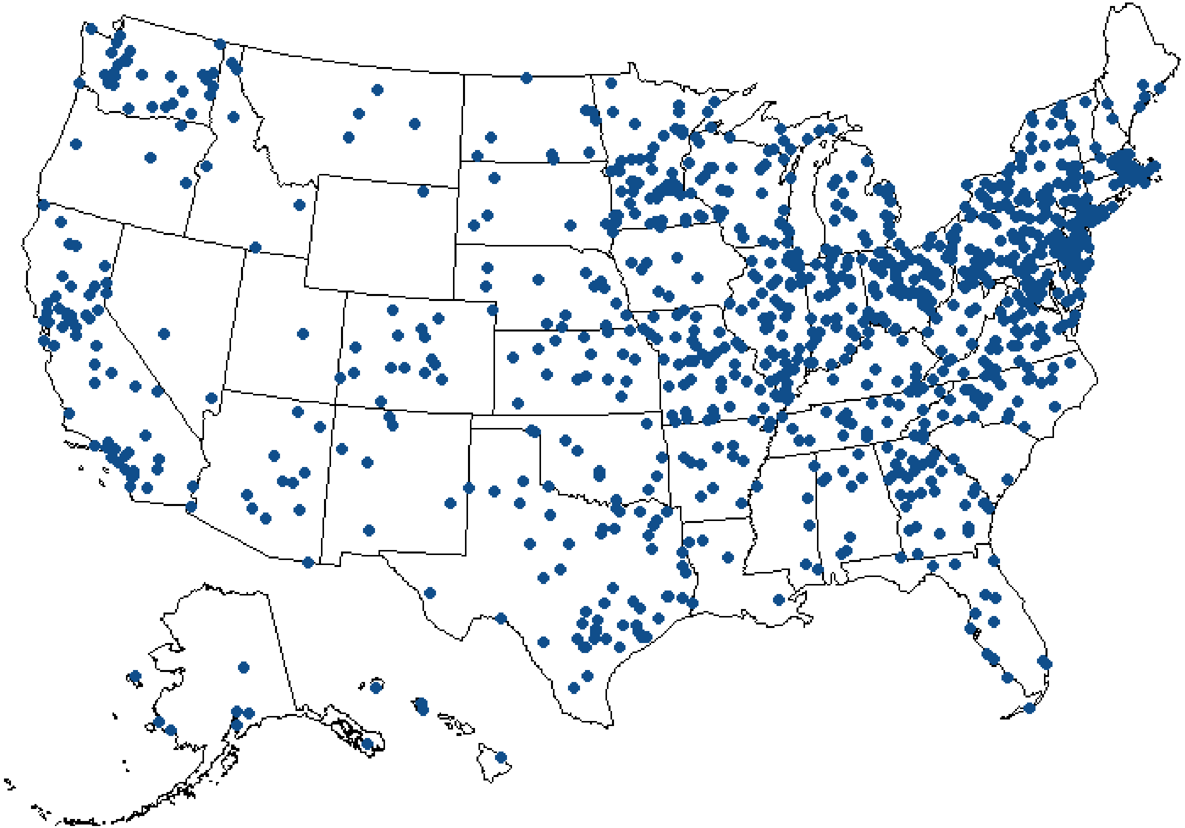
Andrew Moore, RTI

Agenda for today’s webinar

1. Learn about what variance estimation is and why it’s important
2. Differences between direct and indirect variance estimation
   - Including generalized variance function (GVF), Taylor Series Linearization (TSL), and Balanced Repeated Replication (BRR)
3. Two live examples on direct estimation in SAS and SPSS
4. Q&A

What is variance estimation and why is it important?

Sample vs. Census

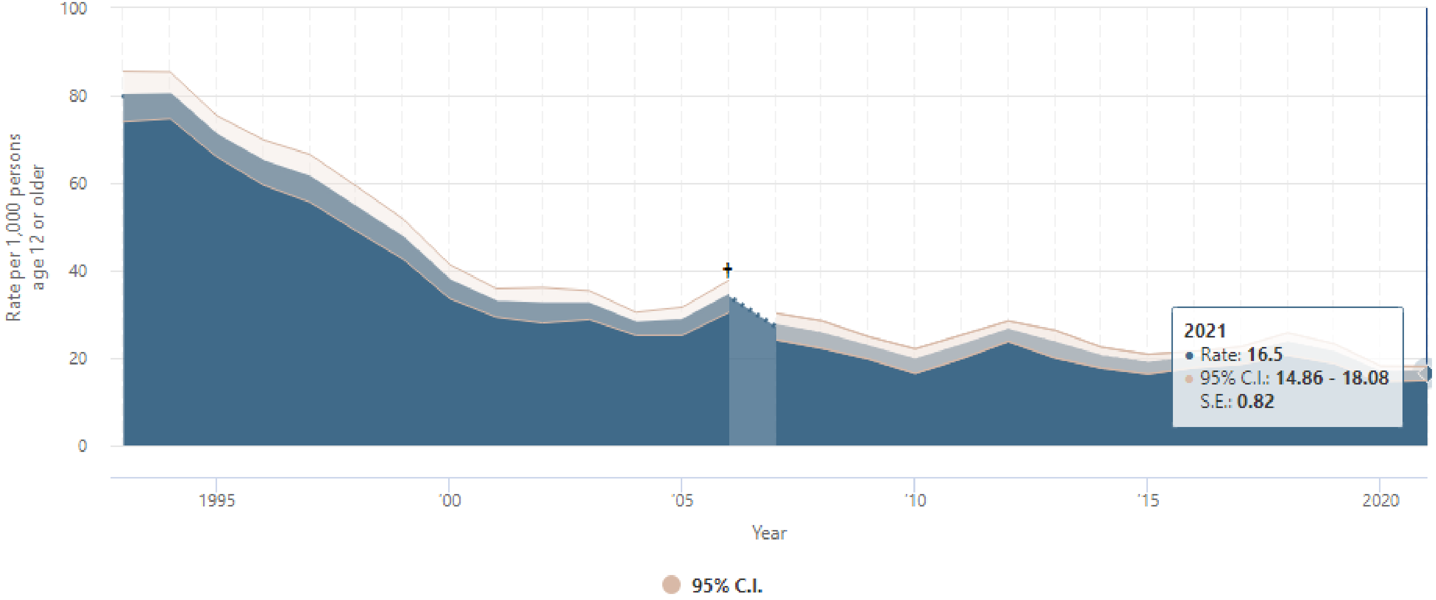


*Points on this map have been randomly generated and do not reflect households or areas sampled by the NCVS.

What is variance?

- Estimates based on a sample have some degree of sampling error. The sampling error of an estimate depends on several factors, including the amount of variation in the responses and the size of the sample.
- The variance for an estimate is a type of sampling error and measures the deviation between the estimate and the average.
- The standard error (SE) is the square root of the variance.
- Standard errors can be used to calculate confidence intervals around an estimate.

Rate of violent victimizations, 1993-2021



"95% C.I.": 95% confidence interval.

"S.E.": Standard error.

† Estimates for 2006 should not be compared to other years. See User's Guide for more information.

In October 2019, BJS released a revised set of 2016 NCVS data. See User's Guide for more information.

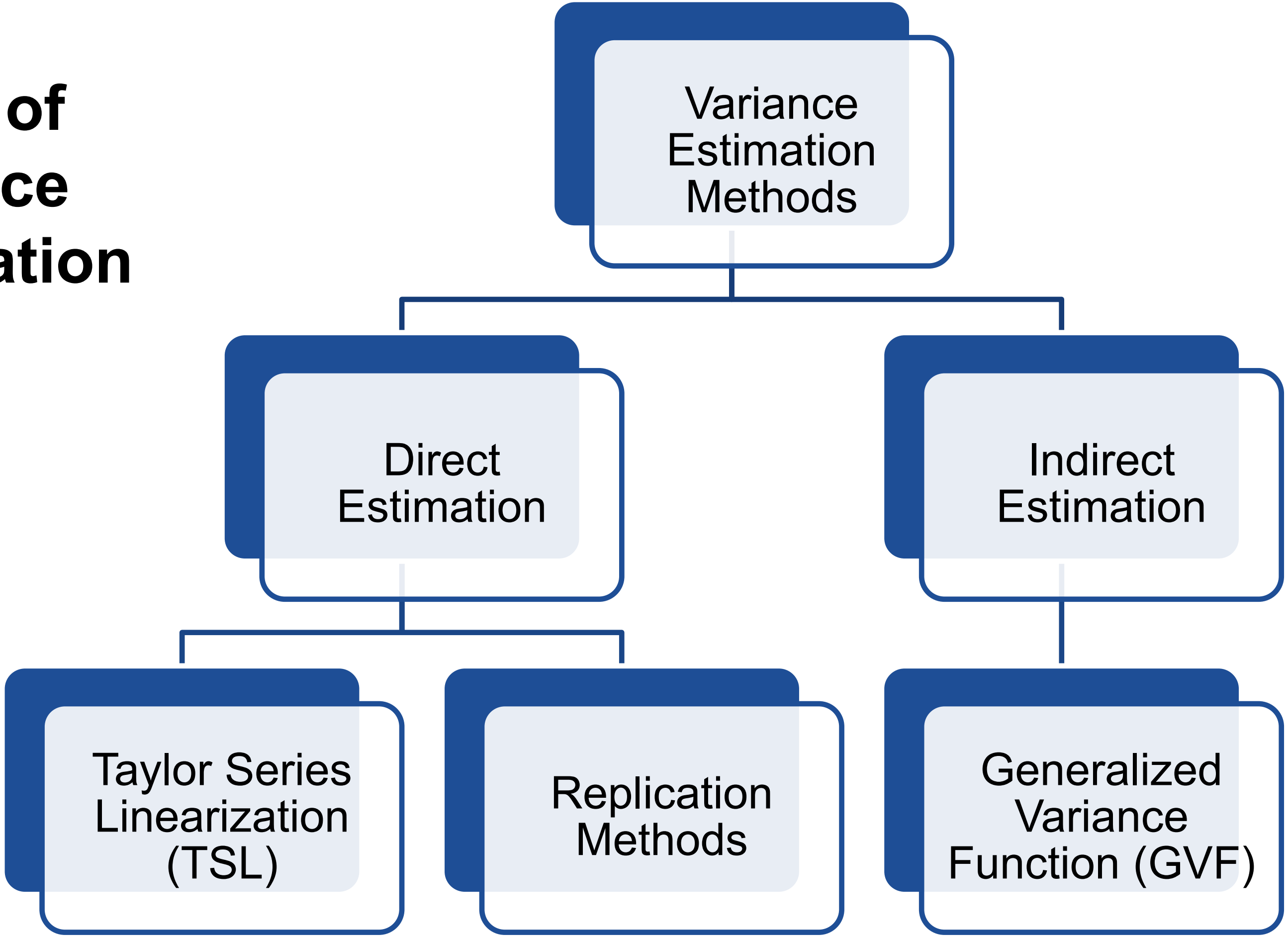
Source: Bureau of Justice Statistics, National Crime Victimization Survey, 1993-2021. https://ncvs.bjs.ojp.gov/quick-graphics

Why is it important?

- Generally, an estimate with a smaller standard error (square root of the variance) provides a more reliable approximation of the true value than an estimate with a larger standard error. Estimates with relatively large standard errors have less precision and reliability and should be interpreted with caution.
- Standard errors help determine whether two estimates are statistically different or not statistically different.
  - In BJS reports, we conduct statistical tests to determine whether differences in estimated numbers, percentages, and rates in this report were statistically significant once the standard error was taken into account.

Types of Variance Estimation

Types of Variance Estimation



Variance Estimation Methods

Direct Estimation

Indirect Estimation

Taylor Series Linearization (TSL)

Replication Methods

Generalized Variance Function (GVF)

Indirect Variance Estimation

How it computes variances?

- A *generalized variance function (GVF)* is produced through a nonlinear model is used to fit the variance estimates
- Users use the resulting model parameters to produce approximation of standard error (correlation parameters needed to compare over time)

Is anything special needed?

- No special software is needed
- Only weighted estimates and model parameters are needed
- Can be computed in Excel

Generalized Variance Function Basics

- Function for an overall total
- $V_t(\hat{t}_D; a, b, c) = a\hat{t}_D^{\,2} + b\hat{t}_D + c\hat{t}_D\sqrt{\hat{t}_D}$
  - a, b, and c are the GVF model parameters
  - $\hat{t}_D$ is the estimated total
- Function for a rate
- $V_r\left(\hat{r}_{C,D}, \widehat{N}_D; b, c\right) = b\frac{\hat{r}_{C,D}\left(1000-\hat{r}_{C,D}\right)}{\widehat{N}_D} + c\frac{\hat{r}_{C,D}\left(\sqrt{1000\hat{r}_{C,D}}-\hat{r}_{C,D}\right)}{\sqrt{\widehat{N}_D}}$
  - $\hat{r}_{c,D}$ is estimated rate per 1,000 for crime c
  - $\widehat{N}_D$ is the weighted population

GVFs are the method traditionally implemented by BJS for the NCVS

Very simple to implement

Especially when file structure is complex like the NCVS

Direct Variance Estimation

How it computes variances?

- Directly from data
- Does not require outside information

Is anything special needed?

- Certain variables on the dataset need to be specified during the estimation process
- The variables to be specified depends on the type of direct variance estimation being used
- Requires statistical software (SAS, SPSS, SUDAAN, R)

Types of Direct Estimation

Taylor Series Linearization

- Utilizes a *population weight* and *design variables*
- Design variables provide details about the complex design such as stratification or clustering (PSUs)

Replication

- Utilizes the *population weight* and *a set of replicate weights*
- Replicate weights are survey weights created when a subset of respondents is excluded and the remaining cases are reweighted to represent the population
- NCVS uses balanced repeated replication (BRR)

Taylor Series Linearization Basics

- NCVS Design Variables
  - PSEUDOSTRATA (V2117)
  - HALFSAMPLE (V2118)
- NCVS Weight Variables
  - Person weight: WGTPERCY
  - Household weight: WGTHHCY
  - Incident weight: WGTVICCY*SERIESWGT

Statistical software can only analyze one dataset at a time

Need to move incident counts from incident file to Household or Person File

Need to specify design variables and population weight in statistical software

BRR Basics

- NCVS Replicate weights
  - HHREPWGTCY1 – HHREPWGTCY160
    - (household replicate weights)
  - PERREPWGTCY1-PERREPWGTCY160
    - (person replicate weights)
- NCVS Weight Variables
  - Person weight: WGTPERCY
  - Household weight: WGTHHCY
  - Incident weight: WGTVICCY*SERIESWGT

Statistical software can only analyze one dataset at a time

Need to move incident counts from incident file to Household or Person File

Need to specify the replicate weights and the population weight

	ADVANTAGES	DISADVANTAGES
Indirect (GVF)	Does not require statistical software package (e.g., SPSS complex survey package) Does not require knowledge of special statistical software Does not require file manipulation of NCVS datasets	Less accurate than TSL/BRR Has to be calculated separately for each outcome Requires additional correlation parameters to make comparisons over time
Direct (BRR/TSL)	More accurate than GVF Can calculate multiple outcomes simultaneously Can conduct comparisons of groups/outcomes over time in same procedure that produces estimates	Requires access to statistical software that can handle complex survey designs Each method has quirks with the NCVS (e.g., no TSL in 2016; BRR more complicated in SPSS) For NCVS, requires file manipulation to combine the incident counts with the person/household file

	ADVANTAGES	DISADVANTAGES
TSL	Computationally faster Easy to implement in SPSS Easier to run analyses that span across Decennial Census Updates	Requires knowledge of sample design For NCVS, requires file manipulation to combine incident counts with the person/household file Cannot run TSL for NCVS in 2016
BRR	Does not require design information Better for disclosure avoidance	Computationally slower More difficult to run in SPSS Cannot be pooled with years that have a different number of replicates Not available on the NCVS concatenated files

Comparison of Design Features: GVF vs. TSL vs. BRR

Feature	GVF	TSL	BRR
Directly estimated from the data		✓	✓
Requires *special* statistical software for complex surveys		✓	✓
Does not require knowledge of special statistical software	✓		
Requires design variables		✓	
Easily run with SPSS	✓	✓	
Requires merging of incidents on person/household files		✓	✓
Can be run for all years[1]	✓		✓
Can be run using concatenated file on ICPSR	✓	✓	
For pooling years, does not require the same number of replicates	✓	✓	

[1] TSL cannot be run in 2016 due to lack of design variables on file

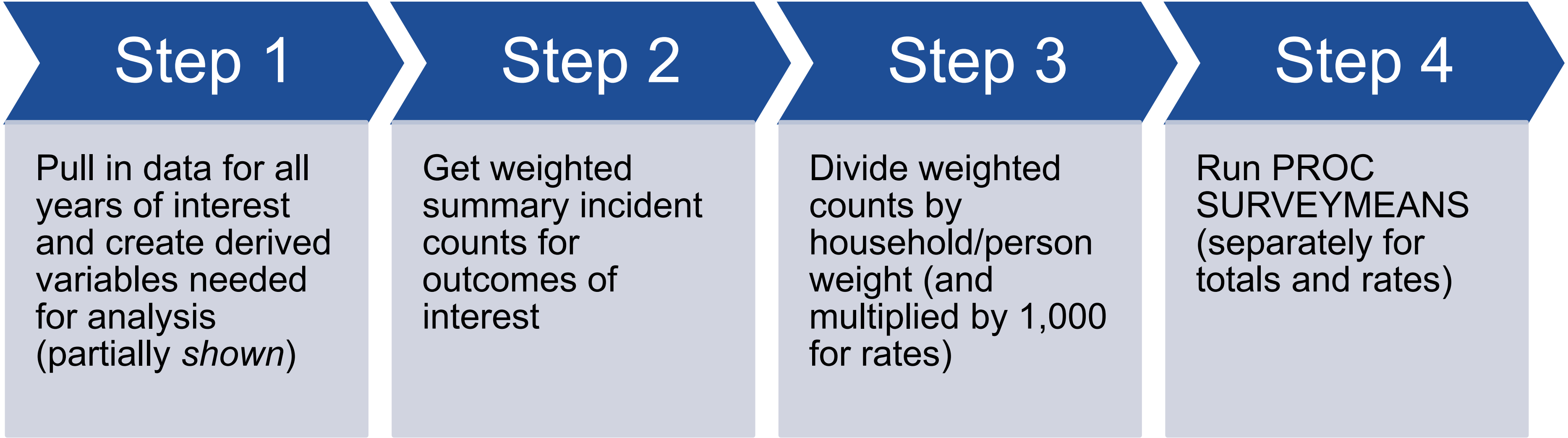
Comparison of Variance Estimation Methods: RSEs (Totals)

	2020			2021		
Crime	GVF	BRR	TSL	GVF	BRR	TSL
Violent Crime	5.5	4.8	5.1	5.0	4.5	4.3
Rape/sexual assault	16.2	15.8	13.3	12.3	12.0	14.7
Robbery	14.1	12.5	11.7	10.8	11.9	13.2
Aggravated assault	5.9	5.2	5.6	5.3	5.2	5.1
Simple assault	10.9	9.1	9.4	9.1	7.7	7.4
Personal theft	6.4	5.7	6.2	5.7	6.0	5.9

NOTE: Relative Standard Errors (RSEs) are the ratio of the standard error and the point estimate times 100

Using BRR to estimate variance in SAS

Steps in Estimation Process (BRR in SAS)



Step 1

Pull in data for all years of interest and create derived variables needed for analysis (partially *shown*)

Step 2

Get weighted summary incident counts for outcomes of interest

Step 3

Divide weighted counts by household/person weight (and multiplied by 1,000 for rates)

Step 4

Run PROC SURVEYMEANS (separately for totals and rates)

Estimation Process: Step 1 (partial for incident datasets)

```
data incident2;
  set incident; /*a*/

   /*b*/
   RSA = (v4529 in (1, 2, 3, 4, 15, 16, 18, 19)); *Rape/Sexual
Assault;
   ROB = (5 <= v4529 <= 10); *Robbery;
   AST = (v4529 in (11, 12, 13, 14, 17, 20)); *Assault;
   SAST = (v4529 in (14, 17, 20)); *Simple Assault;
   AAST = (v4529 in (11, 12, 13)); *Aggravated Assault;
   /*c*/
   VIOLENT = (MAX(RSA,ROB,AST));
   /*d*/
   PTFT = (21 <= v4529 <= 23);*Personal theft;
   /*e*/
   if (v4022 ne 1) then exclude_outUS=0; *exclude incidents
occuring outside of the US;
   else exclude_outUS=1;

o  run;
```

- Comment Annotation
  a) Assumes all years of data have already been set together
  b) Define components of violent crime offense
  c) Define violent crime offense
  d) Define personal theft offense
  e) Exclusions for estimating crimes that occurred in the US

Estimation Process: Step 2

```
*This step creates series weighted
sums for the number of victimizations
per person;
proc means data=incident2 noprint;
  where exclude_outUS=0 and
(violent=1 or PTFT=1); *a;
  by year yearq idhh idper;
  weight series_weight; *b;
  output out=vicsum /*c*/
  sum(VIOLENT RSA ROB AST AAST SAST
PTFT )=
     violent rsa rob AST aast sast
PTFT ; *d;
run;
```

- Comment Annotation
  a) Standard exclusions used by BJS
  b) series incident weight (series_weight= WGTVICCY*serieswgt)
  c) output dataset to be merged back onto person file
  d) List of outcome variables being analyzed

Estimation Process: Step 3

```
data perinc;
  merge vicsum(in=b) person2(in=a keep=year yearq idhh idper wgtpercy wgthhcy pseudostrata
halfsample perrepwgtcy1-perrepwgtcy160 perrepwgt1-perrepwgt160 sex age race region msa) ;
  by year yearq idhh idper; *a;

  *b;
  array viccnts{*} VIOLENT RSA ROB AST AAST SAST PTFT ;
  do i=1 to dim(viccnts);
    if missing(viccnts{i}) then viccnts{i}=0;
  end;
 *c;
  array viccnts2{*} VIOLENT2 RSA2 ROB2 AST2 AAST2 SAST2 PTFT2;
  array viccnts3{*} VIOLENT3 RSA3 ROB3 AST3 AAST3 SAST3 PTFT3;
  do i=1 to dim(viccnts2);
    if WGTPERCY>0 then do;
      viccnts2{i}=(viccnts{i}/WGTPERCY)*1000; /* Rates */
      viccnts3{i}=(viccnts{i}/WGTPERCY); /* Totals */
    end;
    else do;
      viccnts2{i}=0;
      viccnts3{i}=0;
    end;
  end;

  drop i;
run;
```

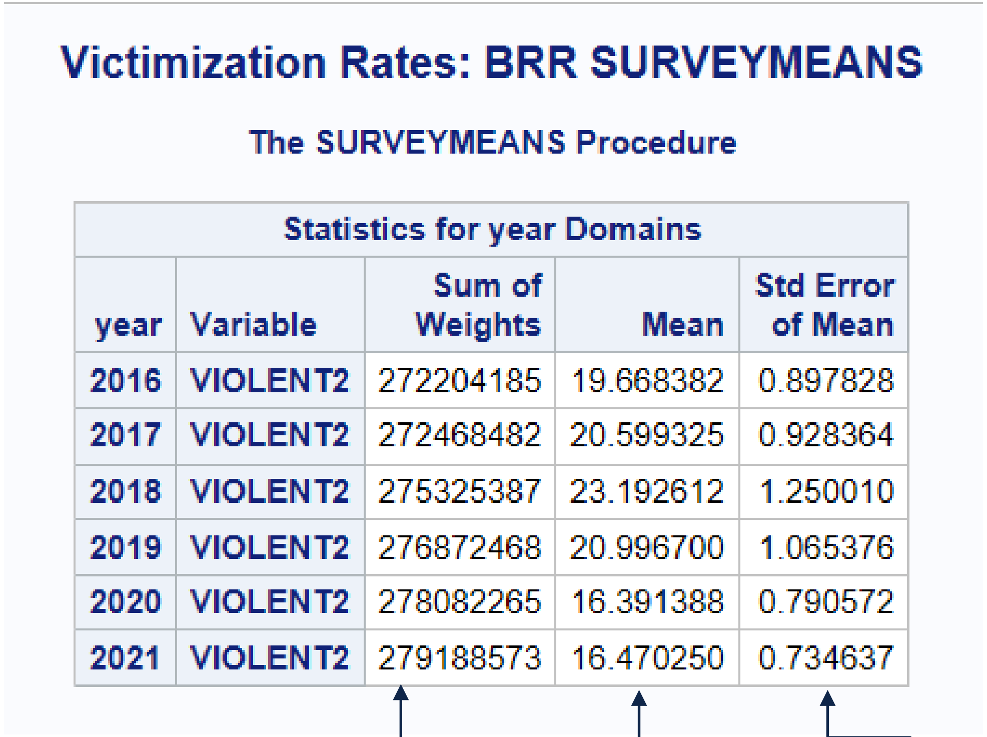
- Comment annotation
  a) Merge summary incident counts onto person file
  b) Set missing incident counts to 0
  c) Divide summary counts by population (and multiply by 1,000 for rates) to prepare for estimation

Estimation Process: Step 4a (RATES)

```
proc surveymeans data=perinc
  varmethod=brr (fay)
  mean sumwgt; *a;
  var VIOLENT2; *b;
  class year;
  domain year; *c;
  ods output domain=est; *d;
  weight wgtpercy; *e;
  repweight perrepwgtcy1-
perrepwgtcy160; *f;
  title "Victimization Rates: BRR
SURVEYMEANS";
run;
```

- Comment Annotation
  a) NCVS uses Fay’s BRR method
  b) List of outcome variable(s) of interest; “2” version used for rates
  c) Specify separate estimates by year
  d) Need to use ODS to save results to an output dataset
  e) Population weight
  f) Replicate weights

Estimation Process: Step 4a Output (Violent Crime Rates)



Victimization Rates: BRR SURVEYMEANS

The SURVEYMEANS Procedure

Statistics for year Domains

year	Variable	Sum of Weights	Mean	Std Error of Mean
2016	VIOLENT2	272204185	19.668382	0.897828
2017	VIOLENT2	272468482	20.599325	0.928364
2018	VIOLENT2	275325387	23.192612	1.250010
2019	VIOLENT2	276872468	20.996700	1.065376
2020	VIOLENT2	278082265	16.391388	0.790572
2021	VIOLENT2	279188573	16.470250	0.734637

Population

Rate

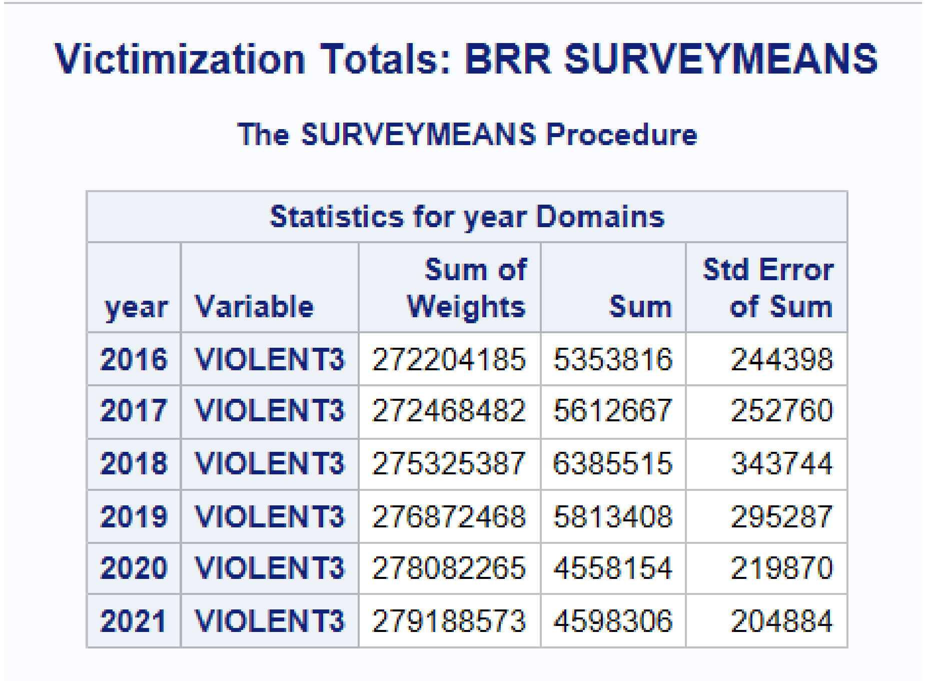
SE of Rate

Estimation Process: Step 4b (TOTALS)

```
proc surveymeans data=perinc
  varmethod=brr (fay)
  sum sumwgt; *a;
  var VIOLENT3; *b;
  class year;
  domain year; *c;
  ods output domain=est_tot; *d;
  weight wgtpercy; *e;
  repweight perrepwgtcy1-
perrepwgtcy160; *f;
  title "Victimization Totals: BRR
SURVEYMEANS";
run;
```

- Comment Annotation
  - a) NCVS uses Fay's BRR method
  - b) List of outcome variable(s) of interest; "3" version used for totals
  - c) Specify separate estimates by year
  - d) Need to use ODS to save results to an output dataset
  - e) Population weight
  - f) Replicate weights

Estimation Process: Step 4b Output (Violent Crime Totals)



Victimization Totals: BRR SURVEYMEANS

The SURVEYMEANS Procedure

Statistics for year Domains				
year	Variable	Sum of Weights	Sum	Std Error of Sum
2016	VIOLENT3	272204185	5353816	244398
2017	VIOLENT3	272468482	5612667	252760
2018	VIOLENT3	275325387	6385515	343744
2019	VIOLENT3	276872468	5813408	295287
2020	VIOLENT3	278082265	4558154	219870
2021	VIOLENT3	279188573	4598306	204884

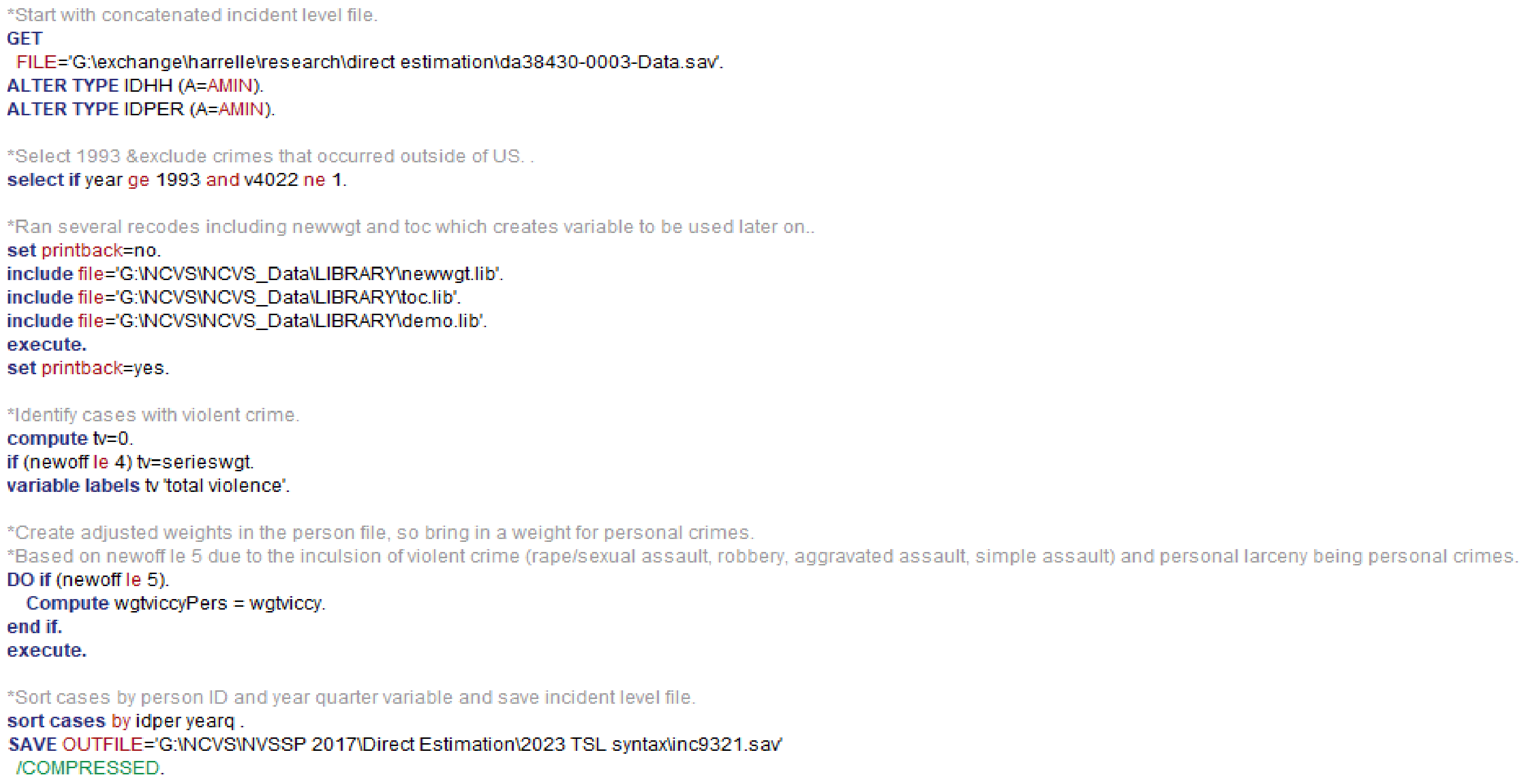
Population

Total

SE of Total

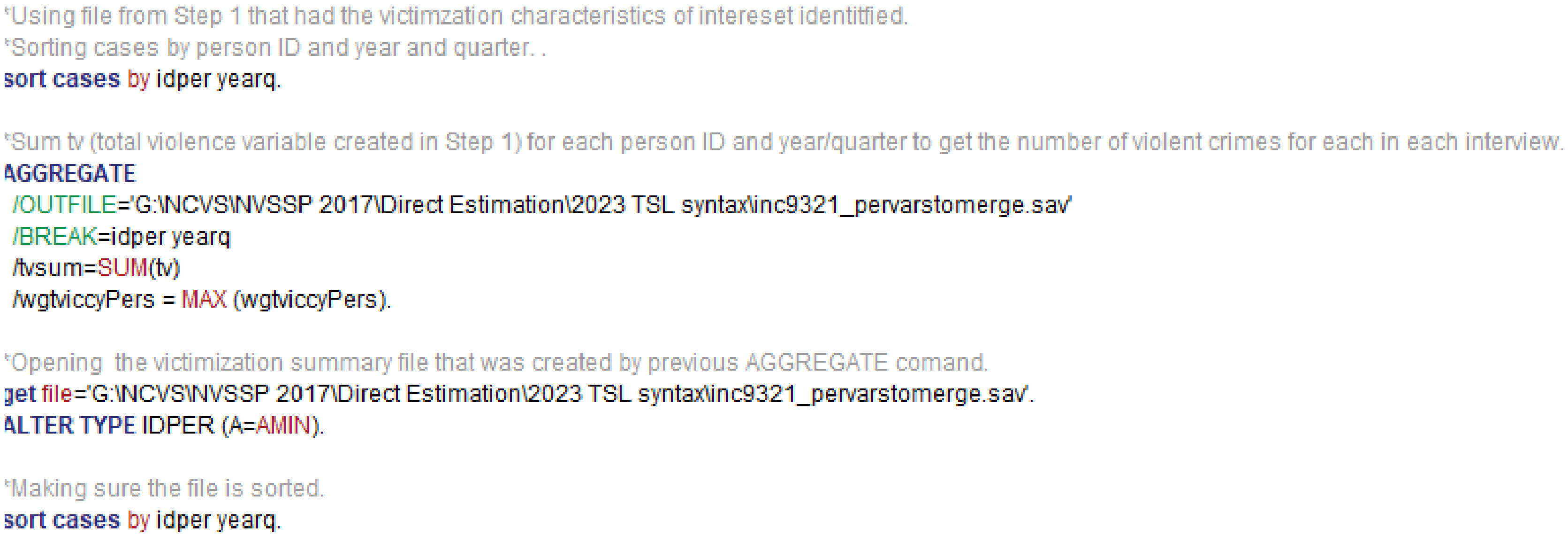
Using TSL to estimate variance in SPSS

Step 1: ID cases with characteristics of interest (violent crime)



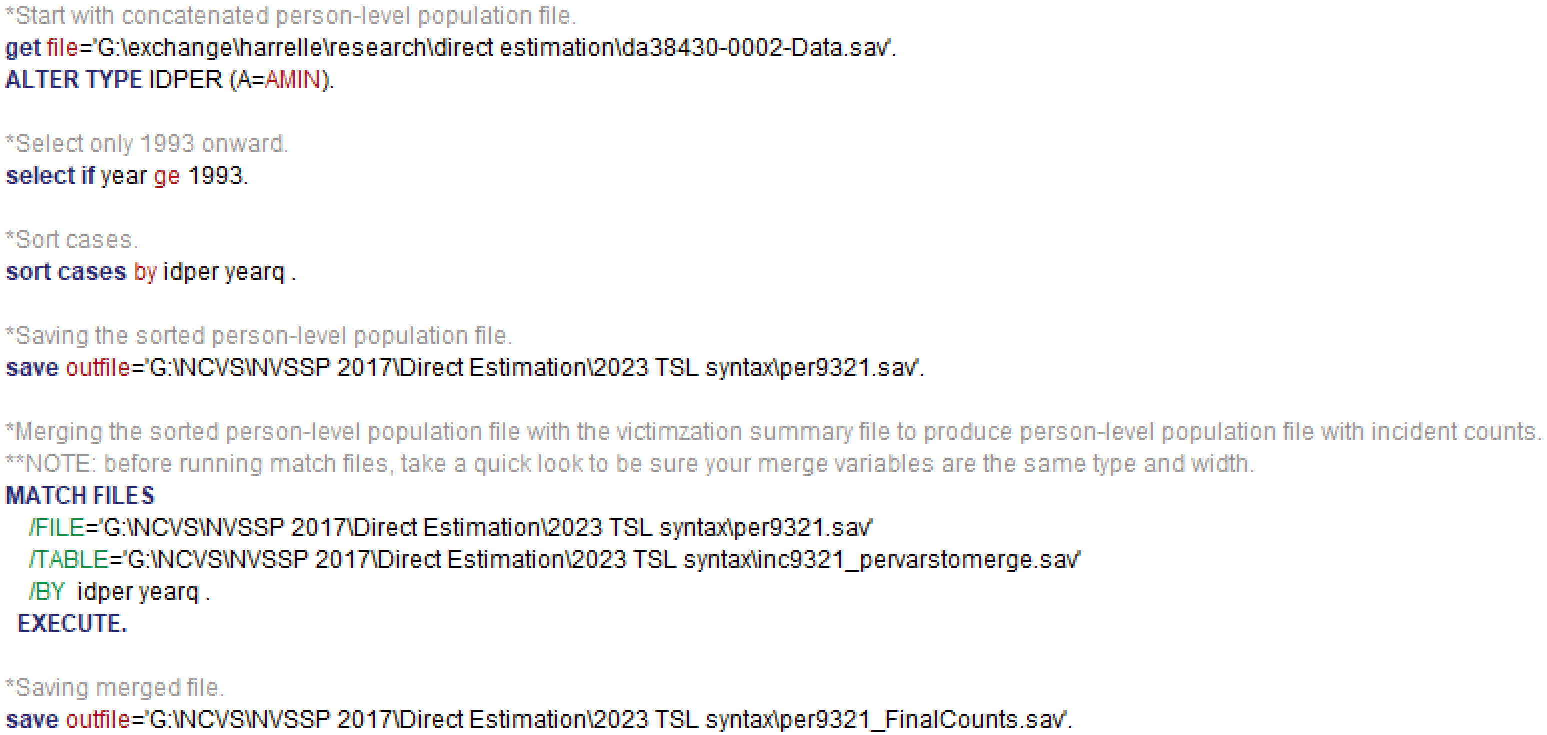
```
*Start with concatenated incident level file.
GET
 FILE='G:\exchange\harrelle\research\direct estimation\da38430-0003-Data.sav'.
ALTER TYPE IDHH (A=AMIN).
ALTER TYPE IDPER (A=AMIN).

*Select 1993 &exclude crimes that occurred outside of US. .
select if year ge 1993 and v4022 ne 1.

*Ran several recodes including newwgt and toc which creates variable to be used later on..
set printback=no.
include file='G:\NCVS\NCVS_Data\LIBRARY\newwgt.lib'.
include file='G:\NCVS\NCVS_Data\LIBRARY\toc.lib'.
include file='G:\NCVS\NCVS_Data\LIBRARY\demo.lib'.
execute.
set printback=yes.

*Identify cases with violent crime.
compute tv=0.
if (newoff le 4) tv=serieswgt.
variable labels tv 'total violence'.

*Create adjusted weights in the person file, so bring in a weight for personal crimes.
*Based on newoff le 5 due to the inculsion of violent crime (rape/sexual assault, robbery, aggravated assault, simple assault) and personal larceny being personal crimes.
DO if (newoff le 5).
   Compute wgtviccyPers = wgtviccy.
end if.
execute.

*Sort cases by person ID and year quarter variable and save incident level file.
sort cases by idper yearq .
SAVE OUTFILE='G:\NCVS\NVSSP 2017\Direct Estimation\2023 TSL syntax\inc9321.sav'
 /COMPRESSED.
```

Step 2: Creating victimization summary file



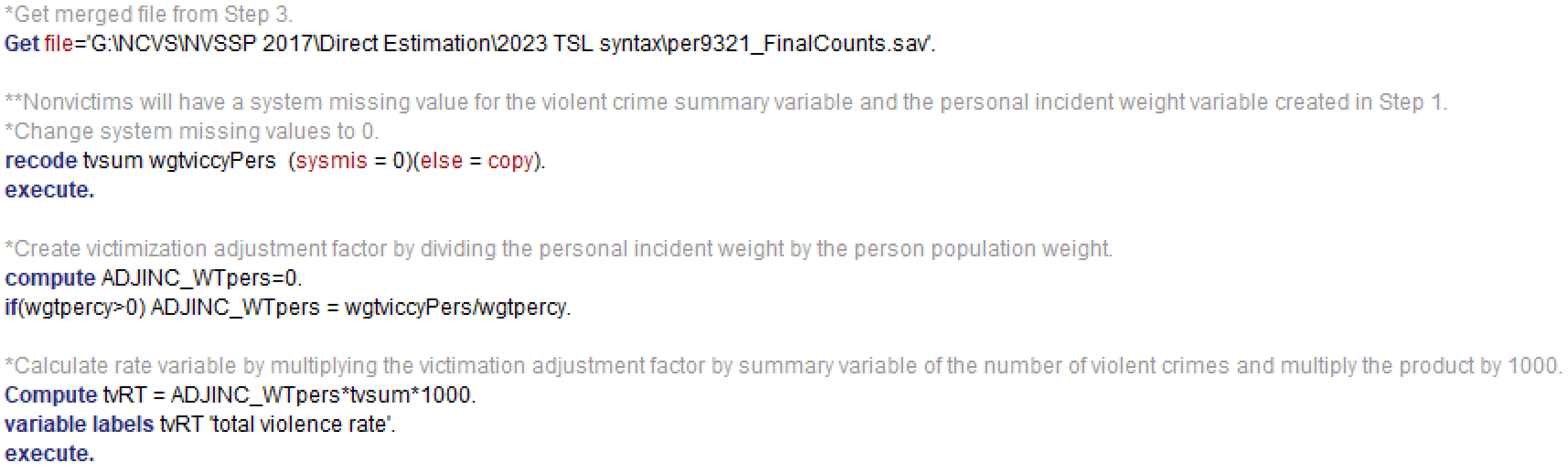
```
*Using file from Step 1 that had the victimzation characteristics of intereset identitfied.
*Sorting cases by person ID and year and quarter. .
sort cases by idper yearq.

*Sum tv (total violence variable created in Step 1) for each person ID and year/quarter to get the number of violent crimes for each in each interview.
AGGREGATE
 /OUTFILE='G:\NCVS\NVSSP 2017\Direct Estimation\2023 TSL syntax\inc9321_pervarstomerge.sav'
 /BREAK=idper yearq
 /tvsum=SUM(tv)
 /wgtviccyPers = MAX (wgtviccyPers).

*Opening  the victimization summary file that was created by previous AGGREGATE comand.
get file='G:\NCVS\NVSSP 2017\Direct Estimation\2023 TSL syntax\inc9321_pervarstomerge.sav'.
ALTER TYPE IDPER (A=AMIN).

*Making sure the file is sorted.
sort cases by idper yearq.
```

Step 3: Merge victimization summary file with person population file



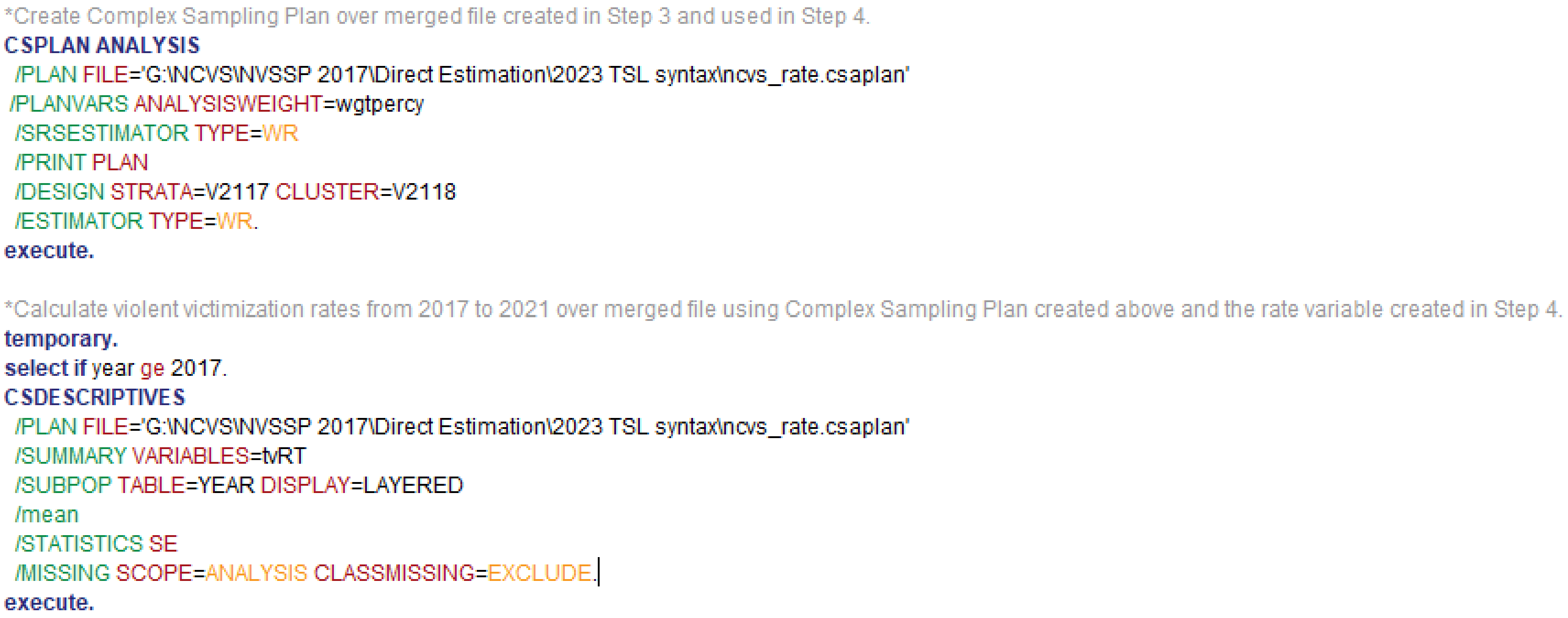
```
*Start with concatenated person-level population file.
get file='G:\exchange\harrelle\research\direct estimation\da38430-0002-Data.sav'.
ALTER TYPE IDPER (A=AMIN).

*Select only 1993 onward.
select if year ge 1993.

*Sort cases.
sort cases by idper yearq .

*Saving the sorted person-level population file.
save outfile='G:\NCVS\NVSSP 2017\Direct Estimation\2023 TSL syntax\per9321.sav'.

*Merging the sorted person-level population file with the victimzation summary file to produce person-level population file with incident counts.
**NOTE: before running match files, take a quick look to be sure your merge variables are the same type and width.
MATCH FILES
   /FILE='G:\NCVS\NVSSP 2017\Direct Estimation\2023 TSL syntax\per9321.sav'
   /TABLE='G:\NCVS\NVSSP 2017\Direct Estimation\2023 TSL syntax\inc9321_pervarstomerge.sav'
   /BY  idper yearq .
 EXECUTE.

*Saving merged file.
save outfile='G:\NCVS\NVSSP 2017\Direct Estimation\2023 TSL syntax\per9321_FinalCounts.sav'.
```

Step 4: Create victimization adjustment factor and the rate variable



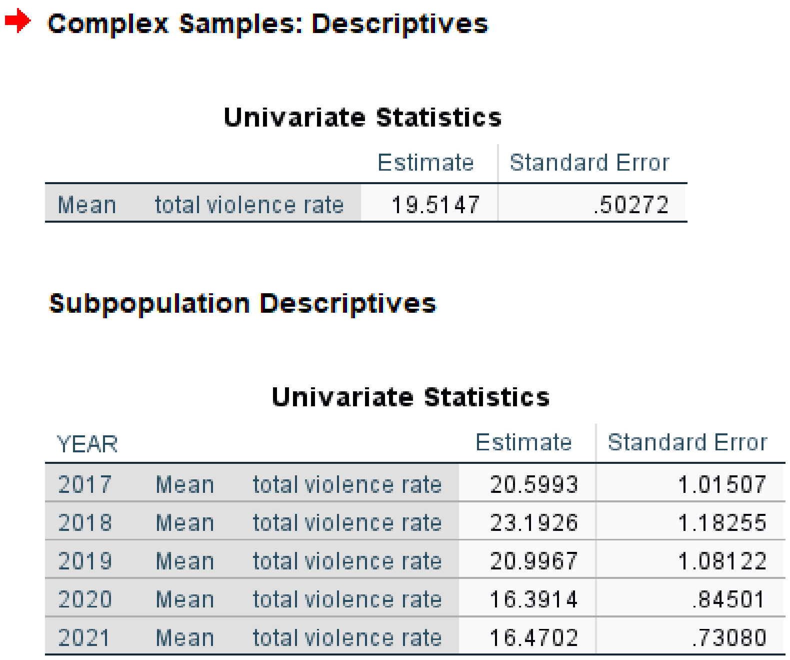
```
*Get merged file from Step 3.
Get file='G:\NCVS\NVSSP 2017\Direct Estimation\2023 TSL syntax\per9321_FinalCounts.sav'.

**Nonvictims will have a system missing value for the violent crime summary variable and the personal incident weight variable created in Step 1.
*Change system missing values to 0.
recode tvsum wgtviccyPers  (sysmis = 0)(else = copy).
execute.

*Create victimization adjustment factor by dividing the personal incident weight by the person population weight.
compute ADJINC_WTpers=0.
if(wgtpercy>0) ADJINC_WTpers = wgtviccyPers/wgtpercy.

*Calculate rate variable by multiplying the victimation adjustment factor by summary variable of the number of violent crimes and multiply the product by 1000.
Compute tvRT = ADJINC_WTpers*tvsum*1000.
variable labels tvRT 'total violence rate'.
execute.
```

Step 5: Generate rates and standard errors



```
*Create Complex Sampling Plan over merged file created in Step 3 and used in Step 4.
CSPLAN ANALYSIS
 /PLAN FILE='G:\NCVS\NVSSP 2017\Direct Estimation\2023 TSL syntax\ncvs_rate.csaplan'
 /PLANVARS ANALYSISWEIGHT=wgtpercy
 /SRSESTIMATOR TYPE=WR
 /PRINT PLAN
 /DESIGN STRATA=V2117 CLUSTER=V2118
 /ESTIMATOR TYPE=WR.
execute.

*Calculate violent victimization rates from 2017 to 2021 over merged file using Complex Sampling Plan created above and the rate variable created in Step 4.
temporary.
select if year ge 2017.
CSDESCRIPTIVES
 /PLAN FILE='G:\NCVS\NVSSP 2017\Direct Estimation\2023 TSL syntax\ncvs_rate.csaplan'
 /SUMMARY VARIABLES=tvRT
 /SUBPOP TABLE=YEAR DISPLAY=LAYERED
 /mean
 /STATISTICS SE
 /MISSING SCOPE=ANALYSIS CLASSMISSING=EXCLUDE.
execute.
```

Output



Complex Samples: Descriptives

Univariate Statistics

		Estimate	Standard Error
Mean	total violence rate	19.5147	.50272

Subpopulation Descriptives

Univariate Statistics

YEAR			Estimate	Standard Error
2017	Mean	total violence rate	20.5993	1.01507
2018	Mean	total violence rate	23.1926	1.18255
2019	Mean	total violence rate	20.9967	1.08122
2020	Mean	total violence rate	16.3914	.84501
2021	Mean	total violence rate	16.4702	.73080

Q&A Session

Please type your questions into **Q&A** selecting **All Panelists**

Alexandra (Lexy) Thompson
Statistician
Bureau of Justice Statistics
Alexandra.Thompson@usdoj.gov

Erika Harrell, PhD
Statistician
Bureau of Justice Statistics
Erika.Harrell@usdoj.gov

Marcus Berzofsky, DrPH
Senior Research Statistician
RTI International
Berzofsky@rti.org

Andrew Moore
Research Statistician
RTI International
amoore@rti.org

810 Seventh Street, NW, Washington, DC 20531 | Phone: +1 (202) 307-0765 | https://bjs.ojp.gov