DARYL FOX: Good afternoon, everyone, and welcome to today's webinar, “Direct Estimation Methods and the National Crime Victimization Survey,” hosted by the Bureau of Justice Statistics. At this time, it's my pleasure to introduce Alexandra Thompson, statistician with the Bureau of Justice Statistics, for some welcoming remarks and to begin the presentation. Alexandra?

ALEXANDRA THOMPSON: Thank you, Daryl, for that introduction to the webinar, and welcome, everyone, again. My name is Alexandra or Lexy Thompson. I'm a statistician in the Victimization Statistics Unit at the Bureau of Justice Statistics. My portfolio work currently focuses on school crime, hate crime, and victim service providers. I've also coauthored various BJS statistical reports, including “Criminal Victimization 2021” and “Hate Crime Victimization 2005 to 2019.” With me today, we also have Dr. Erika Harrell, who's also a statistician in the Victimization Statistics Unit at BJS. During her tenure at BJS, she's written on a variety of topics, including workplace violence, identity theft, and crimes against persons with disabilities. She was also part of the BJS team that created the N-DASH, the data analysis and visualization tool that generates estimates using data from the National Crime Victimization Survey.

We also have Dr. Marcus Berzofsky from RTI International, where he’s a senior research statistician. He has over 20 years of experience with design, implementing, and analyzing complex survey data. For the past 10 years, he has served as the co-principal investigator on the National Victimization Statistical Support Program, which he has helped BJS review and improve methodology used for the NCVS. Dr. Berzofsky has also authored several federal and peer-reviewed publications on NCVS methodology, including reports on producing direct variance estimates with NCVS data. We also have Andrew Moore available, who is a research statistician at RTI International for 14 years for his experience with survey research. His areas of interest also include implementation, weighting, data analysis, and SAS programing. For the past 10 years, he has served as a statistical task leader on the National Victimization Statistical Support Program, where he has contributed to numerous substantive and methodological tasks aimed at enhancing the NCVS. I'll now go over with the agenda for today's webinar. We'll start with learning about what variance estimation is and why it's important. Then we'll go over the differences between direct and indirect variance estimation, including the methods that are listed below on the slide.

We also have two demos on direct estimation, one in SAS and one in SPSS. And as Daryl mentioned, we'll have a Q&A at the end of the webinar. And I'll start us off today by talking about what variance estimation is and why it's important. Before I talk about the exact definition of variance estimation, I want to talk about the structure of the NCVS, because that's really related to how we approach variance estimation for the

data. The NCVS is a survey where we interview a sample of households that are representative for the entire country. This map gives somewhat of a visual representation of that. These points are not based on NCVS data. They are randomly generated, but we're not a census of crime across the country. We interview a sample of persons within a sample of households. And since it's not a complete count of crimes occurred across the country, we need the data from the households that we interview represent not only themselves, but also the households we're not able to interview through the NCVS. And we're able to generate national estimates for crime across the U.S., partially through weighting the data.

And so what is variance? Estimates based on a survey and based on a sample have some degree of sampling error. And sampling error can depend on a variety of factors, including the amount of variation between responses for similar households and also the size of a sample. So if you have a small sample for a particular group, your sampling error may be larger and for a larger sample, your sampling error may be smaller. And the variance which we're talking about today is a type of sampling error, and that measures the deviation between the estimate that we have from the survey and an average. This theoretical average comes from the idea of, if we did this survey with different samples over and over and over again, there would be an—sort of a true average between all these different estimates. So the variance is measuring the deviation from our current estimate compared to this theoretical average. And then the standard error, which you may have heard about before, is the square root of the variance. And standard errors can be used to calculate confidence intervals around an estimate you may have heard of, a 95% confidence interval.

This is a visual representation of a confidence interval. The screenshot comes from the NCVS dashboard or N-DASH. And so, the blue trend line represents the estimate or the rate of violent victimization per 1,000 persons age 12 or older. The brownish shaded area represents the confidence of the 95% confidence interval around the estimate. So that top brownish line in the white is the upper bound, and then the brownish line in the blue area, that is the lower bound of the confidence interval. So for example, in 2021, we estimated that the rate of violent victimization was 16.5, violent victimization per 1,000 persons age 12 or older. And then we estimated the standard error using GVF that was about 0.82. And using that standard error, we calculated the confidence interval that is displayed on screen.

And so, another way to interpret this is, if BJS use the same sampling method to select different samples and compute an estimate for each sample, it would expect the true rate of violent victimization to fall between 14.86 and 18.08, 95% of the time.

Now, why is this important, calculating the variance? It gives you partially an idea of how reliable your estimate is. An estimate with a smaller standard error, which is the square root of the variance, provides a more accurate or more reliable approximation of this true value than an estimate with a larger standard error. And so, larger standard errors may have less precision and less reliability. And then depending on the parameters of your project, an estimate with a very large standard error should usually be interpreted with caution and also may not be usable. Standard errors also help us determine whether two estimates are statistically different or not statistically different.

For example, you can have two estimates where one may be technically higher than the other, such as maybe one estimate of 16.5, and one estimate of 17, whether that's a rate or a count. And so, 17 is technically higher than 16.5. But when taking into account these standard errors, they may actually not be statistically different. And so, in a lot of BJS reports that you'll see or for every BJS report we conduct statistical tests to determine whether our estimated counts, percentages, or rates, whether they were statistically significant compared to maybe another estimate once the standard error is taken into account.

Also, with variance estimation, it's important to account for the NCVS' complex sample design. And if you don't account for the design of the survey, then you may underestimate your variance, which could lead to inaccurate confidence intervals or inaccurate statistical comparisons. And that's done through weighting, the weights that are present on the data file for the survey and also some other variables which Marcus will go into during his portion of the presentation. Speaking of Marcus, I'll turn it over to him to talk about types of variance estimation.

MARCUS BERZOFSKY: Thanks, Lexy. Good afternoon, everyone. I'm Marcus Berzofsky. I will talk about types of variance estimation and sort of compare the different—give advantages and disadvantages of the different types. And then I will also give the example on SAS.

So with that, so, for variance estimation, there are two broad types. There's what's called direct variance estimation and indirect variance estimation. Direct—and we'll define each in the forthcoming slide. But Direct Estimation often can come into two different flavors. One is called Taylor Series Linearization and the other are Replication Methods. And then there are several different types of Replication Methods. And I will talk about the one that the NCVS uses, in particular. On the indirect side, the Indirect Estimation, the main type of estimation is called a Generalized Variance Function. There are other types of Indirect Estimation. So for those that have heard of like small area estimation, that is a type of Indirect Estimation, but that's a very different type, a

very specific, you know, type of Indirect Estimation. But the main one for just a general dataset, like the NCVS, they produce estimates for any type of estimate. It's called the Generalized Variance Function. We'll talk about that. Coming up.

So, we will start with Indirect Variance Estimation. So how does indirect variance estimation compute a variance? Well, the generalized variance function, often abbreviated to be GVF, is produced through a nonlinear model. And I'll show you what we mean by a nonlinear model in a second. And so this nonlinear model is used to fit the variance estimate. So it's a linear model where the outcome are the actual estimates produced through direct variance estimation as your outcome for the model. And then it has this nonlinear function that is used to the model, what the variances are and produce the regression curve that then produces the variance estimates from the GVF.

Once you have that model, once someone else creates that model, you don't have to create that model. Someone else produces that model. Once you have that, then there will be some model parameters.

So just like in a model, there's—they're betas. And so those beta estimates are the model parameters that are given to you that you can then plug into a very relatively simple function that you can use in Excel or whatever. And then that will then generate with some totals that you can get out of the data. You can then generate your standard errors. So, in terms of needing anything special, no. So no special software is needed. In fact, you could use a SAS or an SPSS just to generate the—to summarize the data to get the necessary population counts or counts that you need to enter, to input into the formula. But really, even if you just have the data, and then it would be a little hard, but if you had a data in an Excel spreadsheet, you could just summarize it in there. So you don't really need special software per se. It can make your life a little easier, but you don't actually need anything special. And so that's one thing that's really nice about GVFs. And with that, only the weighted estimates in the model parameters are needed, as I just said, and everything in theory could be done in Excel. But, certainly having a software like a SAS or SPSS, just to summarize stuff will make your life a little easier.

Looking at the basics of it, so I wanted to just—I don't want to get into the nitty-gritty of these formulas. I know they're a little complicated looking here on the left side, but I just wanted to show you what is meant by a nonlinear function. And basically, a nonlinear function by definition just means it's not—it has things like a square or a square root, and it's just not going to produce a straight line. So, basically when you produce variances for any dataset, they're not going to come out in a simple straight regression line. That's simply a straight line that you could fit to that curve. It's going to be some sort of curve or some sort of different shape. And so you need a nonlinear function to

describe that pattern. And so in the NCVS, they're—it's relatively complex. And they're different patterns depending on the type of estimate. So, for a total, there's what's called a three-parameter model. So, there's three parameters. There's this A parameter, the B parameter, and the C parameter. And so when you're using the GVF, you would have an A parameter, B parameter, C parameter. You just basically enter those into the function along with your totals that you get from the data. And you would then—it would then give you the variance for that estimate. And then you could take the square root of that to get the standard error.

The function is different if you're producing a rate per 1,000 persons or households. It's a two-parameter model which is the B and the C parameter and a slightly different function. So basically you can get these values of T or R or N. Those come directly from the data. You then enter those. And then, you know, with the parameters that are given to you in a spreadsheet, you can then just enter them and get your variance. So, it's a really relatively straightforward thing to implement for a single estimate. If you're doing a lot of estimates, it can be a little cumbersome. But in general, for a single estimate, it's a relatively straightforward approach to calculating variances. But I do want to point out that it's an approximation. Because it's a model, it's not getting the exact variance for a certain outcome. But if the model—it's the model approximation of that variance estimate.

So, turning to then direct variants to contrast it, Direct Variance Estimation. So how does it compute variances? Well, it computes it directly from the data. So, it pulls in the data and it will directly compute the variance estimation based on the actual individual, case level data. It doesn't—and as such, it doesn't require outside information. Everything you need to calculate the variances are on or should be on the dataset that you're using for analysis. So, in terms of them what's special that's needed, you do need—but with that said, you do need to have certain variables on the dataset in order to calculate the variance of a complex survey design. And they're usually one of two types of variables, and we'll talk about that. And it depends what variables they are will depend on the type of direct variance estimation you're going to compute. And I will talk about that as we talk about Taylor Series or Replication in a second, because which one you need is varies. I'll talk about that in a second.

But what variables you need will depend on the type of direct variance estimation you need. And to do—and to do it, to get the proper scenarios, you do need a statistical software like SAS, SPSS, SUDAAN, or R. All of those will have complex survey packages that can account for the complex survey design when computing the variances. And these packages are important because regular SAS, base SAS or base SPSS, they will not calculate the correct standard error. Because there are weights and

populations, they will underestimate your standard errors. So, if you're not using a complex survey package in one of these software packages, you will get incorrect standard errors. So, you need to be very conscious of that. So, you could just have base SAS or base SPSS and run a function and get a variance but if you're using complex survey data, that variance will be wrong. It'll be smaller than it should be. And so you need to make sure you're using the complex survey functions or procedures when using these software packages. And we'll talk about the SAS example and SPSS later in the presentation.

So, the two types of direct variance estimation that I mentioned a couple of slides ago are Taylor Series Linearization and Replication without—I'm not going to get into the nitty-gritty of how those two approaches work, but I'll just simply say what is needed to operate them from a practical standpoint.

For Taylor Series, Taylor Series does rely on the survey design. So, you need to be able to specify the survey design. And with that, you need two things. You need the population weight, which is the main weight on the file, and then you need to be able to define the design parameters, which often are a stratification variable or a clustering variable, often called a PSU, primary sampling unit. These variables will be on the analysis file, and you need to specify them in your statistical procedure.

Replication. Replication, when it produces, it accounts for the design. And so you do not actually need the design parameters. You need the population weight. But instead of the design variables, you have a set of replicate weights. So replicate weights are basically—it's dropping some cases, creating a variance, and then doing it again and over and over again so that it captures the variation in the underlying sample data. And so one nice thing about replication is that you don't need to know or specify what the design was. You don't need to have that information. But you do have this series of replicate weights.

So, for example, the NCVS in, the last 20 years I think has sometimes been different. But there are a 160 replicate weights on the file. So, it's not a small number of additional variables that you need to specify. So there are different ways of doing replication. It's just the different ways of creating those replicate weights. Main ones out there are jackknife or delete a group to delete one jackknife and then balance repeated replication or BRR. And the NCVS uses that second one, balance repeated replication. And so it uses that approach to create the replicate weights. And that's important and you will need to specify that in the sampling procedure. And NCVS, in fact, uses a special version of BRR called Fay's BRR, so for Bob Fay.

So, in terms of some basics on Taylor Series Linearization in terms of how to use it, you do, as I mentioned, need to know the Design Variables. And there are two such variables on the NCVS. So, the PSEUDOSTRATA, which is if you're using the public use file, the V2117 and then the HALFSAMPLE variable which is V2118. And that's the clustering variable. And then there are the NCVS Weight Variables. There are three depending on which file you're using, the WGTPERCY, WGTHHCY for the household weight and then the incident weight, which is WGTVICCY times we—in most cases multiply that by the SERIESWGT. So, if it's a series incident, that will inflate the victimization weight.

As I'll talk about, you only can use one population weight at a time. So, we'll talk about what—how you not take the three files that are the NCVS and produce—and create a file that can be used for direct variance estimation. For both Taylor Series and Replication, you'll need to do some data manipulation to combine the incident file with the person or household file to get it to—in a way that can then run in a statistical software package. And I'll talk about that when I get to the example section.

BRR Basics. So, you have the same population weights there below. But above, instead of specifying the design, you have the replicate weights and there are replicate weights on the household file and replicate weights on the person file that you will need. One thing I'll note is that these replicate weights for those that are familiar with the public use files in the NCVS, these replicate weights are only available on the single year annual files. They're not—for those that are used to downloading the concatenated file that already has a lot of years set together, that will not have the BRR weights. So, if you want to use BRR, you need to download each individual file and then stack them together and use them that way.

The other thing I'll just note about for both direct variance estimation methods, so the NCVS every 10 years changes to a new set of primary sampling, a new set of clusters, new set of primary sampling units to take into account the new census. And so when that happens, the definitions of the design variables will change. And so that, you just need to know that and add an extra variable sometimes in your analysis, in your design specification. The replicate weights in the single year weights should be fine, but that's—it's probably more of an issue with Taylor Series. But I'll just note that they don't actually show that in an example. But I thought I would note that because that's an important detail.

So that's sort of the basics of each of these methods. And I thought I would just spend the next few slides sort of just doing some comparing and contrasting of the different methods. I know some of this text is small. I don't know how to get it bigger, but I want

to just start because there's no perfect method. There's no like, "Oh, it's one question out there, like, well, if you had to do one way, which way would it be?" And it really is your user preference, to be frank. But you should know the advantages and disadvantages of each. And so first, I'll just compare Indirect to Direct and then we'll look at, comparison of BRR to Taylor Series. But looking at Indirect to Direct, some advantages—and some of this is repeating things I've said. But looking at indirect you don't need to know how to use a complex survey package. So if you're not program-oriented, if that's not your thing, GVFs are really nice because they allow you to get the correct standard errors without having to learn a software package or even get a software package. So, that's nice. And it doesn't really require any file manipulation, which I just sort of mentioned that you have to do with both the direct estimations. You can leave the three files as they are, get the necessary counts that you need from the respective files for whatever estimate you're producing, and then plug them into the formula.

On the disadvantage side, they are less accurate. It's a model-based approach to producing estimates. So, it is an approximation rather than exact standard error for a specific estimate for a specific outcome. And it has to be calculated for each outcome individually. So, you're going through each one and calculating it one at a time. And if you're comparing years, there are additional parameters I didn't mention. There are correlations because the NCVS in particular, it's a rotating panel design. So, households are in the sample for three and a half years. And so each year, there's repetition in the households, so there's a correlation there in the samples. And so you need to account for that correlation.

And so there are additional parameters that the Census Bureau provides and BJS provides that if you're doing a year-to-year comparison, you want to make sure you take into account. Again, not anything you can't do in a spreadsheet, but just more things that you just need to be aware of and make sure you are taking into account.

Whereas on the Direct side, some advantages there, again, they're more accurate. They're specific using the data specifically for that outcome that you're estimating. You can estimate in a statistical procedure. You can list multiple outcomes simultaneously, get multiple outcomes, look at domain estimates simultaneously, do all that all in one fell swoop. So, estimates can be done a lot more quickly. And so—which is good. And you can do comparisons over time as well. I mean, sort of statistical complex design package will take into account that correlation that I just mentioned. You don't need to require or do anything special to do that. Some disadvantages are, though, if you are not— if programing is not your forte, you know, there is one access to those software packages. And then two, knowing how to use them properly which—so, that can

certainly a barrier to some people. So, the GVFs are nice because I think it doesn't require that barrier.

And just to note, just—these are quirks specific to the NCVS. There are some quirks to both methods. Neither method is perfect. So, when it comes to the NCVS data, unfortunately. So for example, 2016, in particular, was a weird year, that was a year where they were switching over from the census data and the design changed. They expanded the design to allow for state level estimation. So there are a lot of changes going on. And at that point, they had to modify the design variables.

The Taylor Series, that actually not an option for 2016. So, if your analysis includes 2016, and you want to do Taylor Series, you're not going to be able to. Whereas on the flip side, BRR, BRR can be done for 2016 and can be done for all years, but in SPSS, if you're not a super SPSS user, SPSS bought the complex package in SPSS will only do Taylor Series. To do BRR, you can do it, but you have to write your own macro. There's no default BRR option in SPSS. So it can be done, but it does require a macro that you write, to basically to program the BRR function yourself.

And so that—if you're not, again, an expert of SPSS function, that can be a barrier. But once you have that macro, once you have that function, then yes, you could do BRR in SPSS without any issue and you can do it for all years as well. And as I said earlier, when we'll talk about this, there is some file manipulation required because you can't separately analyze the incident file and the householder or person file. You have to combine the incident to those files, and then once the incidents are appended or merged onto the person or household file, then you can analyze—calculate direct variance estimates.

Now, just focusing on the two direct variance estimations now, Taylor Series here on the top and BRR on the bottom, there are some differences. Taylor Series is computationally faster, technically. In today's day and age, what does that mean? I mean, BRR— maybe it takes a couple of minutes longer depending on the memory of your computer. Computers today are all pretty good. So, that is a true statement. But in today's world, I don't think it's a major barrier with BRR. But the Taylor Series technically— will run faster. And Taylor Series is easier to run in SPSS, if you're an SPSS user. Taylor Series is definitely its default. It's the only complex estimation procedure—base procedure in the complex design. Like I said, you can write a macro to do BRR, but if you don't have that, Taylor Series is much easier to implement. And Taylor Series are easier to implement across decennial census if the number of replicates change. So, the last couple decennial censuses, the BRR number of replicates has been steady at 160.

So the last 20 years, it's been, what, it's been at 160. But there were earlier years in—if you go back all the way to the '90s where there were fewer, I think, than 160, and that's problematic, because you can't specify a different number of weights. And so that's problematic. But as long as the number of BRR weights are the same, it's not a problem, but if that changes, it could be problematic. So, I don't—and I can't predict the future of how the census will create the BRR weights going forward. But for a 10-year period, they're going to be fixed. So from the next, you know, starting 2016 to 2025, there's 160, and then presumably, in 2026, when it changes, from what I understand, the census will keep it at 160 and it'll be fine, but if they changed it, that could be problematic.

So disadvantages, though, to Taylor series, you do need to know the sample design, you need to be able to specify it in the procedure so you need to be aware of how to do that. Not necessarily challenging, but if you don't know anything about the design, that could be a little bit problematic. And then we'll talk about more of this problem manipulation. I mean, it's, you know, it's not a small thing, so you need to be aware of it. And then the 2016 issue, just, again, something to be aware of.

BRR is technically better for disclosure avoidance because you don't need to specify the design. You don't need to know anything about it. All that can be embedded in the replicate weight. So, for surveys that are concerned about disclosure avoidance, BRR is definitely a superior approach. Disadvantages, it can be computation a little slower, a little bit more challenging in SPSS if you don't have the macro, and it is not available on the concatenated file, so you have to download each individual year. And, again, if you're pooling years, again, you can't pool years that have more different number of replicates. Okay?

I'm not going to walk through this because this is just sheer repetition, but I did want to have a slide roll, just stay here for a minute, and let you look at it. But I did want to have a nice slide that just, sort of, just ran through everything there, sort of summarized the different features of each of the three methods on one sort of slide and you can sort of see—I mean, each one has—where it has checkboxes. And I wouldn't say, "Oh, TSL has more checkboxes, so it's the best." That's not—was not my intent when I was putting this together at all. It was just going through different features. But they each have their pros and their cons. And I think the key in—if my message—a key message that I'm trying to get across here is that you just need to be aware of them and know how you like to analyze data. And once you understand that, I think that will help lead you to the approach that's best suited towards you. Because I don't think any of these are bad ways to go. They're just different. And so it's really what works best for you, and

the way you like to analyze data, and in particular, analyze the NCVS. So I don't think any one of these is better, or I wouldn't discount any one of them per se. Okay?

Last, in this section, I just wanted to show you just so you can understand how the estimates are different. I have two years here, 2020 and 2021, and I have all the types of violent crime and then personal theft. And the numbers in the table here are what's called a relative standard error. Those who don't know what a relative standard error is, it's the ratio of the standard error that you get over the point estimate times 100, so it's a percentage. And so this is the measure of how—the quality of the precision of the estimate. So the problem with just looking at standard error is that it's relative—how good it is relative to the point estimate. So you might have a big standard error, but if it's a big estimate, it gets to be relatively small. So for relative standard errors, smaller is better. So smaller relative standard errors are—meaning it's a better—has better precision. And what I'll just note here is that there's no—so this is why I said there's no one that's necessarily better, it's not that there's one that's consistently higher.

So the GVF, for example, if you looked at the robbery—if you looked at the robbery row, you'll see in 2020, it had relative standard that's higher than BRR or TSL. But then you go to 2021, and the robbery relative standard error is lower. So going from year to year—so there is no—if you ask me, "Well, are GVFs always higher? Or are they always lower?" The answer is, it changes. So, because it's a different model each year that's being fit and how well that model fits for a different outcome, it's going to vary from year to year. BRR and TSL are going to be closer to each other because they're using direct variance estimations. While they're different, they're still similar. So, they're going to be much closer to each other in terms of their numbers, their relative standard errors. But as you can see, they're not exact. But the GVF, because it, sometimes you're above the regression line, sometimes you're below, or inconsistent in terms of being higher or lower. And it'll vary even within outcome from year to year. So just be aware of that.

Again, so just to my point, that I can't sit here and say one is superior to the others because, it varies from year to year and it's really—there are pros and cons to both approaches.

So, at this point, I think I'm going to pivot to showing an example. And I'm going to show an example using BRR in SAS, and then Erika is going to show an example of doing Taylor Series in SPSS. And so I will demonstrate this. I'm not going to run SAS live. I have a code here that I'll show you. But I've sort of broken down the process into four steps and I'll partially show step one, and then I will show you the code for the other three. But the first step is the most data manipulation, but you need to download all data that you need. So BRR, as I said, is not only concatenated file, so if you're using

multiple years for your analysis, you need to make sure you set all those data, download from ICPSR, all those different data years and set them together. And then you need to create the necessary derived variables that you need for them. So if you're analyzing certain outcome types, you need to make sure you define those and/or domains, like race, ethnicity, make sure you define that the way you want, and things like that. And that's step one.

Step two is sort of a critical step. You need to then summarize the incidents. On the incident file, you need to get summary counts of the number of incidents that occurred per person or per household, because you're going to merge those and back on to the person or household file. So the incident file is for those users on the NCVS, no, it's an incident level file. It's not a single record file. So if a person had multiple incidents during a year, they'll be listed multiple times. And so you need to summarize that count, and then merge that count back on to the person or household file where there's only one record per household or per person. Obviously, you need to merge those counts back onto it. And so that's what step two is, it's getting those summary counts, and then you got to merge them back on.

In step three for SAS, you need to then once you've got those counts, you need to divide them by the weight. In order to get the accurate count for either a total or a person, you need to create an additional function—a conditional value, which divides that count, that weighted count that I just told you, summary count of incidents by the total person weight or divided by the person weight and then multiplying it by 1,000 for weight. So you need to do that in a data step and then you can run it through the survey procedure with—for SAS is PROC SURVEYMEANS. So, SURVEYMEANS, so there is a PROC MEANS, that would get you the wrong values. PROC SURVEYMEANS, because there's that survey there, is accounting for the complex survey design. And that's the procedure you want to use in SAS. Okay?

So walking through that here, and I know some of this is small, so I apologize, but you'll see. So I just have some comments here like comment up here at the set function, you'll see this "A", and that corresponds over here to my comment annotation. So A just—this—I've already set all the years together, so I'm just noting that, but it's not a single file of incidence counts. You're going to need to set multiple years for BRR. Here in B, I'll just note, here's where the definition for things like rape and sexual assault, robbery, overall assault, simple assault, aggravated assault, et cetera, and then the aggregate violent crime, which is the maximum of RSA, robbery, yes, no, for those, and then personal theft. So these are just—you're defining your outcome variables, we're putting in exclusions. BJS typically excludes crimes that occurred outside of the United States. So, that's this extra line right here, so [INDISTINCT] So that's basically setting up and

defining your variables. So that's the part of step— I didn't show you all step one with all the content—setting and stuff but defining some of your outcome variables there.

Step two, the part of step two I'll show you is where you were summarizing. So here we are, just using a regular PROC MEANS because we are just getting a regular count, we're not calculating a variance here or worrying about it. We are just summarizing down the variables of interest. We're using a weight statement here to get the series weights, which again, was the function—the product of WGTVICCY and the serieswgt, so that's what this series right here is. And we're outputting that and we're out, you know, so we're summarizing the counts, the weight accounts for all those outcome variables that we defined in the prior slide.

Step three, I know it's just gotten smaller so—but step three is then the process of creating the variables you need for analysis and I'm focusing here just on violent victimization. We use arrays. You don't have to use arrays in SAS. So if they're a little complicated for you, you don't have to do one. But basically, this is just looping through the different outcome types and it's dividing it by the person—you're dividing by the person weights and multiplying by a thousand, so its—so this VIOLENT2 will be used for rates and the one—and taking the count, the weighted count that you had divided by the population count without multiplying by 1,000, that's the variable VIOLENT3. That's what we'll use when we want to calculate the totals and PROC SURVEYMEANS, so that's sort of the key takeaway here. For each of these outcomes, we're dividing by their weight to then either—so—and multiply them by 1,000 if—for rates to have it be in the form—the structure it needs to be for calculation of weights or calculation of the estimates itself. Okay.

So now 4A is the rates. This is the function, this is the PROC SURVEYMEANS, this is the main function. This is how you would set it up if you're running, calculating rates. So first, the variance method. So, you don't need to know anything about the design, but you do need to say what variance method you're using. And we're using BRR fay, and then, you need to put this in parentheses ‘(fay)’. That's not a comment. That's how you would write it in SAS. Fay’s method, it divides the weights by two. I mean, that's a little bit in the weeds. But the point is if you just said BRR, which would run, that would get you the incorrect variances. So with the NCVS, it's really critical that you do put that in parentheticals next to it. That's what is necessary for the NCVS. The outcomes that we'd get here, the mean here under that mean and sumwgt, we're getting the mean which will, in this case, will get us a weight—a rate and then the sumwgt is the population total. Those are outcomes that we're going to calculate. The variable VIOLENT2, so that was when we calculated for rates, had to multiply by a 1,000. We could list other variables here, so if you wanted to get the rates for RSA, we have that

for rape and sexual assault, we have that RSA variable. So we could get the detailed ones here. We could just list them all. So you don't need multiple—you could just, in a row, put all those variables in this var statement so there's no need to—so you can—but as I said, nice thing about this, you can calculate all this simultaneously. We are going to then, by domain, get our estimates by year. So our domain is year. If you wanted to look at other domains, so you could look at—if you had calculated like age categories like say you have five age—got like an age cat variable, you could put year, star, age cat in the domain, and that would calculate by year and age category. That would calculate them. Or if you didn't—if you wanted to pool your years and do it by age, you would just say age cat. So you could put all that in your multiple domains and calculate that.

To get an output data set, PROC MEANS, SURVEYMEANS, doesn't automatically create an output data set. You do need to use the output delivery system ODS to get that, so you need to line ODS output and then we create the dataset, in this case we call it estimate or EST. So you do need that ODS statement to get an output datasets. There will be output that's populating your output window, but if you want to get an actual dataset, and then, use it later, you need that ODS statement. Weight—so again, in SAS, you—or any statistical procedure, you only can enter one weight. So we move over the weighted incidents, so we don't need the incident file anymore for purposes of direct variance estimation. And we can just specify, in this case, it was [INDISTINCT] and violent crime, we can just focus on the person weight and use it. And then the replicate weights, and then we just need to specify the replicate weights, for the person one through 160. And we don't need to list them all, so—if there's—it doesn't really matter if there's 10 or 300, which obviously in 300, too. So you can just put a dash and then list them, so you don't have to, you know, so you don't have to tediously type every single one of them. You can just put the range.

As long as they all have the same basic root. And the only difference is the number at the end. So the weight name is the same. The only difference is the number at the end, one to 160, then you can just put a dash and list them all.

So this is the SURVEYMEANS function for a rate. And then looking at the outcome, you'll get outcome that looks like this. So we [INDISTINCT] in 20—I've—when I created my dataset, I [INDISTINCT] years for 2016 through '21. So as I noted, I put 2016 on purpose because I want to just remind you that BRR can create proper estimates for 2016. Taylor Series, we would not have been able to do that, but for BRR, we can. And so we get our two outcome measures, we get the sum of the weights, that was that sumwgt function. We get the mean, this is the rate. So 19.66 here in 2016, it's per 1,000 persons. And then we get the standard error of the rate. So the standard error of the

19.66 is 0.897 and so that's the standard error. And if you want to get a confidence interval, you can multiply this number by 1.96 to get the half-width and then plus or minus that number. And so that's the output that you would get. There are some other parts of the output that you can, for the most part, ignore them. I mean, they help you know that things ran properly, but they're not necessary in terms of the outputs itself. Now, quickly, I'll just say in for 4b here, this is just totals. The only difference in this slide is in the bar statement. It says VIOLENT3 instead of VIOLENT2. But otherwise, the setup is exactly the same and you can see them. The difference here is that your sum is now—is now—for VIOLENT3, it's now the total. So 535, you know, million—wait. Am I doing that right? No. There's a five—yeah, anyway, all right. I don't see the comments. It's hard for me sometimes, but, yeah, this is the total number of violent crimes in 2016 and this is the standard error for that. You can see that. And these numbers, I checked all these numbers will match what's in the BJS crime victimization bulletins, but there is your outcome. So with that, I will—I will pass it to Erika so that she can give you an equivalent example of Taylor Series and SPSS.

ERIKA HARRELL: Okay. Thanks, Marcus. What I'm going to do here is run through some SPSS syntax. that will be using Taylor Series Linearization to generate rates and standard errors. And with this example, I'm going to be showing how to generate rates of violent victimization per year from 2017 forward. As Marcus said, you can't do it for 2016, so I'll just be looking at 2017 forward. It's a few steps. Majority of this code is just set up for the file. Once you get through setting it up, running it and actually getting the estimates is actually very—is kind of short, a very short set of code.

Right now, first you have to start—we start with the concatenated incident level file. This is a full file. You can use a shortened version selecting certain years, but you have to be very careful that the same number of years you select here will be the same number of years later on in the code. First, we start with the concatenated incident level file. We exclude crimes that occurred outside of the U.S., select year greater than or equal to 1993. These are recodes that we use within our own unit that create a lot of variables that we use on a regular basis. Newwgt creates, like, incident level variables. TOC, that's our type of crime recode. And demo is like a demographic—that has all of our basic demographics that we have. After that, we have to identify cases with violent crime and newoff, which was created in our TOC recode. Le 4,—those are all of our violent crime values one through four. So we get those cases and we get the series weight that value for the case and identify in the incident level file. We also create a weight, an adjusted weight, actually based on the new—our newoff—which is—our newoff variable, which comes from the TOC recode. This is based on all of the person level crimes. The violent victimization is a person level crime. We need a victimization weight for person level crimes, which includes the violent crimes and personal larceny.

So newoff le 5 gets all of those violent crimes and personal larceny and getting that incident weight—that victim—I'm sorry, the victimization weight for those cases. And we sort the cases by person ID in year quarter. This is very important. Later on, you'll see that that is necessary to have these cases sorted for just every data file in this code.

Now, we also have to create a victimization summary file from that incident file. So using that same incident file that was created and say we create a—we add—we do an aggregation and get the number of violent crimes, that total violence variable, the number of violent crimes for each person ID and year quarter. We have to have that for each in order to get the total number of crimes that a person has had during the year and during the quarter. We just have to have that and we also have to sum up the victimization weight to make sure that we have the correct weights when we get ready to merge and also get ready to analyze the data. This code actually saves the victimization summary file so we have to get the file and actually open it. We open it. There's an alter type here that has to do with our merging that's going to take place in the next step. The variables that are used to merge have to be the same type. So you have an alter type command here for the person level Identification variable.

Now, we're also here, we're sorting cases. Again, we have to keep sorting cases on—or else the merge that occurs in step three will not happen. You will run into an error. Now, in step three to merge the victimization summary file with the person level population file, you have to do this because SPSS can only process a single data file at one time. Therefore, you have to merge the incident level file with the person level population file into a single file in order to run the Taylor Series Linearization. So this is our concatenated person level population files less than the same year. We're still sorting cases, saving that sort of person level file. And down here with the match files, it's pretty important that you do the files in this order in SPSS or else you will run into an error and there won't be a merged file. The file statement is for the population file for the person level population file and the table statement that has to equal the victimization summary file. If you get those mixed up, you will have an error. Also, the idper and yearq variables, they have to be there. These two data files have to be sorted by those two variables or else the merge will not work. You will run into an error. Saving a merge file, now moving to step four, we create what's called the victimization adjustment factor and the actual rate variable. So getting that first—getting that merged file from step three, you recode that violent victimization variable and the weighting variable, because when we merge victimization summary file with the person level file, there will be non-victims in that merged file. And SPSS, when they do a merge and there's no data for a case on a variable, if it automatically makes "system missing," this code actually changes the "system missing" to "0" and keeps everything else.

There's a victimization adjustment factor. It divides the person incident weight that was created back in step one and divides by the WGTPERCY, which is the person population weight that's on the person level population file. When you did the merge, it carried the population weight over, so it's on the file. Now we have to calculate a rate variable. We have to use the victimization adjustment factor, multiply it by tvsum which is the violent victimization count and multiply it by 1,000 because that's how we present our rates in our BJS report. That was the total violent victimization rate.

And generating rate is our last step. Like Marcus stated earlier, you can't just use SPSS, regular SPSS and get these rates via Taylor Series Linearization and get standard errors. You have to have a complex sample. You have to have a way of acknowledging the complex sampling design of the NCVS. So SPSS does have a complex sampling package, which is a set of commands that will allow us to take into account the NCVS sampling structure, sample design, and make it so that we can accurately reflect how the sample in the NCVS is structured.

So in order to do that, first of all, SPSS has to have the complex sampling package installed with your version of SPSS that you have. If you don't, then these commands, these two—the CS ANALYSIS and CS DESCRIPTIVES, this code will not work. First, you have to create what's called a complex sampling plan. This is where you tell SPSS how the sample is created. And it uses a variable—first of all, it has to be run over the data file that you want that gets the rate from, which for us will be that combined incident person level population file with all of the rate variables and victimization adjustment factor and all that. That's what it has to be run over in order to be created. It can't be created without a dataset.

The analysis weight would be since we are looking at a personal victimization, which is violent crime, we need the personal population weight, which is just WGTPERCY. We have two variables here, the PSUEDOSTRATA and the HALFSAMPLE variable, V2117 and V2118, which has to be used in order to create the sampling plan. And next, finally, is to actually generate the actual rate. What I'm doing with this code is generating rates of violent victimization from 2017 to 2021 over that merged file that we created earlier. Using the complex sampling plan that was done above this plan right here, created by CSPLAN ANALYSIS, and also using the rate variable that was done above.

Here, I'm selecting—since I used the full concatenated file, I am selecting year to be 2017 going forward. This plan file—this is the plan and it has to be here. This is the plan that was created earlier. This tvRT, this is the rate variable that was created earlier. You have to have a subpop table if you're doing annual rates with the concatenated file because if not, it will just give you estimates for the entire table, which is like an average

of all—the rate would be the average of everything, of all years that are in the dataset. So here we have to have a year layering it so that we can get annual rates going from 2017 to 2018. And with statistics, we have SE, which stands for Standard Error.

And now I'm going to run this to show you the actual standard errors. And it will take a minute because we're running it over the full concatenated file. The person level population file is the largest file that we—the personal level population, the concatenated file is the largest file that we have, so it takes a while to run it. So if you want to just run it for a year or so, you can actually cut down the incident level and the person level file to whatever year you're working with, except for 2016, and merge them together and it should be able to take a shorter time to run.

Now, this SPSS output, it shows that complex samples, the descriptive output. Now this, very statistic, this is what I was talking about earlier about creating a mean across all the data. This is the rate for all of the years that I was—I had selected from 2017 to 2021, it's overall rate. But here are the annual rates. One for each year, you have the rate, the estimate. These are the rates and the standard error. And if you notice, these rates are the same as what you would find in here.

This is table one from the 2021 "Criminal Victimization Report." The rate that you would find there in SPSS are the rates that you will find in this first row of the table, the 20.6, 23.2, 20.0, 16.4, and the 16.5, the same rates. And I will stop sharing. And I think at this point we are ready for our Q&A section.

ALEXANDRA THOMPSON: Thank you, Erika and Marcus, both for those demonstrations. So we can go to Q&A. I don't see any currently. Let me just check the chat. I don't think the chat—any questions in the chat. I don't see any questions in the Q&A so far. I'll give maybe people a moment to write some in.

I know it's a lot of information today, so definitely—we'll display our emails before we log off in case you have any questions that you think of after the presentation. And one of the things I wanted to bring up was when Marcus was talking about the GVF, the Census Bureau does create GVFs, specifically for the NCVS. And GVFs for each year are available in the code book which is available through the National Archive of Criminal Justice Data, NACJD. So if you go to NACJD to download the public use file, that should include the code book and within the code book, there's a Source of Accuracy statement from the Census Bureau, which contains the GVF parameters if that's the method that you wanted to use.

And then, with anything related to the NCVS when you're generating your own estimates, if you're wondering whether you're doing something correctly in your own calculations, some of the easiest ways to check are just comparing against published reports, kind of how Erika showed during her section, you can look at the victimization bulletin and check the estimates are calculated correctly and check the standard errors. In that particular bulletin, we did use BRR rather than TSL but maybe sometimes different publications use GVF or BRR depending on what was available at that time. So I wanted to throw that out there for anyone who might be thinking of projects that they're working on.

MARCUS BERZOFSKY: And one question has now come in.

ALEXANDRA THOMPSON: Do you see it, Marcus? I don't see anything in the Q&A panel.

MARCUS BERZOFSKY: So the question I saw—I'm seeing. “What if one has a regression model, which procedure should one use?” So the GVFs are only for descriptive statistics, so there aren't GVFs for regression models. So in that case, you would have to use one of the direct variance estimation methods. So you can use those. You can use either Taylor Series or BRR for a regression model. PROC SURVEYMEANS would not be your function. I think in SAS, it's proc survey reg and if you have a dichotomous outcome I think it's like proc survey logistic. So there are survey procedures for regressions in SAS. I know. I imagine too in SPSS. But you have to use one of the direct variance estimation procedures, either one will be allowable.

And so the second question I see here is, “Can R also be used?” Yes. We just gave examples in SAS and SPSS that R has complex survey functions as well, procedures and functions and you most definitely can use R as well. Yes. SUDAAN is another software package out there that I know will work. Stata, if people are Stata people, any of those software packages that can account for complex survey design. As long as your—the function—the procedure you're using is a complex survey procedure and R Stata, SUDAAN, Stata all have those. So you can use any of those. We're just limited in time what we can show examples. There are actually examples that BJS has published, the user's guide. There are user's guides, the direct variance estimation that are out there and you can get it through Google—the Google search. There are appendices and I know we did SUDAAN, SAS, SPSS and others. So there are examples of it. So, there are example codes out there that BJS have published.

ALEXANDRA THOMPSON: I think it's under the National Crime Victimization Survey Data Collection page. I think there's a document subsection and I think those are under if I remember correctly.

MARCUS BERZOFSKY: And that document will also talk through how to merge the files that you need do. So it'll walk through that and then there are appendices for each of the different software packages with examples how to do it.

ALEXANDRA THOMPSON: I don't see any other Q&A questions on my end. Do any other presenters, have one that I didn't feel free to speak up. If that's all the questions, if you think of anything after this presentation that you didn't think of now, all four of our emails are displayed on the screen. We're happy to answer any questions that we can. And thank you again from both BJS and RTI for attending this webinar. We hope it was helpful in that you're now thinking of all the great ways you can use the NCVS data to learn more about crime in the U.S. So thank you again. And I think our emails are also—Tammy posted our emails in the chat as well and I hope everyone has a wonderful day. Thank you again to the panelists and everyone for attending.