Introduction to Statistics
January 14, 2016

Stan Orchowsky: Good afternoon, everyone. My name is Stan Orchowsky, and I am JRSA's director of research. Welcome to the first in our Training and Technical Assistance Webinar Series on Statistical Analysis for Criminal Justice Research.

Stan Orchowsky: I want to tell you just a little bit briefly about JRSA. Justice Research and Statistics Association is a national non-profit membership association. Our core members are estate statistical analysis centers that collect and analyze criminal justice data across the country. I invite you to visit www.jrsa.org to learn more about the organization and to become a member.

Stan Orchowsky: Before we go any further, I want to thank our partners at the Bureau Justice Statistics for helping to make this webinar possible. I'm going to be starting out the webinar today, and then I'm going to be turning things over to Erin Farley, who is our research associate at JRSA.

Stan Orchowsky: Let me just cover a few logistical items quickly. We'll be recording today's session for future playback. The link will be posted on the website and emailed to attendees. Today's webinar is being audio cast via, both, the speakers on your computer and teleconference. If you have speakers on your computer or headphones and are not a presenter, we recommend listening to the webinar using your computer speakers or headphones. To access the audio conference, select audio from the top menu bar and then select audio conference. Once the audio conference window appears, you can view the teleconference, call in information, or join the audio conference via your computer. If you have questions for the presenter or would like to communicate with JRSA staff, we encourage you to submit them using the chat feature on the right side of your screen. Please select host from the drop down menu next to the text box.

Stan Orchowsky: The session is scheduled for one hour. I would not be surprised if we went over an hour. If you need to drop off at 3:00 Eastern, please feel free. If you can stay with us and we do go over, that would be great. If you're having technical difficulties or get disconnected during the session, you can reconnect to the session using the same link that you used to join, initially. You can also call WebEx tech support at 1(866) 229-3239.

Stan Orchowsky: At the end of today's webinar, we're gonna ask you to complete a short survey. The information you provide will help us plan future webinars, so we appreciate if you would take a few minutes and fill that out.

Stan Orchowsky: Today's webinar is titled Introduction and Overview of Descriptive Statistics. This is a first in what we've designed to be a series of webinars to provide some basic information to folks out there who, perhaps you took a class in statistics and grad school, and it's been a whole, probably not as long as it's been for me, but a while, and you need a refresher. Perhaps some of these concepts will be new to you. You're not a researcher, but you need to look at research, assess research, and so this series is designed for those kinds of folks. This will provide you with a very basic overview, today's session, of descriptive statistics, which is
going to be our starter. At the end of today's webinar, we're gonna give you a little preview of some of the upcoming webinars that will fill out the rest of this series.

Stan Orchowsky: This is going to be very introductory. Obviously, any of the information that you see on any of the slides today, we could spend a whole hour talking about any one of those things. We want to make sure that you understand these basic concepts before we move on to more advanced issues, and we want to make sure that we give you some practical examples of how to use this information. All right, we've covered all that. We've covered that.

Stan Orchowsky: What we want to do today is talk about some key concepts and descriptive statistics, and we want to demonstrate how to run descriptive statistics using SPSS, which is the Statistical Package for the Social Sciences, which you may be familiar with. This is commonly used by researchers and gives you the capacity to do an awful lot of a statistical analysis, but we also want to show you how you can do this in Excel, which is a little bit easier, more accessible, and more likely to be sitting on your desktop computer.

Stan Orchowsky: Let's begin here with one of many possible conceptual models of research and the research process. We begin here at the top with an area of interest with an idea or with a theory. Perhaps for some of you, certainly for our statistical analysis centers, this may also be a policy question. It may be a program that needs to be evaluated and so forth.

Stan Orchowsky: From that, we have a branch down to the choice of a research method. Basically, it's the design, the design of your study that's going to determine how you're going to address this, let's just call it, a policy question. You can see there, you have a number of possible methods that you're going to choose from, which, in part, are going to end up determining what your data look like, what the observations are, and what do your data look like, in turn, going to influence the analysis that you do.

Stan Orchowsky: The parallel to that design process, over on the left, you can see that we have a process related to specifically observations and measurement consisting of two steps, one, conceptualization, and one, operationalization, so talking about the how we're going to take this policy question and conceptualize it in terms of, what are we going to study? For any particular issue, there are going to be a number of ways to look at this, and so we want to get our key core concepts down. If we're talking about recidivism, for example, what do we mean by that? Or a better example, if we're talking about rehabilitation, what do we mean by that? How do we conceptualize that? Or if we talk about risk, how do we conceptualize that? These are concepts that we often talk about, obviously, in criminal justice.

Stan Orchowsky: Once we have that, we need to operationalize the concepts. We need to understand the operations or the specific variables that we're going to be
measuring, the specific way that we're going to be defining those things. Recidivism, for example, we may define in any one of a dozen different ways. How are we going to operationalize that? Those two things in parallel, the conceptualization, operationalization, and the choice of research method, as I said, will determine the observations, the data that we collect for analysis and interpretation, and how we analyze those data, as well as the application, how we report the results and how we assess their implications.

Stan Orchowsky: Then, of course, over there on to the right, we have the whole issue of sampling, and who we're going to study, and how we're going to study that. That will be a subject for another day.

Stan Orchowsky: Let's talk a little bit about two different kinds of statistics, descriptive and inferential. The descriptive statistics, these are the things, this is what we're going to cover today. These are ways of summarizing, organizing, making sense of a set of observations, describing the characteristics of a sample. When we talk about, for example, looking at criminal history records and producing summaries of what we find in there, that would be an example of descriptive statistics. We're looking at who is sitting in our jails or prisons in Maryland, would be an example of descriptive statistics.

Stan Orchowsky: Inferential statistics allow us to take measurements from a sample and to infer, or use the information drawn from the sample, to make estimates about the characteristics of the larger population from which the sample was drawn. For example, we might do a victimization survey. We can't, obviously, contact every single victim in a state, but we can certainly do a random sample of people and then draw conclusions from that random sample about the entire group of people who were victimized. Or, perhaps we're trying to evaluate the impacts of a program, and so we have essentially two samples from two separate populations, one of which receives the program, the so-called treatment group, and one of which doesn't receive the program, the so-called control group. These will obviously have implications, the type of analysis we're doing of whether it's descriptive or inferential will have implications for how we proceed in terms of analysis. As I said, we're gonna focus on the descriptive in today's webinar.

Stan Orchowsky: Oh, two data types. When we talk about operationalization, we can obviously either have quantitative data or qualitative data. I'm sure you're all familiar with these concepts. Quantitative data, basically numbers, factors that ... For example, we can create scales and quantify various phenomenon measures if we're interested in looking for example at the probation or parole officer's decision to revoke. We can identify what we think those important factors are. We can create a scale that would represent those factors, and then we can give that scale to a sample of probation officers across our state, and calculate their scores, and develop a quantitative measure of the factors that affect their decision.
Stan Orchowsky: Qualitative data, on the other hand, is non-numbered data that comes from talking to people or looking at other sources. For example, if we’re looking at that same question of, what affects probation officers’ revocation decisions? We can convene a group of probation officers in a focus group, for example, sit down with them and ask, "What is it that affects your decision regarding revocation?" Then through a series of guided questions, we can collect qualitative data that would also get at the same question.

Stan Orchowsky: Another way we can look at types of data is to think about the different types of variables that we talk about, a variable being ... Well, the definitions they are, I think most of you are probably familiar with variables. So, we can talk about, at the top there, you see the distinction between latent and manifest, so variables that are not directly observable and can be measured indirectly versus variables that can be observed. Here, you might think about things like, as latent variables, things like feelings or things like traits, such as self-esteem, for example, whereas at the manifest level, you might think about specific behaviors that we could look at or measure.

Stan Orchowsky: If we look at another way of characterizing variables below, the distinction between reliability and validity. Some of you may be familiar with these concepts. Looking first at reliability, which refers to the consistency or repeatability of a particular measure or concept. The easiest way to think about these is to use the example that you may remember from stats 101, which is the scale example. For example, if you have a scale that’s designed to measure your weight, and every time you step on the scale, you get a different weight, then that’s not a reliable measurement tool because the scale will vary every single time that you step on it, and so you won’t get the same result. If the scale is reliable, then every time you step on it, if you step on the scale and it measures, oh, let’s just, for laughs, say 138 pounds, and then you step off the scale, and step on it again, it should say 138 pounds if it’s reliable.

Stan Orchowsky: Validity refers to the degree to which the variable accurately measures the underlying concept or construct that it’s designed to measure. For example, using the same scale example, we might have a reliable scale in that every time we step on it, it reads the same way, but it reads consistently off. Instead of 138, it constantly reads 128. It’s consistent. It’s reliable, but it’s not valid because it’s not giving us a true picture of what we’re trying to measure, which in this case is weight. These, obviously, become important factors in scale construction when you’re trying to create a scale or use a scale that has been created. Again, we just want to introduce. These concepts are beyond the scope of what we’re gonna talk about today.

Stan Orchowsky: Then finally, a concept that hopefully you’re all familiar with, variables can be defined by the roles they play in a hypothesis testing or where there’s a causal relationship between two variables implied. You should all know this distinction, obviously. A dependent variable being usually ... We think of in terms of outcomes, what we hope to change, or behavior, or attitude, or something, so
what we want to measure, and then independent variables and the classic sort of experimental design way are the variables that are under the control of the experimenter are the variables that we manipulate in order to observe the effect on the dependent variable. For most of us, who are not necessarily doing randomized control trials or experiments, you can think of the independent variables as being sort of factors related to the dependent variable or predictors, if you will, of the dependent variable.

Stan Orchowsky: All right, a little bit about collecting quantitative data. There are a number of issues that we can think about here. One of them is the unit of analysis, if you look on the left there. Normally, when we’re doing our research, our unit of analysis will be people, will be individuals, but not necessarily. It can be other things. It can be states. It can be localities. For those of you who are familiar with the National Incident-Based Reporting System, NIBRS, one of the issues there if you have multiple units of analysis, depending on which piece of that you’re looking at.

Stan Orchowsky: For example, in NIBRS, one unit of analysis is the incident, so you have an incident level of data. That’s one possible level, but in an incident, you may have multiple offenders.

Erin Farley: Okay [inaudible 00:16:49] backing up.

Stan Orchowsky: Within an incident, you may have multiple offenders. There’s also a unit of analysis that can be a defender level, so we can look at all of the offenders that were involved across a number of incidents. Within another unit of analysis is victims, so you can actually have, obviously, a number of victims across incidents. You can have a number of victims for a given offender and a number of offenders for a given incident. So, it’s important to understand and recognize what unit of analysis you’re dealing with and for any particular data collection.

Stan Orchowsky: Another issue is this measurement issue. We kind of discussed this before when we talked about conceptualization and operationalization. This would be the idea of operationalizing and assigning numbers to observations, for example, the development of a likert-scale to measure something, which is the idea of putting things on a scale from, say, very important, somewhat important, not very important, very unimportant, that sort of thing.

Stan Orchowsky: Then finally, looking at just another issue to consider is the issue of missing data, so when we don’t have meaningful information for given observation. The important thing here is to think about why the data may be missing. So, for an individual, if we’re collecting observations or data on individuals, the data may be missing because there’s no response. The data may be missing because a response is not applicable to a particular question. The data may be missing because the person gave a response, but it was not a valid response, and a whole host of other reasons, and of course, when you’re dealing with large datasets, that becomes an even bigger issue.
Stan Orchowsky: Okay, finally, this also should be the levels of measurement issues should be something that you may recall from stats 101 that we have nominal, ordinal, interval, and ratio kinds of variables, with nominal variables being the categorical types of variables, so for example, gender, male, female, or race would be examples of nominal variables. Ordinal variables, which indicate difference and can be ranked along the scale from low to high or from having less of something to more of something. If you think about, for example, categorizing offenses according to level of seriousness, so trying to develop a measure where we have homicide at the top as the most serious offense, and then rape, sexual assault has maybe a little bit less serious. Robbery is a little less serious than that, perhaps burglary on down. Then that becomes sort of an ordinal level, so we can roughly have a sense that a homicide is a more serious offense than a simple assault, but we don't have any way of gauging how much more serious.

Stan Orchowsky: An interval level variable, however, gives us a sense of that. An interval level variable, you have the same continuous measurement, but the size between the intervals is the same anywhere on the scale. The easiest example to think about here is in terms of temperature. When I left my house this morning, it was 28 degrees, and now it's about 48 degrees. So, that's 20 degrees difference. That's the same interval as if I lived in Orlando, and the morning started out at 48 degrees and it's now 68 degrees, so having those intervals that are the same. You could also think of intelligence as another example. It is the hallmark of the interval level.

Stan Orchowsky: Then finally, the ratio level, which is the same, except for we have a true zero point that actually is a meaningful sort of a thing. When we think about temperature, for example, zero is a valid temperature, but it doesn't mean lack of heat, whereas on a ratio level, zero means zero. So, any sort of account that you could imagine, a sentence length, a number of prior offenses would be examples of ratio level measurement where zero offenses means zero offenses. It means we have no criminal record. If we have two offenses, we have half as many as someone who has four offenses.

Stan Orchowsky: All right. I believe that's it for me. Thank you. I am going to turn this over to Erin.

Erin Farley: Okay. Hello, everyone. Stan, can you check and see if you can hear me? Oh, hello, welcome. Okay, great, wonderful, all right. I am taking over from here and going to first be talking about frequency and distribution, so moving into a discussion on how to summarize your data.

Erin Farley: Frequency distributions are an effective way to summarizing various types of quantitative data. Frequency distribution presents the count of how many observations fall within each category. Cumulative frequency is for every given score, the total number of cases in a distribution at or below that value. Then cumulative percent is for a given score, the percentage of cases in a distribution at or below the value.
Erin Farley: Before jumping into descriptive statistics, you want to first examine the frequency and distribution of your data, and so I have prepared a number of examples. This is what you would see in SPSS. What I did for this first example, is I pulled the monitoring the future 2014 dataset off of ICPSR, which some of you may be familiar with. Looking at this screenshot, I want to draw your attention to analyze on the top of the taskbar. It’s highlighted, and then if we’re interested in frequency, you can see that you would first go to descriptive statistics and then it would create a new list of options, and you would click on frequency.

Erin Farley: Then it would be followed by something that would look similar to this. You can see all of your variables and your dataset are listed on the left. You would find whatever variable you were interested in and use the arrow button to click it over to the right. Then for right now, we’re just very simply looking at a frequency, so we would go ahead and press okay. Just to give a little further information, the measure that I selected that, as you can see, is Aver Grade, that actually is a question. Which one of the following best describes your average grade in this school year? Okay.

Erin Farley: And so, here is what I see when I run the frequency distribution. As a side note, anybody on the [inaudible 00:23:25] who might be familiar with monitoring the future data, if you downloaded this, actually, they're pretty good at handling missing data, and you would not see missing data. They've already ... Or questionable data, like you see, negative 99, they've already sent that missing. They've already cleaned the dataset. But for today, I wanted to [inaudible 00:24:44], and so I pulled it back out and put it in the dataset.

Erin Farley: If you look at this variable, you can see that the responses range from D to A. Actually, if you look at the data, D is represented by a one, all the way to A, which is represented by a 9. Before interpreting the percent, the valid percent, and the cumulative percent, however, you first have to handle that missing data that is 1,202 cases. What you want to do is send them missing. You can point and click and do that in [inaudible 00:25:21], or you can write syntax.

Erin Farley: For just an illustration, I have included the recode syntax that I utilized to send those missing cases or non-response cases missing. You can see the output to the right, and now you see that that is gone and it’s actually been moved or relocated to the bottom. It starts with D, which has those responded that their average grade was D all the way to A. Okay.

Erin Farley: If you look at this SPSS output, if you look, you can see that the students who reported A, that's 5,116. That's 18%. Comparing, if you want to look at preparing the percent and the valid percent, what you can see is that the missing is still included in the percent column. What you want to do is you want to look. As a rule of thumb, always look at the valid percent. Whether you have missing data or not, I just think it's a very good rule to stick by. You can see that the largest percentage of kids reported that they had an A, an average grade, at
18.7%, followed pretty closely by A- with 17%, and then a B+ with 17%. Now, if you looked at the cumulative percentage, you could say that approximately 50% or 47.3% of students reported that they had an average grade of B or less.

Erin Farley: Now, this is the test. To begin my Excel example, let's see if I can get this right. Part of this presentation is presenting screenshots that I did of SPSS to avoid any possible trouble, but for Excel, it's a little more nuanced, and so I actually am going to attempt to go live with an Excel example, so let's see how this goes.

Erin Farley: What I did was, I pulled ... If you look to the left, you can see students. It's just numbered. I just numbered each value, grade, and it's 1 to 40. I just pulled 40 values. I got rid of all the missing, so it ranges from one to nine, but say you did not know what your range was. What you wanted to do is ... Let me pop up my menu here. Okay. What you wanted to do is go to data, and then let me make sure I get this right. Bear with me. Okay. If you highlight all of these, it looks a little different than what I was working with, so just give me one second. Okay. Go to data. It looks a little different. Okay, so we might skip ... Oh, wait here we go. All right, everything is okay. All right, off to a rough start, but we're doing okay.

Erin Farley: Go to advance. Click advance. Here we go. Here, it says you want to copy to another location. Here is your list of range. Those are all the values of the grade, the scores. Then you're gonna copy. Since I have done this exercise before, it's gonna pop up in the E column right there. I want unique records only, so then you press okay. Let me try that again. Control, shift, here's another way, control, shift, down, and then go to data advanced, copy to another location, unique variables only. One more time, and then we'll move on. Okay, so it's being cranky, which is expected since we're live, so I'm gonna cheat and I'm gonna enter it in myself, but theoretically, it would put all of this right there.

Erin Farley: We know that the values are one to nine, but they're not really one to nine. What they do is they represent a grade. Here's the grade. Okay. Here, if you wanted to find your frequency, what you would do is I'm using the array function, so you would highlight the whole section. Then you would press the equal sign. Start to type in a frequency. You would double click on frequency, and it's gonna ask you for your data array and then your bins array. Your values are your data array, so you're gonna highlight those, comma. Your bins array are your values right here, your elements. Then you're gonna close it and then enter, and there it is. It's one. Let me try that again. The dangers of going live, okay. Highlight ... I know what I did wrong. Okay, this is a good lesson. For using the array function, you actually have to use control, shift, and enter. There you go. Okay. There is the frequency.

Erin Farley: If you wanted to do the percent frequency, it's quite similar. You would highlight again. You would start by typing in equal, type frequency, there it pops up, you double click on that, then go back to your values, highlight values, there it is, comma, go back to your bin elements right there, but then what you would do is
forward slash count. Then you would go back to your parentheses, and then you would enter this in, there it is, end, close the parentheses, and then again, you control, shift, and enter. There are your percentage frequencies.

Erin Farley: Then I'm gonna flip this a little bit. That's the array function, but if you wanted to do it just one by one, which I'm just only gonna show the first example, it's pretty simple. You would do equal, and you would do one, forward slash. Then you would do ... Actually, you know what? I missed a step, which is not a problem. Easily, you want to sum your frequency, which you can do by equal and alt. There you go. Here, if you wanted to do it one by one, 'cause there are a million ways to calculate these measures, and so I'm just picking a select few. Those of you who are on the webinar today watching the webinar, you may be very experienced in Excel. This is for those individuals who may be less experienced in Excel. Here, just again, one example is you would type equal, click on that box, forward slash, and then click on the total, and then here, it would be enter. There you go. Okay. That covers frequency distribution.

Erin Farley: Now, I'm gonna hold off discussing cumulative distribution. When I present example, [inaudible 00:34:05]. I want to go back to the slides. There we go. Okay, and discuss measures of central tendencies. You have mean, median, and mode, and so I'm sure, again, many people are familiar with mean, median, and mode. Mean is the average group of scores, the limitation. The benefit of a mean is that, of course, it accounts for every value that you have. That is also a drawback because it does take into account every single score. If your dataset or your values have an outlier, you're gonna have possibly some misrepresentation possibly in the value of your mean. That leads us to median and mode.

Erin Farley: Now, median is the score that divides the distribution of scores exactly in half. Your scores first, they must be rank ordered. With an odd number of scores, the median is the score that lies exactly in the middle of the distribution. With even numbers, the median is the average of the two middle scores. The median remains the same, no matter how skewed the data are. It is insensitive to extreme scores, and so that is why it is preferred in terms of measures for central tendency. Disadvantage of the median is that by taking every score ... Oh, that's actually ... Never mind.

Erin Farley: Moving on, sorry, I was reading a different line, moving on to mode. Mode, in the distribution of scores, either categorical or continuous, the mode is the value that is most frequent. It's an important to note you can have more than one mode. If you have one mode, it's called unimodal. If you have two modes, that is bimodal distribution. Advantages to the mode is that it has no assumption regarding level of measurement, but again, a disadvantage is it fails to take into account all levels of data. So when you present data, it really is just a good rule of thumb to present all of these so that it gives a good sense of what the overall data looks like.
Erin Farley: Here, moving on to another example, here is an SPSS example for finding the mean, median, and mode. Again, I'm using monitoring the future data. I picked a different question. This is perceptions of safety or feelings of unsafety when going to and from school. When you, again, go to analyze frequencies, you would want to click on statistics, and then this additional table would pop up. You can see a lot of things that you can pick. Right now, we are interested in the upper right hand corner of the box, and here you can see that the mean, the median, and the mode are checked. There, you would press continue, and from there, okay. This is what would pop up.

Erin Farley: Here, you can see this is actually an untampered with variable from the dataset. You can see where the missing and the not asked on this form has already been taken care of. You can see the distribution, primarily a large segment of the response is never, 59.3%. Then above that, though, you see the box that carries the mean, median, and mode. The mean is 1.59. The mode is 1. Or, I'm sorry, the mean is 1.59. The median is 1, and the mode is 1.

Erin Farley: I wanted to provide an additional example where the data is a little bit messier. This is data that I pulled from the victim segment of NIBRS data. I aggregated the data just to some number of victims for homicide or non-negligent manslaughter by state. You can see the range is 1 to 624. One state had 1 victim, and another state reported 624. Now, remember, for the NIBRS, one offense can have multiple victims. There's a wide range here. We only, again, have 35 states, so we're missing a good number of states, as well. If you look to the left, you can see the mean, median, and mode. Now, the mean is 105, but the median is 42, and the mode is 12. Actually, if you look at the footnote, this is a bimodal distribution. So, not only is the mode 12, you can actually scroll down and you can see that another value, 138 also contains, essentially, two states reported 138 victims of homicide or non-negligent manslaughter.

Erin Farley: I also, just as a reference, I included the syntax above. Here, we see that there is about those lower numbers. Specifically, I just want to highlight the state that reported 624 are really pulling that mean up when you look at the median and the mode. What I did as just an illustration of how it can impact or how you can change your data, depending on, again, your research question, the quality of data, and your analysis plan, you can alter things by removing values. To the left, you see the summary statistics that were on the other page, 105 for the mean. The median is 42. The mode is 12. I included, again, the syntax above each of these boxes for anybody who is interested.

Erin Farley: What I did was, you can see in the right syntax, select a state does not equal 21. That state that had a value of 21, that was the state that was reporting over 600 homicides, so I took them out, and then I ran the descriptive statistics again. You can see noticeable change in the mean. The median doesn't change too much. The mode stays exactly the same, but the mean drops 15 points, so brings it back into ... Just tempers the value of the mean a little bit, but if you can remember what was in that distribution, you still had a handful of states that
had values over 300. So again, deciding how to handle outliers all depends on what data you're working with, what your research question is, and also what your analysis plan is going to be.

Erin Farley: Okay, so another Excel example, I have good feelings about this one now. I'm going to change files. Give it one second. Okay. Here it is. Here, I pulled the 35 states and their aggregated count of homicide, non-negligent homicide, and then I also, just for another comparison, I pulled adaptive motor vehicle parts. Okay. The first section is just mean. What you want to do, equal. Now, for Excel, mean is actually average, so you want to start to type in average. You can see that. You just double click on that. Then you highlight your values. There it is.

Erin Farley: So then, you do the same thing for median, and you just start to type in median, and it should pop up. You double click. You highlight your value ranges, your responses. You close it, and then you press enter. That's probably something I should mention, is you want to ... If you're dealing with an array function, which again, some of you may be more or less familiar with, you want to click control, shift, enter, but for these very simple functions, you're just pressing enter. So again, mode pops up, select your values, enter, okay. Then standard deviation, we'll leave that for the next example.

Erin Farley: Now, here's why I put in theft of motor vehicle parts and accessories as just another example, is what you can do is you can highlight this mean for homicides, but you see in the corner, if you put your cursor over it, it turns into a plus sign, a black plus sign. If you drag that over and unclick, it provides you with the same calculation for the second column, so you don't have to keep repeating yourself if you're doing this for large amounts of variables. Again, for the median, if I just pull the corner across, 936 is the median for motor vehicle thefts, accessory thefts, and then pull the mode, and here, this is an NA. That actually is correct. The reason why it is an NA is 'cause there is no mode in that dataset. That is mean, median, and mode.

Erin Farley: I'm going to quickly hop back to my PowerPoint slides to continue. Let's see. It just takes a second. Okay, so the next area to discuss for summarizing and describing your data are measures of dispersion, which include range, mean deviation, variance and standard deviation. Range is the easiest measure of dispersion to calculate. It is the difference between the maximum and the minimum score. The mean deviation measures the average distance of each score from the mean. Now, the variance is the square of the standard deviation, and we generally don't use variance as an index of spread because it is in squared units. So instead, we use standard deviation. Standard deviation is the square root of the variance. It measures the spread of a set of observations. The larger the standard deviation is, the more spread out the observations are.

Erin Farley: Just a couple other characteristics to describe a standard deviation, the smallest possible value of a standard deviation is zero, but this would only happen if every value was identical. You're probably not going to find that standard
deviation very frequently. The standard deviation is affected by outliers. That is because it is based on the distance from the mean. For examples today, we are just going to focus on the range and the standard deviation. Okay.

Erin Farley: Part of, both, talking about the standard deviation, as well as reaching back to discussing the mean, median, and mode leads into also understanding the skewness of your data. Here are a few figures to demonstrate how your data may or may not look if you run a figure, or histogram, or examine the distribution of your data in comparison to the first image, which is an example of a normal distribution. As you can see, the mean, median, and mode are all exactly the same value, which is probably highly unlikely in real data, but that's what one would look like. The mean would be zero in this example, and the standard deviation would be one.

Erin Farley: To the right are two examples of different types of skewed data. The negatively skewed data is the first one. If your data is negatively skewed, it means the mean is less than the median, and the long tail on the left side of the distribution is where you see the tail on the left side, and this is where there are low scores. In contrast, if you have a positively skewed dataset, which is the one all the way to the right, C, the mean is greater than the median, and the long tail would be on the right side of the distribution. Again, if you have skewed data, it is important that you present all these descriptive statistics, the mean, median, and mode so that it does present an accurate picture of your data.

Erin Farley: Okay, so here's another example in SPSS. I am using the offense measure from the NIBRS victim segment data. Here, I went the same path that I had earlier, analyzed down to descriptive statistics to frequencies, but here, when I clicked on statistics and this additional box pops up, you'll see that I have clicked down on the lower left hand corner of the box, standard deviation range, minimum, and maximum. This is the result. You have a range from 623, which is 1 minus the maximum, which was 624. Standard deviation is 142.96. Then I go back. Okay, there might be ... Let me hop back and make sure I didn't miss. Okay, sorry, I almost did skip this.

Erin Farley: Here, if you want to actually present it in a graphical format and see how it looks in the histogram, you would then go to the same exact path, but below the statistics button is a charts button. You have your options, none, bar charts, pie charts. Here, you want to hit histogram. You can show the normal curve on the histogram, and I checked that, press continue, and then something like this will show up. This is, again, the NIBRS victim segment data based on the 35 states that reported information on victims of homicide or non-negligent manslaughter.

Erin Farley: As you can see, there are a large segment of states that are lower to the left side of the center of the table, but you see that there is a long right tail. This would actually, leading towards the higher values where you can see sticking out at 600 is that one single state. Here, this dataset, you would describe as
being positively skewed, but you know it's really that one state that's dragging it out, and then you can see the states that had the values, about 300 that I had referenced, as well. You can see those are kind of sticking out, as well. Someone might want to have a critical discussion on, again, what the purpose of the research is, what they're going to be doing with the data and if they want to keep those states in or not.

Erin Farley: I wanted to provide an additional example since the state data is kind of unique. Here, I went back to the monitoring the future, and I pulled that unsafe variable. Again, I clicked standard deviation, mean, mean, maximum, and range. Here, you can see it puts out the frequency distribution, as well as the box above it, that lists the standard deviation is .861. The range is 4. The minimum and maximum is 1 to 5. Anybody who has done survey research is gonna recognize that as a likert scale from never to every day, which is frequently used in a lot of social science research.

Erin Farley: Now, one thing to ask is, what does that standard deviation mean? We know that the larger the value gets, the more variation we have, but what might be interesting, for example, again, depending on what you are looking at is, here we are looking at schools. One might want to look at how students answered this question in school A in comparison to how students answered the same question in school B, and then examine utilizing the measures of dispersion to see the variation in how students across different schools actually answered questions like this. That's one way to compare the value of standard deviation. Oh, I jumped a little bit.

Erin Farley: Okay, so here is the histogram. Again, I printed out a histogram. See the values from one to five. Again, this data is positively skewed with most of the students reporting a value of one, feeling pretty safe in their school or going to and from school. Here, I wanted to add just one additional histogram that is a true continuous variable. This is the age range that I pulled from the NIBRS data. I did, however, as a caveat, I did lap off anybody who knows NIBRS data, I should mention that they have a category for newborns under one, and they have categories of people over 98.

Erin Farley: What I did was I actually lapped those categories off, sent them missing to create a continuous variable to actually show you what a very large dataset with a continuous age variable might look like. This is what it would look like. Again, it has positively skewed. You can see that a lot of the ages are sort of lumped on the lower scale, 20s and 30s, which makes sense. We're talking about victims of crimes. As we know, the victims of crimes, a violent crime, like homicide, are usually younger in age, no surprises there. Okay, so moving on to my Excel example. We're getting close to that one hour mark, so let's see how fast I can do this. It makes me click twice. Okay, here we go.

Erin Farley: I put in standard deviation twice, so ignore that. We'll start here at the second box. Standard deviation is pretty much the same way you would calculate mean,
median, and mode. You want to do equal. Then it's STDEV, double click on that. Highlight your data. Close it and press enter. There you go. Same way, equal, max, double click, highlight your data, close it, and then enter, and then your minimum, minimum, oh, I forgot to double click. Hold on. Min, go up, scroll down, there you go, one and now to do range, what I was doing is just click equals and then do max, minus sign, then click on the minimum box, and then press enter. There you go, 623, and so we know all of that is right 'cause we've seen it before in the SBSS output. Then what you can do, if you wanted, again, additional variables, you want to do the same exact thing. You want to save some time, just pull on the corner, drag it, and it'll calculate the same things for the additional data. Okay.

Erin Farley:

Then for creating the histogram, what you want to do is you want to go to a data analysis pack, and so you want to make sure you have this. What you can do is if you go to your main sort of office button right here, you can go to Excel options, go to add ins, and you can check and see right there. You can see that there is an analysis tool pack, so I added that in, so I have my little analysis right here. If I click on that, it's supposed to, there we go, okay. Here, you have a variety of things that you can do in Excel now that even I didn't know you could do before I looked at this analysis pack, like covariance, and [inaudible 00:56:58], and things of that sort, so there's amazing things that you can do in Excel, but for today, we're just doing the histogram.

Erin Farley:

You click on histogram. You click okay, and then you want your input range, again, is just your values, your response values. Then you have your bin range, and I use it interchangeably, bin range elements, these are your bin range. Oh, click in your bin range first. Then enter it. Okay, and then a new worksheet, that's okay, or you can do a new workbook, whichever, and then cumulative percentage, and then okay. Oh, I have the same ones. Oh, I see, there is an error. Okay, hold on one second. Let's just see if we can quickly ... I do have it saved in case we have technical difficulties. Okay.

Erin Farley:

Okay, so there it is. The frequency pops up, and then the cumulative percentage here, the distribution by, again, one to nine. One, remember, is D. Nine is A. There were, in case anybody is that observant, there were no D pluses in this group. It goes from a D to a C minus. Two is actually a C minus. Here, you can see the distribution. I see that the histogram did not immediately pop up, but what I can do ... I just got rid of it. Hold on. Bear with me. Okay, okay. It showed here you can see the bin, the frequency, and the cumulative distribution, it's exactly the same. Here's the histogram. You would actually have a histogram that looks something similar to this. Okay. Oh, did I pause? Having me in control of any technology is always cause for an interesting story. Okay, okay.

Erin Farley:

I actually believe that is it. We are just a couple minutes over. If anybody has any questions, specifically maybe that last component about histograms, please do not hesitate to email me or make a comment, and we will follow-up on any
questions that anybody has. I know I did go a little fast, and so yeah. I appreciate any questions. I'm going to pass it off to Stan.

Stan Orchowsky: Thanks, Erin. We're gonna go ahead and open up the poll now, the evaluation poll. If you could take a few minutes and do that while we're waiting for questions to come in, and let me ... As I mentioned at the outset, this is designed to be the first in a series of webinars addressing statistical analysis for criminal justice research. I wanted to give you a little sort of preview of what's coming up in the future.

Stan Orchowsky: The next thing we'll be talking about is just sort of the basics of sampling. Again, all of these are gonna be designed to be roughly an hour to an hour and 15 minutes, so we're not gonna be able to get into a whole lot of detail about any of these, but you'll get the idea. Significance testing, specifically comparing proportions, and then significance testing comparing means, correlation and simple linear regression. We're gonna take a little pause at that point and sort of switch gears and talk a little bit about displaying data, so some of the things that Erin was talking about today, we want to look at in addition to histograms, bar charts, and pie charts, and what you should use, and when you should use it, and so forth and so on. We'll take a little pause there and talk about displaying data, and then we'll go on to somewhat more advanced topics, multiple linear regression, logistic regression, which is used when you have dichotomous dependent variables, and then finally, finishing up with exploratory data analysis, looking at things like factor analysis and discriminate analysis. That should provide.

Stan Orchowsky: Then we might add in a few topics based on the feedback that we get, and certainly, if you have any, we're happy to hear that, but that will give us, I think, a sense when we get through all of that, that will give you pretty good grounding in terms of at least touching on some of the really important topics. There are obviously many more specialized topics and more advanced topics that we can get into, and we may get into those in the future. Paralleling that, we hope to have, at some point, some discussion of experimental design, which is a whole other set of issues, but this is the future topics for the statistical analysis series.

Stan Orchowsky: All right. Again, the poll will be open, I think, for a few more minutes. I think that wraps up our first statistical analysis webinar. Thank you all very much for your attention, and we will see you next time.

Stan Orchowsky: (silence)