

Elevate CBOs Webinar Series: Data 101 Workshop

We will be starting soon, thank you for joining us!



Attendees are muted, so please share comments and ask questions in the **Q&A box**



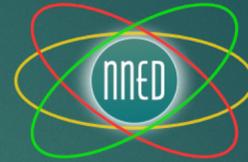
Closed captioning is available through Zoom using the **CC button**



Resources and the recording will be available following the webinar at share.nned.net



Provide us with feedback in the post-webinar survey



Elevate CBOs Webinar Series

Data 101 Workshop

MAY 26, 2022

LOGISTICS



Share comments
and ask questions
in the **Q&A box**

LOGISTICS



Closed captioning is
available through Zoom
using the
**CC button / Full Live
Transcript** option

LOGISTICS



**Resources and the
recording** will be
available following the
webinar at
share.nned.net

LOGISTICS



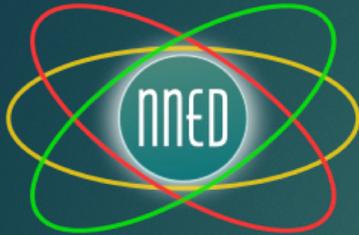
Provide us with
feedback in the post-
webinar survey



Angel Villalobos

FACILITATOR, NNED NATIONAL
FACILITATION CENTER

CHANGE CONSULTANT,
CHANGE MATRIX



National Network to Eliminate Disparities in Behavioral Health

A network of over **5,500 individuals**, including more than

1,300 community-based organizations striving for behavioral health equity for for all individuals, families, and communities.

DISCLAIMER

Some views, opinions, and content expressed in this webinar may not necessarily reflect the views, opinions, or policies of the Office of Behavioral Health Equity (OBHE), the Substance Abuse and Mental Health Services Administration (SAMHSA), or the U.S. Department of Health and Human Services (HHS).



Marlon Daniel

STATISTICIAN, CENTER FOR BEHAVIORAL HEALTH STATISTICS AND QUALITY (CBHSQ), SUBSTANCE ABUSE AND MENTAL HEALTH SERVICES ADMINISTRATION (SAMHSA)

Statistics 101

MARLON G DANIEL, MPH MHA CPH

POPULATION SURVEYS BRANCH

CENTER FOR BEHAVIORAL HEALTH STATISTICS AND QUALITY

Statistics

THE **SCIENCE** OF COLLECTING, DESCRIBING,
AND INTERPRETING DATA.

The Research Process

- ▶ Ask a research question.
- ▶ Decide what information is needed to answer the research question.
- ▶ Collect the information.
- ▶ Assemble the information in a way that answers the research question.
- ▶ Analyze data
- ▶ Interpretation

Two areas of statistics

Descriptive Statistics: collection, presentation, and description of sample data.

Inferential Statistics: making decisions and drawing conclusions about populations.

Basic Vocabulary

15

▶ **Element:** The individual for whom the information is collected

▶ **Variable:** A characteristic about each individual element of a population or sample.

▶ **Data (singular):** The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.

▶ **Data (plural):** The set of singular data values collected for the variable (s) and/or from each of the elements.

▶ **Data set:** The entire data collected.

▶ **Database:** The program/tool used for the data entry.

I.D	Age	Gender	Biostats	CHOL
1	23	M	Yes	249
2	24	F	No	205
3	35	F	Yes	286
4	28	M	No	187

Example:

A researcher want to use the NSDUH survey to determine the average number of times in the last 30 days a respondent drank alcohol

- ▶ The *variable* is the number of times alcohol was consumed
- ▶ The *element* is each respondent.

- ▶ The *data* in the *singular* form would be the number of times for each respondent.

- ▶ The *data* in the *plural* form would be the set of values in the sample.

Two kinds of variables:

17

Qualitative, Attribute, or Categorical Variable:

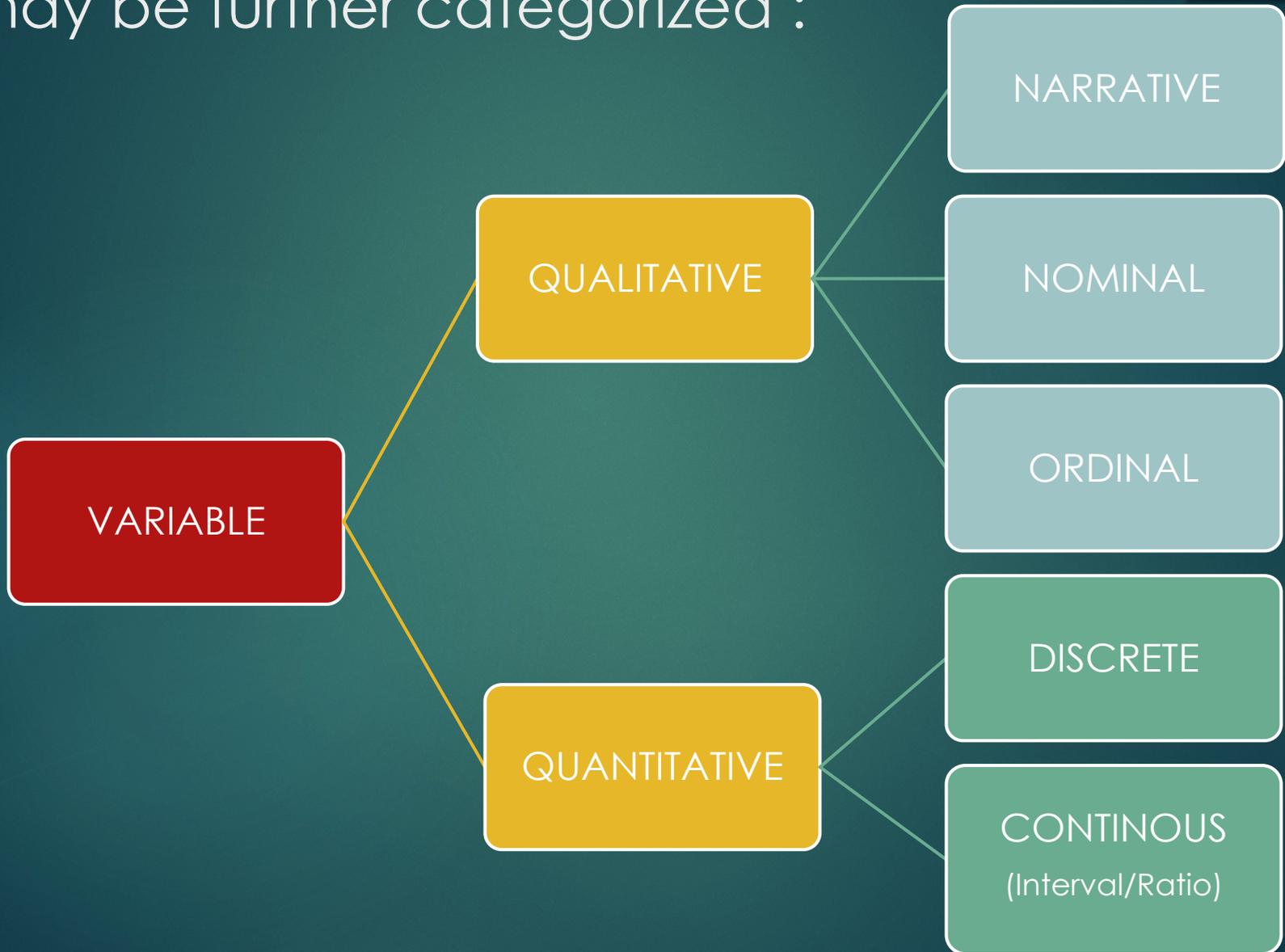
A variable that categorizes or describes an element of a population.

▶ *Note:* Arithmetic operations, such as addition and averaging, **are not** meaningful for data resulting from a qualitative variable.

Quantitative, Numeric, or Interval Variable: A variable that quantifies an element of a population.

▶ *Note:* Arithmetic operations such as addition and averaging, **are** meaningful for data resulting from a quantitative variable.

Qualitative and Quantitative variables may be further categorized :



Narrative variables

- ▶ Information recorded in narrative form
- ▶ Usually difficult to summarize in numeric form
 - ▶ Example 1: comments made on evaluation form
 - ▶ Example 2: response to an open-ended survey question

Nominal variables

20

- ▶ Variables with categories used to classify elements into groups
- ▶ Categories cannot be ranked and/or ranking is subjective
 - ▶ Gender (0=male, 1=female, 2 = Non-Binary)
 - ▶ Race/Ethnicity (1=White, 2=Black, 3=Hispanic)
 - ▶ Blood type (1=type A, 2=type B, 3=type AB)
 - ▶ Presence or absence of a Mental Health Diagnosis (1=yes, 2=no)

Ordinal variables

21

- ▶ Variables with categories that can be ranked
 - ▶ Socioeconomic Status(1=poor, 2=middle class, 3=rich)
 - ▶ Cancer Staging (I - IV)
 - ▶ Wong-Baker Faces Scale



Quantitative Variables

22

- ▶ Discrete – Numeric variables with only whole numbers possible
 - ▶ Number of ICU admissions within a month
- ▶ Continuous – Numeric variables that have an infinite number of values possible in each interval
 - ▶ Ratio- a continuous variable that is anchored by an absolute zero, where at zero the property being measured is absent
example: weight
 - ▶ Interval- There is no absolute zero point where at zero the property being measured is absent
example: temperature

Descriptive Statistics

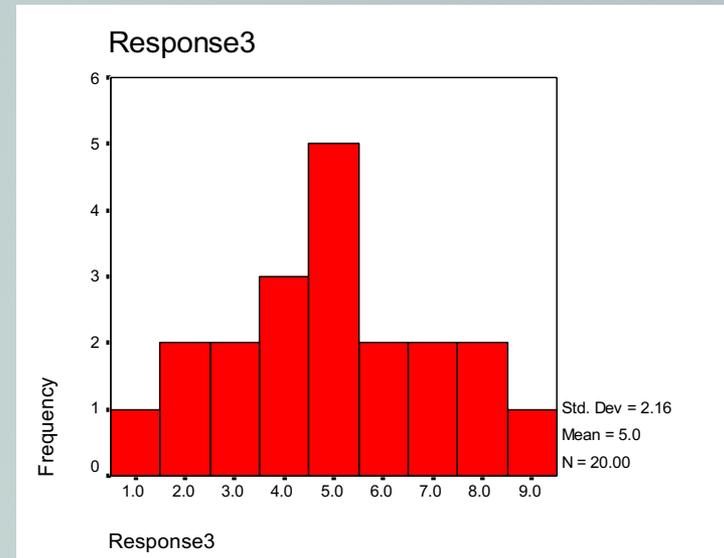
MEASURES OF CENTRAL TENDENCY
AND
MEASURES OF VARIATION

Measures of Central Tendency

- ▶ **Numerical values used to locate the middle of a set of data, or where the data is clustered**
- ▶ **Also used to summarize the data**
 - ▶ Mean
 - ▶ Median
 - ▶ Mode

Mean

- ▶ Typically referring to the arithmetic mean
- ▶ Most commonly used measure of central tendency
- ▶ The mean is affected by extreme values/sensitive to outliers
- ▶ Used for bell-shaped/normal quantitative data



$$\text{Mean} : \bar{X} = (X_1 + X_2 + X_3 \dots X_n) / N$$

where: X_1, X_2, X_n are independent values and
 N is the total number of measurements

Mean - Example

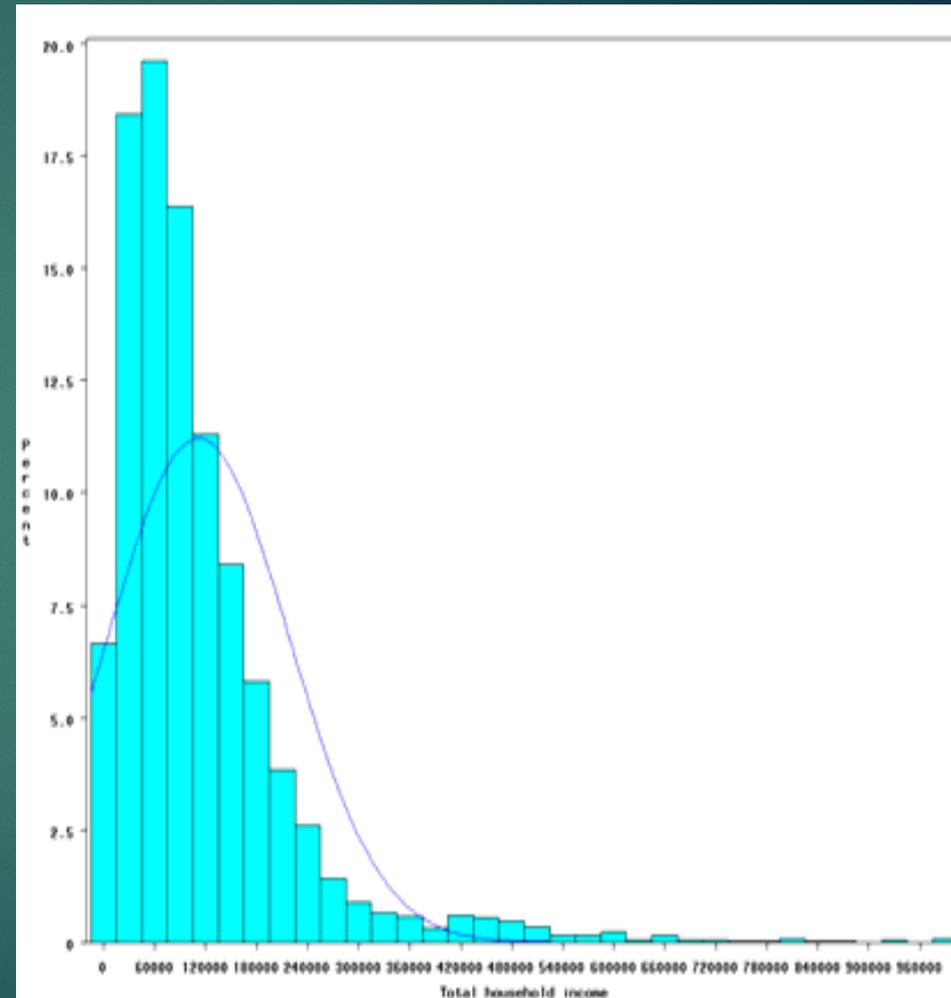
- ▶ Number of Motorcycle Crashes Per week over a 10-week period
 - ▶ 8,5,3,2,7,1,2,4,6,2

$$MEAN : \bar{X} = \frac{(8+5+3+2+7+1+2+4+6+2)}{10}$$

$$\bar{X} = 40 / 10 = 4$$

Median

- ▶ The middle value of a data
- ▶ The value which divides a ranked set into two equal parts
- ▶ Used for skewed/non-normal
- ▶ It is less sensitive to extreme observations than the mean
- ▶ Order the data
 - If n is even, take the mean of the two middle observations
 - If n is odd, the median is the middle observation



Median - Example

- ▶ Given an **even** number of observations ($n = 10$)
 - ▶ 1,2,2,2,**3,4**,5,6,7,8
 - ▶ $M = (3+4)/2 = 3.5$

- ▶ Given an **odd** number of observations ($n=11$)
 - ▶ 1,2,2,2,3,**4**,5,6,8,10
 - ▶ $M = 4$
 - ▶ $(n+1)/2 = (11+1)/2 = 6^{\text{th}}$ observation

Mode

- ▶ The value which occurs the most frequently in a set
- ▶ Used mostly for qualitative data whether nominal or ordinal
- ▶ When there are two modes the data is multimodal (more than two modes it is called multimodal)
- ▶ A modal class is an interval that has the largest number of observations

Mode and Modal Class Example

- ▶ Example: 1, 2, 2, 2, 3, 4, 5, 6, 7, 8
- ▶ Mode = 2

Age Distribution of Individuals Sampled in the National Survey on Drug Use and Health amongst Hispanic Immigrants (n = 568)

	n	%
Less than 18	29	5.11
18 to 24	47	8.27
25 to 34	68	11.97
35 to 44	74	13.03
45 to 54	141	24.82
55 to 64	113	19.89
Greater than 65	96	16.90



Skewness

- ▶ Skewness defines the degree of asymmetry of a distribution
- ▶ In symmetric (bell-shaped) distribution the skewness has a value of ~ 0
- ▶ When a distribution has the larger distribution of values on the right (the mean is less than the median) it is considered to be negatively skewed (also known as skewed to the left)
- ▶ When a distribution has the larger distribution amount of values on the left (the mean is GREATER than the median) the distribution is considered to be positively skewed (skewed to the right)

- ▶ If the MEAN, MODE and MEDIAN are the same then one can determine that the distribution is **normal distribution** and has a skewness of 0.

Measures of Variation

Measures of variation (deviation) of the values around the mean

- If the numbers are near the mean, the variation is small
- If numbers are far from the mean, the variation is large

There are three measures of variation:

1. Range: largest value minus smallest value
2. Variance
3. Standard deviation (square root of the variance)

Range

Range = Maximum Value minus Minimum Value

Concentration of a therapeutic agent in 10 vials of product:

200, 205, 205, 201, 199, 195, 202, 205, 205, 20

$$\bar{X} = 202.4$$

$$\text{Range} = 207 - 195 = 12$$

Many times in the literature you will find the minimum and maximum values presented instead of the difference between the minimum and maximum values.

i.e The mean systolic blood pressure of the patients was 135 (69 – 178)

Variance

- ▶ Used to measure how far data points are from the mean on the average (dispersion)

$$\text{variance} = \sigma^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}$$

The variance in this equation is signified by the Greek letter sigma (σ) squared. The reason for this is that the sum of the deviations from the mean always sum to zero making this calculation somewhat problematic. You will only see sigma when calculating the variance for a POPULATION. SAMPLE variance is represented by s^2 . With the following formula. Due to the variable nature of a sample, a better estimate of the population variance can be changed by subtracting one from the Denominator.

$$\text{variance} = s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$$

Standard Deviation

- ▶ As noted with the variance formula, the unit of measurement is the squared units of the original variables' unit of measurement
- ▶ To allow dispersion to be interpreted as the original units, take the square root of the variance.
- ▶ This process gives you the standard deviation
- ▶ Which leads to the question, which is the correct measurement for dispersion?

Standard Deviation

Population Standard Deviation = $\sigma = \sqrt{\text{variance}}$

$$\text{Standard Deviation} = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}} = \sqrt{6} = 2.45$$

Population Standard Deviation = $s = \sqrt{\text{variance}}$

$$\text{Standard Deviation} = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} = \sqrt{6} = 2.45$$

Percentiles and Quartiles

- ▶ A percentile is the percentage of a distribution equal to or below a particular number
 - ▶ For instance, you might have a score that seems great on the GRE, but 80% of the distribution scored better than you so you are at the 20th percentile 😞
- ▶ Percentiles are usually used to compare individual values with a norm
- ▶ The 50th percentile is also the Median

Percentiles and Quartiles

- ▶ Quartiles are an extension of the percentiles
- ▶ Your distribution is divided into 4 equal sized groups (quartiles)
 - ▶ The first quartile is the 25th percentile
 - ▶ The second quartile is the 50th percentile (Median)
 - ▶ The third quartile is the 75th percentile
 - ▶ The fourth quartile is the complete data

Interquartile Range

- ▶ The Interquartile Range is another useful measure of dispersion.
 - ▶ It is particularly useful when there are extreme observations in the distribution
 - ▶ It is defined as the difference between the third and first quartiles ($Q3 - Q1$; 75th percentile – 25th percentile)
 - ▶ The interquartile range contains the middle 50 percent of the distribution

Coefficient of Variation

- ▶ The coefficient of variation is used to compare variability amongst different datasets
- ▶ By taking a ratio of the standard deviation and the mean, you attain a 'unitless' value that does not depend on the magnitude of the mean values or units of measurement of the original data for comparison

$$\text{Coefficient of Variation} = CV = \frac{SD}{\bar{X}} \cdot 100$$

When to use which?

- ▶ Here are some ~~'rules of thumb'~~ best practices to when to report different measures of dispersion
 1. The **SD** is used when reporting the mean (MEAN ± SD)
 2. **Percentiles** and the **Interquartile Range** are used when
 - a. When the Median is used (ordinal data or skewed data)
 - b. When the Mean is used but the objective is to compare individual observations with a set of norms
 3. The **Interquartile range** is used to describe the central 50% of the data regardless of the shape (whether skewed or normal)
 4. The **range** is used with numerical data when the purpose is to present the potential effect of extreme observations
 5. The **coefficient of variation** is used when the intent is to compare distributions measured on different scales or different mean values

Examples

- ▶ For the following systolic blood pressure measurements **sampled** from the CBHSQ leadership team during the NSDUH recomplete
 - ▶ The Mean
 - ▶ The Median
 - ▶ The Range
 - ▶ The Variance
 - ▶ The Standard Deviation

Blood Pressure Values							
100	98	101	94	104	102	108	108

$$\bar{X} = \frac{(94 + 98 + 100 + 101 + 102 + 104 + 108 + 108)}{8}$$

$$\bar{X} = \frac{815}{8} = 101.86 \approx 102$$

$M = 94, 98, 100, 101, 102, 104, 108, 108$

even numbers so take the mean of the middle values

$$M = \frac{101 + 102}{2} = 101.5$$

Range = max – min

$$\text{Range} = 108 - 94 = 14$$

OR

$$\text{Range} = 94 - 108$$

X_i	\bar{X}	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
94	102	-8	64
98	102	-7	16
100	102	-2	4
101	102	-1	1
102	102	0	0
104	102	2	4
108	102	6	36
108	102	6	36

$$\sum_{i=1}^n (X_i - \bar{X})^2 = 161$$

Since we are dealing with a sample we will use the sample estimate of the variance

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{161}{8-1} = 23 \text{ mm/hg}^2$$

To calculate the sample estimate of the standard deviation we take the square root of the sample estimate of the variance

$$s = \sqrt{s^2} = \sqrt{23} = 4.80 \text{ mm/hg}$$

Table 1. Erythromycin contents of 10 tablets from 2 tablet setting machines and the calculations of their Standard Deviations

Machine A	Machine B
Erythro content	Erythro content
(mg)	(mg)
249	251
242	247
252	257
235	250
257	254
244	251
264	252
249	255
255	244
240	250
Mean = 248.7	Mean = 251.1
SD = 8.72 (mg)	SD = 3.78 mg (SD)

Let us calculate the coefficient of variation from these two samples collected from a tablet setting machine.

$$CV_A = \frac{8.72}{248.7} \cdot 100 = 3.51\%$$

$$CV_B = \frac{3.78}{251.1} \cdot 100 = 1.51\%$$

There is more variation in samples collected from Machine A than Machine B

Populations and Samples

Populations and Samples

49

- ▶ One major purpose of research is to infer information from a sample, to a theoretical population
- ▶ Definitions :-
 - ▶ **INFERENCE** - Make a statement, deduction, conclusion, suggestion etc about all such persons
 - ▶ **POPULATION** – a set of elements of a finite size that will be studied; a large set or collection of items that have something in common
 - ▶ **SAMPLE** – a subset of a population, selected to be representative of the population; a group from which measurements are sought that will be inferred to the population
 - ▶ **GENERALIZABILITY** – whether the results/findings from a study can be assumed to be true for a target population

Populations and Samples...

- ▶ Values measured in a population are represented by Greek letters and are called PARAMETERS
- ▶ Estimates from a sample that are used to make an inference on a population are called 'STATISTICS' and are represented by Latin letters

	Population	Sample
Mean	μ	\bar{X}
Standard Deviation	σ	s
Variance	σ^2	s^2
Correlation Coefficient	ρ	r

Sampling

- ▶ Why do we sample??
 - ▶ Samples can be studied more quickly than populations
 - ▶ Studying samples is MUCH LESS EXPENSIVE
 - ▶ Studying populations is usually impossible
 - ▶ Samples can at times be even more accurate than population-based studies
 - ▶ If a sample is collected correctly, the error of the estimates can be estimated
 - ▶ Samples can reduce heterogeneity

How Samples are Selected

52

- ▶ Probability Sample (Random Sample)
 - ▶ Selecting a group of subjects (sample) for study from a larger group (population) so that each individual is chosen entirely by chance
 - ▶ each has an equal chance of being selected
 - ▶ Random number table:
 - ▶ no guide to the sequence of numbers
- ▶ Non Probability Sample (Convenience Sample)
 - ▶ Simply because of convenience
 - ▶ Selected at will, self-selected

Probability Sample

- ▶ Probability Sample
 - ▶ Where every element has a nonzero probability of being selected in a sample;
 - ▶ SRS – Simple Random Sampling
 - ▶ Systematic Sampling
 - ▶ Stratified Sampling
 - ▶ Probability Proportional to Size
 - ▶ Cluster Sampling
 - ▶ Multistage Sampling

Non Probability Sample

- ▶ Non Probability Sampling
 - ▶ Not based on probability models. Usually used when a probability sampling is very difficult to implement
 - ▶ Convenience Sampling
 - ▶ Intercept Sampling
 - ▶ Snowball Sampling
 - ▶ Respondent Driven Sampling
 - ▶ Time Location Sampling

Random??

- ▶ A random selection means that all human influence, both known and unknown, is removed from the selection process
- ▶ When sampling two issues can potentially arise:
 - ▶ Sampling Bias – systematic failure to observe some elements due to the design of the sample
 - ▶ Sampling Variance – how the estimates attained from a sample can vary from sample to sample

Simple Random Sampling

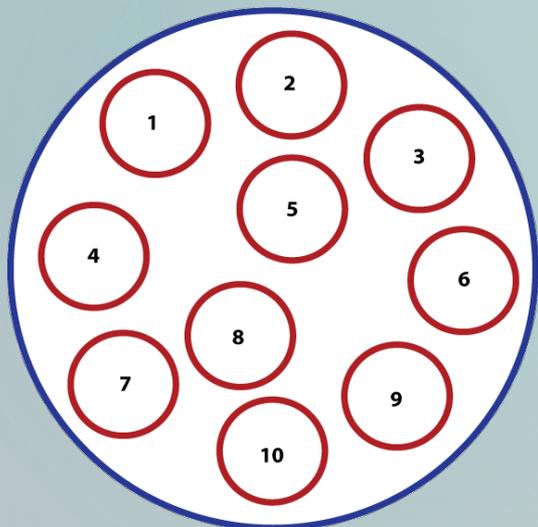
- ▶ Often used as the basic design of which all other sampling designs compared
- ▶ Assigns an equal probability of selection to each element (unit) in the sampling frame
 - ▶ Sampling Frame : list of all elements (units) in a target population
- ▶ Sample designs assigning equal probabilities to all individuals elements (units) are called EPSEM (Equal Probability Selection Method)
- ▶ Units can be selected using a random number table (either generated by a computer or from the appendix of many statistics textbooks)

Random Numbers

51469	33716	60419	80309	47197	46533	21966	22225	74564	90990	15153	69223
07012	67087	13421	47860	71100	48265	43329	31579	66828	61082	04876	73496
94195	05540	98063	20493	53201	55078	96331	99131	17289	62814	26239	82850
13016	75632	87786	24766	40383	20089	80972	71764	72832	69628	79241	50401
36920	50806	09148	60678	32647	63623	70696	76036	60014	08080	63882	23034
92463	84177	62151	01672	83109	65355	92058	11948	78836	78173	53606	27307
79645	22630	46792	74305	38652	72168	45060	52942	02740	79904	74968	63219
98467	92722	36515	78577	25834	37179	29702	52133	58283	86718	27970	30770
48928	67896	57878	14489								

Now suppose that the mean content of Calcium (mg) in a batch of dietary supplements ($N = 10$) is 250.38.

If the sample is done correctly then the mean of the sample of 4 should be close to the overall mean of the batch \pm some error



$$\bar{Y} = (Y_9 + Y_5 + Y_1 + Y_2) / N$$

$$\bar{Y} = (248.2 + 249.6 + 251.7 + 250.3) / 4$$

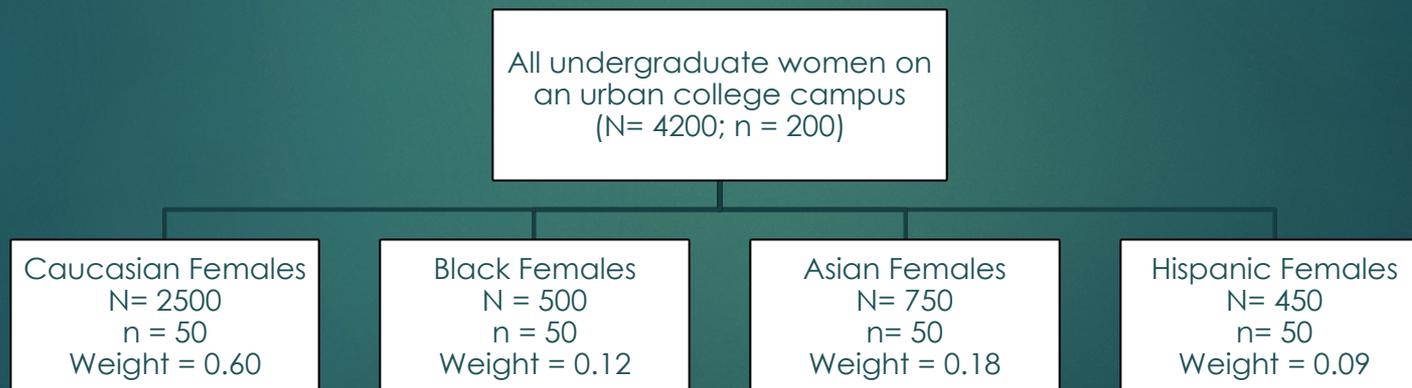
$$\bar{Y} = 249.95$$

Suppose you wanted to randomly select **4 elements** from the figure. We can start in the 3rd column. Since our finite population is 10, we will use the two last columns of that integer and select values between 1 and 10. Keep going down columns until you have selected 4 numbers that correspond to the value of the elements

Stratified Sampling

- ▶ Used to ensure representation of a particular group or subgroup (homogenous group) in a heterogeneous population
- ▶ Superior to SRS because it can reduce sampling error
- ▶ Mutually exclusive groups in a sampling frame
- ▶ These groups are called strata. Elements (units) are within each strata
- ▶ Simple Random Samples are then taken within each strata
- ▶ Used when some naturally occurring categories are smaller relative to others

Dr. Tenecia Smith is interested in the knowledge, attitudes and beliefs regarding obesity and depression amongst female undergraduates at an urban college campus. However, minority groups might be underrepresented when a SRS is taken, so a sample stratified by Race might be more useful



Probability Proportional to Size

- ▶ Also known as PPS sampling
- ▶ A technique in which the probability of selecting a sampling unit (county, food outlet, food item) is proportional to the size of its population
- ▶ This method is useful when sampling units vary considerably in size
 - ▶ This will ensure that units (clusters) with larger sizes have the same probability of selection in the sample as smaller units (clusters)

Systematic Sampling

- ▶ A method where every k^{th} sample (where k is a predetermined selection interval) is selected
 - ▶ Every 5th record in a database
 - ▶ Every 4th house in a neighborhood
- ▶ First the interval is determined, based on the ratio of the target population and sample
- ▶ Then based upon that interval, start with a random number and select at an interval after that number

In this map of a neighborhood, we select homes at an interval of $k = 6$. So every 6th home gets selected. We started with a random number and select our homes



An example of PPS sampling

63

Obs	County - State	Cumulative Sales By County
1	Adams County, NE	50245000
2	Alameda County, CA	2694835000
3	Allegheny County, PA	2717780000
4	Cook County, IL	8624577000
5	El Paso County, CO	1065870000
6	Emmet County, IA	19890000
7	Essex County, MA	1315639000
8	Halifax County, NC	66092000
9	Harris County, TX	6986928000
10	Hennepin County, MN	2186704000
11	Henry County, GA	354250000
12	King County, WA	4436237000
13	Kings County, NY	2190825000
14	Los Angeles County, CA	14957228000
15	Mecklenburg County, NC	1688765000
16	Mercer County, KY	32045000
17	Middlesex County, NJ	1168245000
18	Oakland County, MI	2044471000
19	Pasco County, FL	830362000
20	Pima County, AZ	1816074000
21	San Diego County, CA	5300672000
22	Shelby County, TN	1427556000
23	Tarrant County, TX	3076892000
24	Warrick County, IN	67314000

First we calculate the cumulative value of all units with a primary cluster. In many cases this might be the total population. In our example it would be the sales at food outlets.

Total food sales in our selected counties are \$65,119,496,000.00

An example of PPS sampling

64

- ▶ In each county we want to select 2 outlets, without replacement
- ▶ Let us look at Adams county, with total sales of \$50,245,000.00
- ▶ Our selection interval would be our cumulative sales divided by the number of selected outlets, in this case, 25122500
- ▶ We can select a random number between 1 and our selection interval for our random start. Let's use the number **4060974**
- ▶ We then can create the following series of numbers...

An example of PPS sampling

65

Store Name	Adjusted Sales	Cumulative Sales	County-State
Sunmart Foods	\$5,850,000.00	\$5,850,000.00	Adams County, NE
Russ Market	\$8,190,000.00	\$14,040,000.00	Adams County, NE
Allens Food Market	\$14,040,000.00	\$28,080,000.00	Adams County, NE
Wal Mart Supercenter	\$22,165,000.00	\$50,245,000.00	Adams County, NE

Using our selection interval and a Random start we can create this list of intervals for Adams County

Selection Series

Random Start	\$4,060,974.00
Random Start 1*Selection Interval	\$29,183,474.00
Random Start 2*Selection Interval	\$54,305,974.00
Random Start 3*Selection Interval	\$79,428,474.00
Random Start 4*Selection Interval	\$104,550,974.00
Random Start 5*Selection Interval	\$129,673,474.00

An Example of PPS Sampling

66

- ▶ Let's try Los Angeles County...
- ▶ We have total sales of \$14,957,228,000.00
- ▶ Our Selection interval will be 7478614000
- ▶ Our Random Start can be 7279306540

Store Name	Adjusted Sales	Cumulative Sales County-State
Save A Lot	\$7,800,000.00	\$1,279,278,000.00 Los Angeles County, CA
Ralphs Grocery	\$15,210,000.00	\$3,835,858,000.00 Los Angeles County, CA

Cluster Sampling

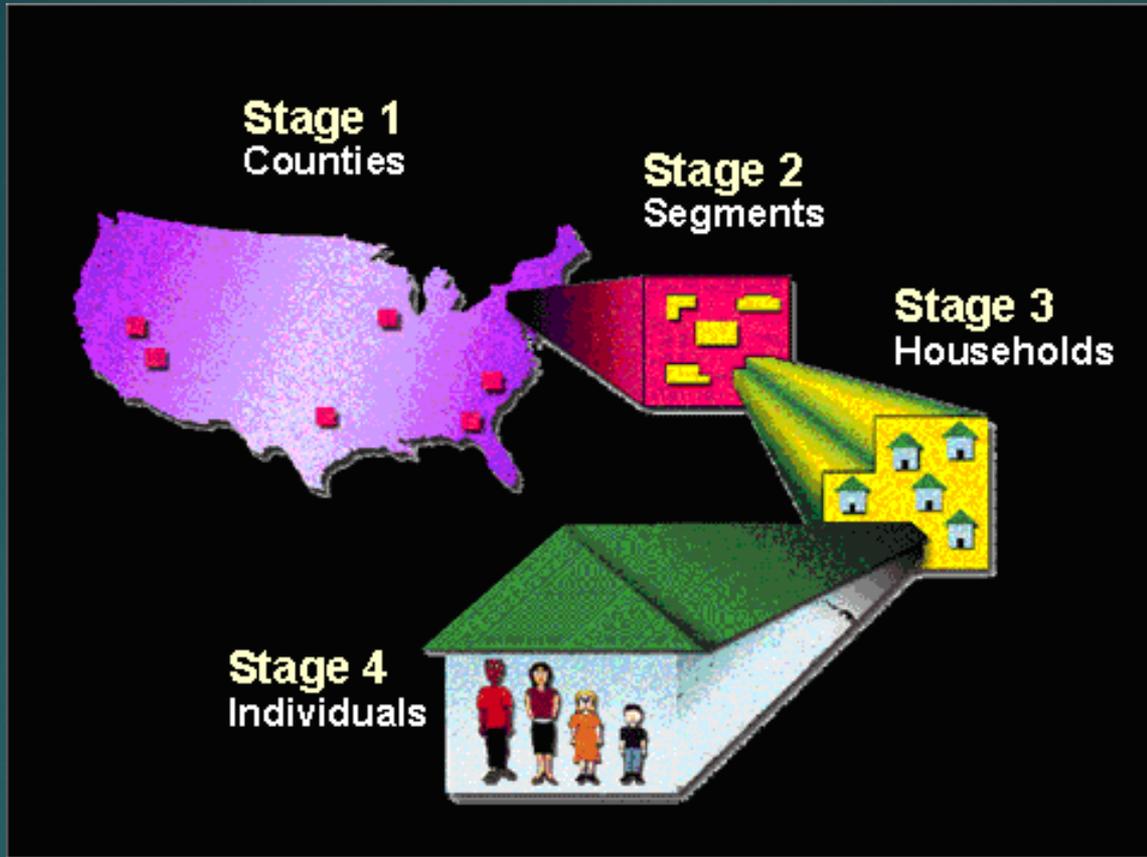
- ▶ One draw back to SRS is that it can be extremely expensive and time consuming
- ▶ Cluster sampling samples clusters of elements (and in turn these clusters become the sampling units)
- ▶ The sampled clusters then have a frame constructed within the cluster and elements are then sampled
- ▶ A major drawback is that there is increased sampling error

Multistage Sampling

68

- ▶ A complex form of cluster sampling (sometimes called Complex Multistage sampling)
- ▶ At the first stage (many times) a population is divided into naturally occurring strata or naturally occurring clusters are selected
- ▶ At progressive stages clusters within clusters are randomly selected
- ▶ At the final stage a systematic or SRS is done to select final elements
- ▶ Very common in government surveys (NSDUH, NHANES, CPS, MEPS, GSS)

The sampling methodology of the NHANES (2005 – 2006). National Center for Health Statistics



Convenience Sampling

70

- ▶ Used in exploratory research
- ▶ Useful to get an inexpensive approximation of a population phenomena
- ▶ Most effective during the preliminary stages of a research project such as pilot testing a questionnaire (i.e. to help determine a sample size)
- ▶ Cannot make scientific generalizations using this method

Intercept Sampling

- ▶ Very widely used, mostly at Shopping Malls and Public Events
- ▶ Cost effective, and easy to implement
- ▶ Not very representative of the general population
- ▶ Usually demographically skewed

Snowball Sampling

- ▶ A type of convenience sample
- ▶ Individuals from a convenience sample are asked to recruit individuals for the study
- ▶ Useful when attempting to study hard to capture and/or rare populations (Drug Addicts, Prostitutes, Honest Politicians)
- ▶ Susceptible to biases and still not very generalizable

Respondent Driven Sampling

73

- ▶ A newer method implemented to try to 'randomize' a nonprobability 'snowball' sampling
- ▶ Also used to sample difficult and hard to reach populations
- ▶ A researcher gets a group of initial contacts. The researcher then asks each initial contact to recruit a fixed number of contacts. These contacts are then asked to recruit a fixed number of contacts. These 'waves' are continued for several researcher determined iterations
- ▶ Research has found that the estimates generated are unbiased

Time Location Sampling

74

- ▶ Another new method to try to capture high risk, hard to capture groups
- ▶ Based on the knowledge that certain groups or high-risk populations tend to gather in certain locations (street corners, bath houses, IDU shooting galleries)
- ▶ TLS maps and enumerates these locations and puts them in a sampling frame
- ▶ These Locations are then randomly selected during a randomly chosen period of day
- ▶ Some debate as to whether or not this is a probability or nonprobability sample, since there is no way to determine if all the sites are all POSSIBLE sites (exhaustive)

The END

FOR NOW.....



QUESTIONS & ANSWERS

Q&A

What resources made the biggest difference for you when you were learning about how to use data?

Q&A

How do you share
data in an
approachable way?

Q&A

How do you analyze data without solely comparing Black, indigenous, and people of color (BIPOC) to the white population?

Q&A

How can small organizations use data and connect it to implementing service delivery?

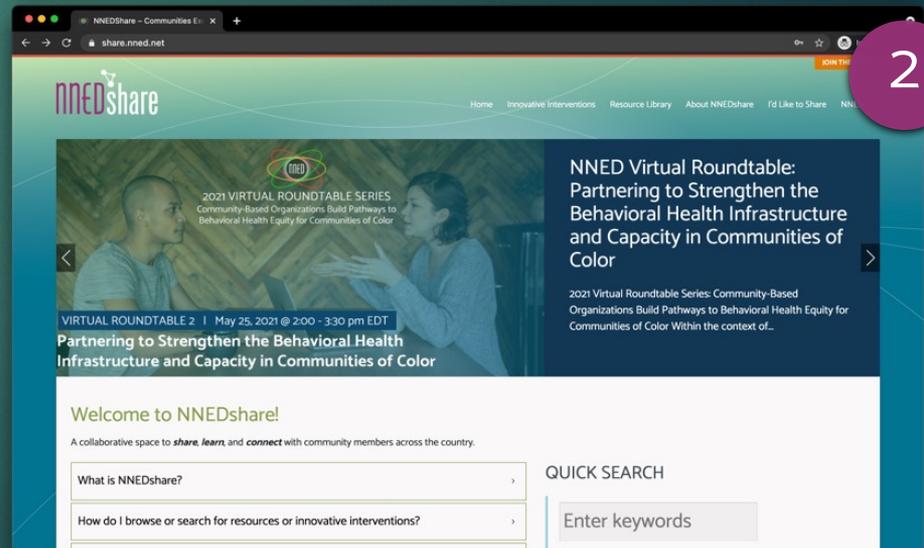
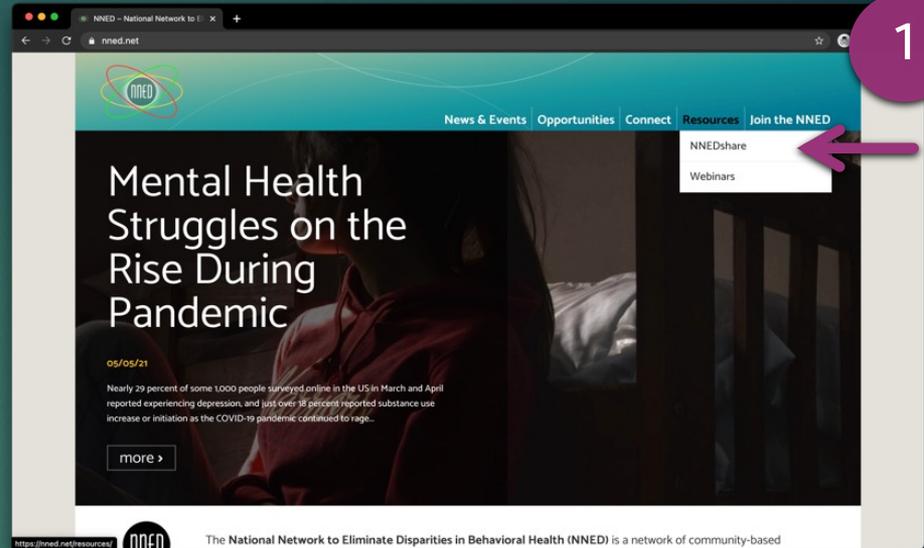
Q&A

How can
data/information
gathered be used to
apply for government
grants?

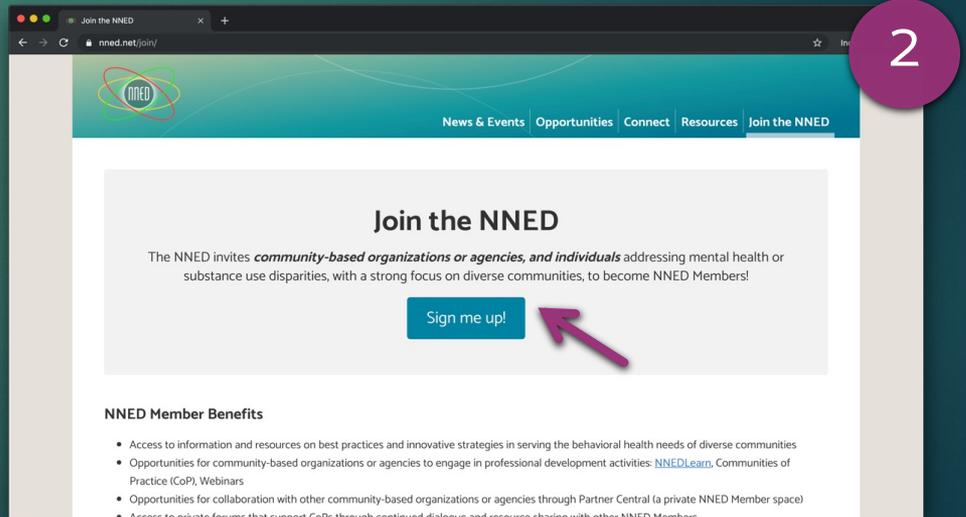
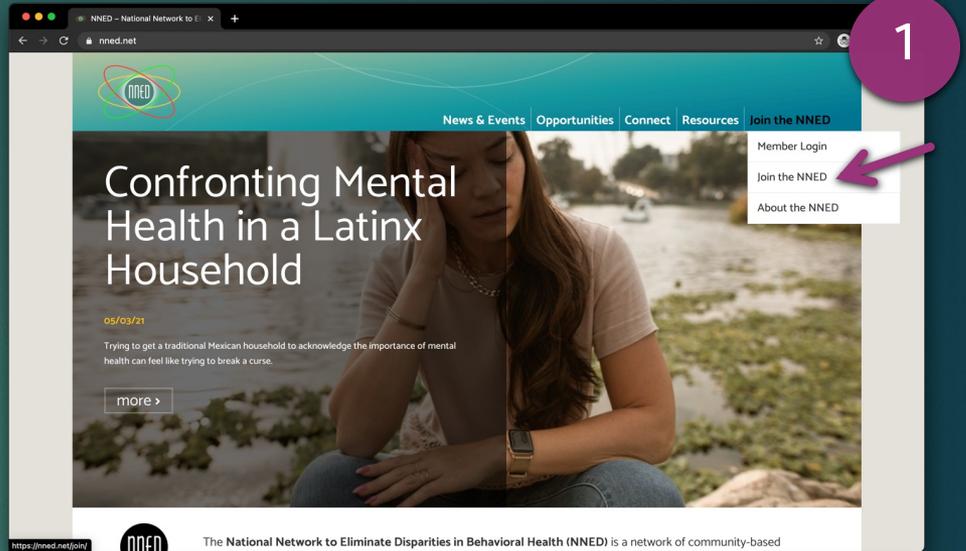


WRAP UP & RESOURCES

VISIT NNEDSHARE:
[SHARE.NNED.NET](https://share.nned.net)



JOIN THE NNED:
[NNED.NET/JOIN](https://nned.net/join)



NNED SPEAKERS BUREAU:

[NNED.NET/SPEAKERS-BUREAU](https://nned.net/speakers-bureau/)

The screenshot shows a web browser window with the URL nned.net/speakers-bureau/. The page features a teal header with the NNED logo and navigation links: News & Events, Opportunities, Connect, Resources, and Join the NNED. The main heading is "NNED Speakers Bureau". Below this is a dark blue section containing the "NNED Speakers Bureau Interest Form". The form includes the NNED logo and the text: "National Network to Eliminate Disparities in Behavioral Health". The form text reads: "The National Network to Eliminate Disparities in Behavioral Health (the NNED), in collaboration with the Substance Abuse Mental Health Services Administration (SAMHSA) is creating a Speakers Bureau of subject matter experts in behavioral health and disparities who represent diverse backgrounds, expertise, and lived experience. The Speakers Bureau will expedite identification of speakers for NNED Virtual Roundtables, SAMHSA events, and may serve as a resource to additional partners, such as the SAMHSA Technology Transfer Centers and Centers of Excellence, for their activities. If you are interested in being a part of the Speakers Bureau, please fill out the form below." The form fields are: "First Name *", "Last Name(s) *", "Pronouns (E.g., They/Them/Theirs, She/Her/Hers, Zie/Zim/Zir, etc.)", "Email *", and "Organization (if applicable)".

News & Events | Opportunities | Connect | Resources | Join the NNED

NNED Speakers Bureau

 National Network to Eliminate Disparities in Behavioral Health

NNED Speakers Bureau Interest Form

The National Network to Eliminate Disparities in Behavioral Health (the NNED), in collaboration with the Substance Abuse Mental Health Services Administration (SAMHSA) is creating a Speakers Bureau of subject matter experts in behavioral health and disparities who represent diverse backgrounds, expertise, and lived experience.

The Speakers Bureau will expedite identification of speakers for NNED Virtual Roundtables, SAMHSA events, and may serve as a resource to additional partners, such as the SAMHSA Technology Transfer Centers and Centers of Excellence, for their activities.

If you are interested in being a part of the Speakers Bureau, please fill out the form below.

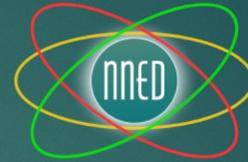
First Name *

Last Name(s) *
If you have post-nominals that you would like to include, please add them here.

Pronouns
(E.g., They/Them/Theirs, She/Her/Hers, Zie/Zim/Zir, etc.)

Email *

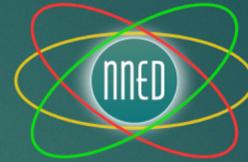
Organization (if applicable)



**Thank you for joining us!
We hope you'll join us for the rest of
the Elevate CBOs Webinar Series.**

QUESTIONS?

**Email the Office of Behavioral Health Equity:
perry.chan@samhsa.hhs.gov**



Please provide feedback on the
Data 101 Workshop:

bit.ly/Data101Feedback