

**TRAINING PACKAGE FOR USING SOCIAL SCIENCE IN COMMUNITY ENGAGEMENT AND/OR COMMUNICATIONS ACTIVITIES**

**SESSION 4.5:** Quantitative data analysis in operational social science research

SESSION CONTENT

**Learning approach:** Real-time presentation, individual and group exercises, case examples

**Delivery mode:** Online and offline, 120 minutes approx.

**Essential sessions to have completed before this session:** 4.3

**Summary:** This session gives an overview of how to apply quantitative approaches to analyse data in operational social science research.

**Learning outcomes:**

* Learn how to analyse data collected with quantitative tools such as Knowledge, Attitudes and Practice/Perceptions (KAP) surveys using descriptive statistics
* Become aware of the inferential statistics that might be helpful to apply to this type of data

FACILITATING THE SESSION



**TRAINING PACKAGE FOR USING SOCIAL SCIENCE IN COMMUNITY ENGAGEMENT AND/OR COMMUNICATIONS ACTIVITIES**

Introduction: (5 minutes total)

Talk through session summary and learning outcomes.

Position this session in the question flow.

1. How to ensure that this information goes back to communities? To inform community-level actions and decision-making of the broader response?
2. What methodology and tools should be used to collect and analyse this information?
3. How to track the information used to ensure that it effectively contributes to operational and strategic priorities?
4. Who can collect this information?
5. Does this information already exist? Is there a related needs assessment or study?
6. What information is needed?

**DATA TO ACTION:**

Key questions in social science research

1. Who needs this information?
2. How to ensure that the information is used to make operational and/or strategic decisions?

Quantitative Research (25 minutes total)

|  |  |
| --- | --- |
|  | Question to participants (5 minutes):  What is quantitative research in social science?  If participants have completed session 4.3 explain that this is a recap question  Online: Invite the participants to write the answers in the chat function and summarize  Offline: Ask two or three participants to share their thoughts |

Qualitative methods are used to understand ‘the why and the how’ – why do people behave in a certain way, how do they perceive the situation and what are their priorities and capacities. These methods are important to get an in-depth understanding of underlying social and behavioural drivers (for example, why do some people not want to be treated in an Ebola treatment centre or avoid being treated there; why do people trust some information sources more than others during a conflict; why do people prefer certain types of COVID-19 vaccines over others, etc.?)

Quantitative approaches

* Work with numbers
* Test theories you have about a situation: e.g. pregnant women are less likely to use water chlorination tablets than non-pregnant women
* Can establish causal relationships – e.g. X causes Y: trust in health system contributes to greater vaccine uptake
* Produce generalizable results – those which are widely applicable – using statistical methods
* The more representative the sample is – when it has characteristics of the whole population including vulnerable and marginalized population groups such as ethnic minorities, people living with disabilities etc., the more likely it can be generalized to a wider population
* Use large sample sizes to get a representative picture of the situation

Quantitative data analysis

Data analysis is the process of systematically applying statistical and/or logical techniques to describe, condense and/or evaluate social science data.

Quantitative data analysis simply means analysing data that is numbers-based or data that can be easily converted into numbers without losing any meaning. It is used to describe the data, to measure differences between groups within the data, to assess relationships between the different things you are measuring in your research, and to test theories (or ‘hypotheses’) we bring to the research – e.g. those with less knowledge on healthy foods will have children with poorer nutritional outcomes – in a rigorous way.

|  |  |
| --- | --- |
|  | Question to participants (5 minutes):  Does anyone have specific experience in running quantitative data analyses? Has anyone analysed quantitative social science data, e.g. from KAP surveys?  Online: Invite the participants to write the answers in the chat function and summarize  Offline: Ask two or three participants to share their experiences |

Descriptive and inferential statistics

There are different methods of data analysis depending on the type of research carried out. Some common methods are: descriptive statistics, diagnostic analysis, predictive analysis, prescriptive analysis, text analysis and inferential statistics.

In this session we will be focusing on two main types of quantitative data analysis:

1. Descriptive statistics
2. Inferential statistics

The choice of using descriptive statistics in research, or a mix of descriptive statistics and inferential statistics depend on the research questions, aims and objectives.

|  |  |
| --- | --- |
|  | Question to participants (5 minutes):  What is the difference between descriptive and inferential statistics?  Online: Invite the participants to write the answers in the chat function and summarize  Offline: Collect ideas from the group |

Descriptive statistics focus on describing the data and the sample of the population we are studying.

Inferential statistics are used to make conclusions from data; specifically, to understand connections between the different factors you are measuring, and to decide whether something could have happened by coincidence or whether the findings are genuine and could be generalizable to the entire population.

We will mainly focus on descriptive statistics. If you are new to analysing quantitative data, then it is better to start here and seek expert input for conducting inferential statistics. It is, however, still important that you understand when and why you might use inferential statistics.

Variables

Before we begin, it is first important to understand what we mean by a ‘variable’.

A variable in research is a person, place, thing or phenomenon that you are trying to measure.

* Categorical variables **>** represent data that can be put into different groups
* Example 1: A survey question on ‘Nationality’ may have 4 categories: 1. Host community, 2. Syrian, 3. Palestinian, and 4. Other
* Example 2: A survey question on COVID-19: ‘What kind of information have you received about Ebola?’ may have 12 categories: 1. Routes of transmission, 2. Symptoms of Ebola, 3. Where to seek treatment, etc. (4–12)
* Continuous variables **>** represent amounts that are measurable in some way
* Examples of continuous variables include: age, weight, height, BMI…
* Example 1: A survey question on COVID-19: ‘Ever since the COVID-19 response was initiated, have you been involved in community actions to address COVID-19?’ can be continuous if the answers range from ‘Never’ to ‘Always’
* Independent variable **>** the one you think might be the cause (as opposed to the effect)
* They are variables you work and directly change in a piece of research
* Example 1: In research looking at the effect of hand-washing training on behaviour change, the training would be the independent variable
* Dependent variables **>** the ones you think might be the effect (as opposed to the cause)
* They are variables that represent the outcome of an experiment/session
* Example 1: In research looking at the effect of hand-washing training on behaviour change, changes in the frequency of hand-washing would be a dependent variable
* Example 2: Most demographic variables are dependent.

Quantitative data analysis steps (5 minutes total)

There are a number of steps to follow during quantitative data analysis:

1. **Data management –**This involves screening the data (e.g. checking all the data is present in the database and whether there any unusual data points that should be checked), cleaning the data (e.g. removing or correcting data that has been entered by mistake or in error) and entering the data into an appropriate program. Data from a quantitative data set (i.e. KAP surveys), is usually loaded into a program such as Excel or the statistics software SPSS, STATA, R, etc., which enables researchers to quickly create tables and charts in order to examine findings.
2. **Understanding variable types – Variable** types are treated differently during data analysis, so it is important to be able to identify which variables you are dealing with, as described above.
3. **Running descriptive statistics –**Descriptive statistics are used to summarize the basic features of a data set through measures of central tendency–
4. **E**.g. what is the middle value of the data set, using mean (average), mode, and median dispersion
5. **E.g**. how far is the data spread out, using range, quartiles, variance, standard deviation, and distribution
6. e.g. what is the frequency of the different values, using skewness
7. **Running appropriate inferential statistics –** Inferential statistics allow researchers to draw conclusions that extend beyond the immediate data. They are used to understand *connections between variables*, and to decide *whether something could have happened by coincidence* or whether *the findings are genuine* and could be *generalizable to the entire population.* For example, inferential statistics are used to determine if a sample represents the population, if there are differences between two or more groups of people, if there are changes over time, or if there is a relationship between two or more variables.

Often, people will first view top-level findings using descriptive statistics before using inferential statistics to test theories or hypotheses.

Inferential statistics involves:

1. **Selecting the right statistical test –**This relies on knowing the nature of the variables in the data set, how they are distributed, and the types of question to ask. Using an example from the Lebanese Red Cross RCCE response to COVID-19 following the Port of Beirut explosion of 2020, **Handout 1** provides a detailed example of study questions asked, how the quantitative tool was constructed, how data was collected, and the methods selected for data analysis (e.g. descriptive statistics).
2. Looking for statistical significance – This is generally captured through a ‘p-value’, which assesses the probability that the findings are more than coincidence. The lower the p-value, the more confident researchers can be that findings are genuine.

Descriptive analysis (55 minutes total)

As already discussed, a first step in analysing a data set is usually to view top-level findings using descriptive statistics before performing inferential statistics to test theories or hypotheses we have about the data, e.g. ‘pregnant women are less likely to use water chlorination tablets than non-pregnant women’.

* Descriptive statistics serve a simple but important role in research, which is to describe the basic features of the data in a study. It is the simplest step that researchers can use to draw conclusions from data and generally forms the basis of deeper quantitative analysis steps taken (e.g. inferential statistics).
* Descriptive statistics do not aim to make conclusions or predictions about the entire population; they are only focused on the specific sample.

|  |  |
| --- | --- |
|  | Brainstorm (5 minutes):  What are some of the most commonly used descriptive statistics?  Online: Invite the participants to write the answers in the chat function and summarize  Offline: Collect answers from the group |

Below are some of the commonly used descriptive statistics:

* Mean: The mean is the numerical average of a range of numbers – calculated by adding the data points together and dividing by the total number of data points.
* Median: The median is the middle value available in the data (midpoint). It helps the researchers estimate where the midpoint of the data is. The data first needs to be sorted to find the median from it. If the data set makes up an odd number, then the median is the number right in the middle of the set. If the data set makes up an even number, then the median is the midpoint between the two middle numbers.
* Mode:The mode is the most commonly occurring number in the data set.
* Range: The range of a distribution is the difference between the lowest and highest value.
* Frequency: The frequency is the number of times a value is found.
* Percentage: To express a number as part of a whole – e.g., how many respondents relative to the whole group of respondents gave certain answers.
* Standard deviation: Standard deviation indicates how widely spread out a range of numbers is, meaning how close all the numbers are to the mean (the average). In cases where most of the numbers are quite close to the average, the standard deviation will be relatively low. By contrast, in cases where the numbers are scattered all over the place, the standard deviation will be relatively high.
* Skewness: The skewness indicates how symmetrical a range of numbers is. In other words, do they tend to cluster in the middle, or do they skew to one end or the other?

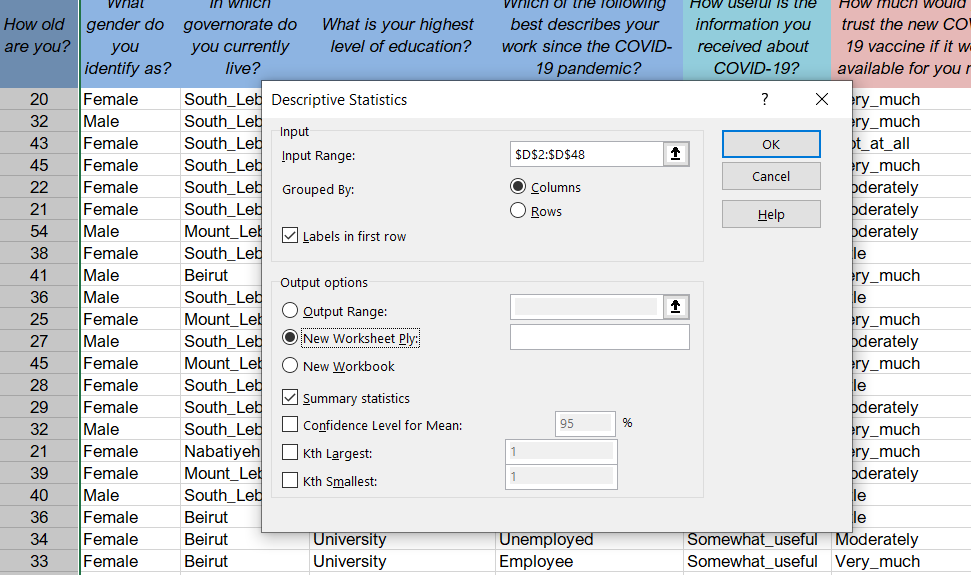
We will now explore descriptive statistical analysis using Excel while providing quick steps on how to conduct it using the [SPSS](https://www.ibm.com/products/spss-statistics) software. For this purpose, we will be referring below to a sample of the database of a study on the perceptions of risk communication and community engagement for COVID-19 implemented in Lebanon as requested by the International Federation of the Red Cross (IFRC) and the Lebanese Red Cross (LRC). Note that the data in this sample has been edited for confidentiality purposes. See **Handout 1** for additional details.

**Data set for the facilitator to run the example ‘live’**

Download **Worksheet 1**. In Excel go to Data **>>>** Data Analysis. If you do not see "Data Analysis" option you need to install it, go to File **>>>** Options **>>>** Add-Ins, and click on “Manage Excel Add-Ins” and select “Analysis ToolPack", then press OK.

Then run Descriptive Analysis as follows:

Data **>>>** Data Analysis **>>>** Descriptive Statistics **>>>** select “Age” for example and select “Summary statistics”.

Figure 1: Summary Statistics

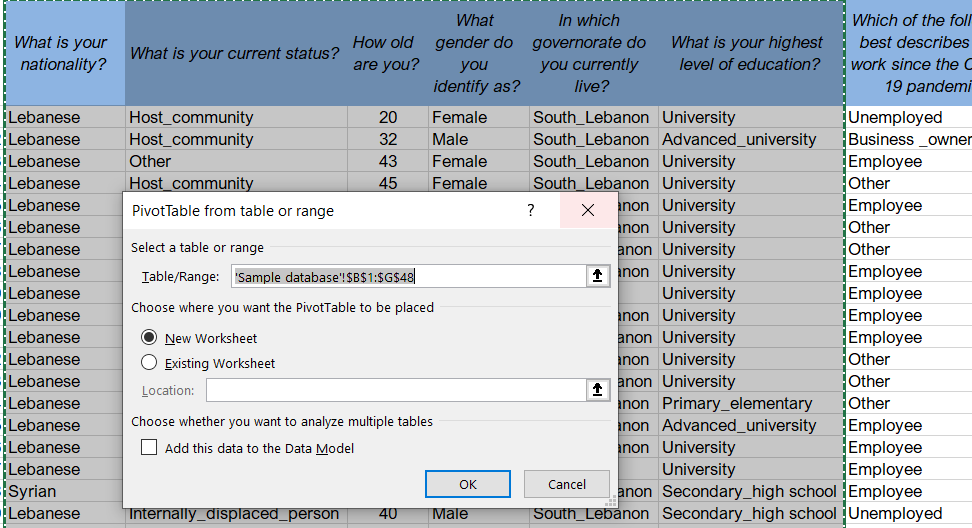
You will get the following table:

Figure 2: Summary Statistics (continued)

|  |  |
| --- | --- |
| Age | |
| Mean | 38.67 |
| Median | 39 |
| Mode | 32 |
| Standard Deviation | 10.58 |
| Skewness | 0.20 |
| Range | 45 |
| Minimum | 19 |
| Maximum | 64 |
| Sum | 1779 |
| Count | 46 |

After looking at the descriptive statistics, you can use cross-tabulations to dig deeper into a data set. Cross-tabulations or crosstabs are useful to better understand the specifics of a data set and the factors that may influence a data point, for example gender and/or education level. It is a way to show the relationship between two variables and is often used to compare results by demographic groups. You can create them on Excel as follows:

Figure 3: Pivot tables



Insert **>>>** Pivot table and select the range of data you would like to analyse.

We will look at ‘nationality’ as an example. Drag and drop ‘what is your nationality?’ under ‘values’ in the pivot tables fields and you will get the frequency. To add the percentages, you need to drag and drop again ‘what is your nationality’ under ‘values’, right click on ‘count of what is your nationality?’ **>>>** value field settings **>>>** show values as **>>>** choose ‘% of column total’ instead of ‘no calculation’.

Figure 4: Pivot tables (continued)

|  |  |  |
| --- | --- | --- |
| *What is your nationality?* | | |
| Row Labels | Frequency | Percentage |
| Lebanese | 44 | 94% |
| Palestinian | 1 | 2% |
| Syrian | 2 | 4% |
| Grand Total | 47 | 100% |

Crosstabs can also be created to examine one data point against another. For instance, you can have a cross-tabulation between nationality and gender by selecting ‘What gender do you identify as?’ in the PivotTable Fields.

Figure 5: Crosstabs

|  |  |  |
| --- | --- | --- |
| *What is your nationality?* | | |
| Row Labels | Frequency | Percentage |
| Lebanese | 44 | 94% |
| Female | 29 | 62% |
| Male | 15 | 32% |
| Palestinian | 1 | 2% |
| Female | 1 | 2% |
| Syrian | 2 | 4% |
| Female | 2 | 4% |
| Grand Total | 47 | 100% |

You can also add other variables, such as ‘highest level of education’ and get the detailed table below:

Figure 6: Crosstabs (continued)

|  |  |  |  |
| --- | --- | --- | --- |
| *What is your nationality?* | | | |
| Row Labels | Frequency | Percentage | |
| Lebanese | 44 | | 94% |
| Female | 29 | | 62% |
| Advanced\_university | 5 | | 11% |
| Primary\_elementary | 1 | | 2% |
| Secondary\_high school | 4 | | 9% |
| University | 19 | | 40% |
| Male | 15 | | 32% |
| Advanced\_university | 4 | | 9% |
| Secondary\_high school | 3 | | 6% |
| University | 8 | | 17% |
| Palestinian | 1 | | 2% |
| Female | 1 | | 2% |
| University | 1 | | 2% |
| Syrian | 2 | | 4% |
| Female | 2 | | 4% |
| Secondary\_high school | 2 | | 4% |
| Grand Total | 47 | | 100% |

It may also be useful to show data in a diagram within your results, and charts and plots can be produced very easily on Excel as per the below examples provided for nationality (bar chart) and gender (pie chart):

Figure 7: Bar chart

Figure 8: Pie chart

In order to engage a variety of stakeholders when presenting the data, the representation of data should be kept simple. It is important to choose the right chart/graph for your data. Avoid using complicated or misleading graphs and make sure you focus on key study questions with clear visualization to strengthen the message.

Some suggestions for charts and graphs are provided in the table below:

|  |  |
| --- | --- |
| Data to present | Chart to use |
| Comparison of two or more categories | Bar chart, doughnut |
| Binary data (i.e., yes/no responses) | Pie chart, bar chart |
| Change over time | Line graph, area graph |
| Correlation between two variables | Scatter graph |
| Frequency | Bar chart |
| Percentages | Bar chart, pie chart |
| Dispersion (how spread out your data set is) | Box and whisker plot, bar chart |

In SPSS, the steps to follow to conduct descriptive analysis are as follows:

* For categorical variables use: Analyze **>>>** Descriptive Statistics **>>>** Frequencies **>>>** Choose the variable
* For continuous variables use: Analyze **>>>** Descriptive Statistics **>>>** Descriptive **>>>** Choose the variable

Refer back to discussion on variables above, if necessary.

*To run quantitative data on SPSS you would need the input of someone with experience using this program*

|  |  |
| --- | --- |
|  | Individual exercise (15 minutes)  Download **Worksheet 1** and run descriptive statistics to identify the mean, median and mode. Then cross-tabulate nationality and gender. You can also practise running other cross-tabulations and tests using the database.  Online: Invite the participants to share the findings in the chat function and summarize  Offline: Collect answers from the group and discuss |

*To conduct inferential analysis of quantitative data you will need the input of someone with in-depth experience in this type of data analysis*

Inferential analysis (15 minutes total)

We will briefly talk about inferential data analysis, so you can be familiar with the different approaches.

Inferential analysis shows the relationships between multiple variables to generalize results and make conclusions/predictions about the entire population. These tests do not give absolute answers, they just indicate how likely it is that a relationship or a difference exists. It is very important to note that if the sample does not accurately represent the population being researched, then findings will not necessarily be very useful beyond telling you about that group of participants.

Bivariate analysis can compare two variables to study their relationships, while multivariate analysis can compare more than two variables.

The most common conclusions/predictions researchers try to make using inferential statistics include:

* Predictions about differences between groups – for example, boys are taller than girls – looking at height differences between children grouped by their gender.
* Relationships between variables – for example, women have less access to COVID-19 information – looking at the relationship between gender and access to information about COVID-19.

|  |  |
| --- | --- |
|  | Question to participants (5 minutes):  Which questions or issues might these different analyses help you to understand in your current work?  Online: Invite the participants to write the answers in the chat function and summarize  Offline: Collect two or three answers |

* Regressions look at the extent to which changes in a predictor variable (gender) result in changes in outcome variable(s) (access to information about COVID-19).
* Correlation describes the statistical relationship between two variables. It measures how far they are related and whether one causes the other or not.

For example, the age and height of a person are highly correlated. If the age of a person increases, height is also likely to increase. This is called a positive correlation. A negative correlation means that upon increasing one variable, the other one decreases. An example of this would be height above sea level and temperature.

The most common statistical tests are listed in **Handout 2**. Depending on your research question and the data you have, you might decide to carry out just one test, or several different tests as part of your data analysis.

*Additional information on inferential analysis and a resource list is available at the end of the session*

Presenting findings and discussion (10 minutes total)

The findings of the study should be written in a succinct and precise format. Graphs, tables, charts and other non-text elements can help the reader understand the data, especially complex findings. We have already discussed which might be appropriate for which types of findings. Make sure that they do not stand in isolation from the text but are being used to supplement the overall description of the results and to help clarify key points being made.

A discussion section of a report using quantitative data should try to provide the following information:

* Interpretation of results – State the research problem being investigated and compare and contrast the findings with the research questions – e.g., *‘The study set out to understand how information on COVID-19 was perceived and used by community members. The majority of survey participants thought the information received was applicable and realistic and 99% stated that they used the information, either by applying prevention measures, monitoring their health and identifying symptoms, sharing messages with others, or managing COVID-19 cases.’*
* Description of findings – Present comparison of groups, or relationships among variables – e.g., explain all unanticipated (surprising) and statistically insignificant findings, which are still important. For example, *‘We had expected to see a significant positive correlation between access to a health facility and knowledge of COVID-19, but no significant positive correlation was found.*’
* Discussion of implications – Highlight key findings, noting findings that you believe are especially important, and what these mean for practical, policy, and/or advancing knowledge in this area – e.g., *‘These findings suggest that the organization should continue to disseminate information in this way.’*
* Limitations – Describe any limitations or unavoidable bias (see Session 4.3 for discussion of bias) in your study. Examples of limitations might include low sample sizes due to budgetary and time constraints, the possible exclusion of people with no access to phones or the internet if the survey uses one of these formats, etc.

The conclusion section should include a summary of the main findings and actions to take forward. Tie key findings to any programme and/or policy recommendation. If appropriate, provide recommendations for future research based on the study’s limitations and any gaps that were identified in the course of the study.

Your recommendations are likely to be one of the most important parts of your report. Good recommendations will mean that your study findings are more likely to be used.

Specific guidance on transforming data to actionable findings and recommendations is covered in Session 5.2. Specific guidance on communicating and feeding back findings is covered in Session 5.3.

Additional articles and links for quantitative data analysis using Excel and other statistical software

* [Quantitative Analysis of Questionnaires: Techniques to Explore Structures and Relationships](https://www.routledge.com/Quantitative-Analysis-of-Questionnaires-Techniques-to-Explore-Structures/Humble/p/book/9780367022839)
* [A Quick Guide to Quantitative Research in the Social Sciences](https://repository.uwtsd.ac.uk/id/eprint/1218/18/A%20quick%20guide%20to%20quantitative%20research%20in%20the%20social%20sciences.pdf)
* [Selection of Appropriate Statistical Methods for Data Analysis](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6639881/)
* [Using Excel for Statistical Data Analysis – Caveats](https://people.umass.edu/evagold/excel.html)
* [Excel Easy: Data Analysis](https://www.excel-easy.com/data-analysis.html)
* [Choosing the Correct Statistical Test in SAS, STATA, SPSS and R](https://stats.oarc.ucla.edu/other/mult-pkg/whatstat/)
* SPSS Tutorials [here](https://www.spss-tutorials.com/) and [here](https://www.spss-tutorials.com/which-statistical-test-should-i-use/)
* [SPSS Tutorial for Data Analysis: SPSS for Beginners](https://www.youtube.com/watch?v=Bku1p481z80) (YouTube)
* [Learn How to use SPSS Statistics to Analyse your Data](https://statistics.laerd.com/features-spss-procedure.php)

Wrap-up/summary (5 minutes)

* Quantitative data analysis is used to describe the data, to measure differences between groups within the data, to assess relationships between the different things you are measuring in your research, and to test theories (or ‘hypotheses’) we bring to the research.
* Descriptive statistics focus on describing the data and the sample of the population we are studying.
* Inferential statistics are used to make conclusions from data; specifically, to understand connections between the different factors you are measuring, and to decide whether something could have happened by coincidence or whether the findings are genuine and could be generalizable to the entire population.
* The most frequently used descriptive statistics are the mean, median, mode, range, frequency, percentage standard deviation and skewness. These can be analysed through Excel, SPSS or other software (e.g. SAS, STATA, R).
* The most common conclusions/predictions researchers try to make using inferential statistics include predictions about differences between groups and relationships between variables.
* Regressions look at the extent to which changes in a predictor variable (e.g. gender) result in changes in outcome variable(s) (e.g. access to information about COVID-19). Correlation describes the statistical relationship between two variables, measuring how far they are related and whether one causes the other   
  or not.

ACKNOWLEDGEMENTS:

Soha Karem and Barbara Muzzolini (Anthrologica) developed the session content. It was reviewed by Theresa Jones and Eva Niederberger (Anthrologica) and Ginger Johnson (Collective Service).