IMPACT Data Analysis Guidelines for Probability Samples

Annexes

Introduction

This document gives the minimum standards and guidelines to be applied across all IMPACT research using probability samples. Its aim is to provide clear guidance on the process and the requirements for preparation and validation of data analysis, including the formal tools required at each step. The guidelines refer to the appropriate techniques for quantitative data analysis as per academic research standards. The focus of this document is on identifying **what** statistics are appropriate in what type of case and how to interpret the results. For each case, it specifies what needs to be done and why. It does not specify how it is done with a specific software, nor does it function as a handbook on how to carry out statistical tests. The guidance should be consulted together with the Research Design, Data Cleaning, Qualitative Data Analysis Guidelines and the Reporting Guidelines.

How to use this document

These guidelines are structured to first establish an understanding of the application of IMPACT‘s analysis minimum standards and processes, and second to enable quick access to specific cases of analysis in a step by step guide.

Chapter 1 outlines the requirements that must be fulfilled in all analysis done in all IMPACT probability samples analysis (For non-probability samples please refer to the qualitative data analysis guidelines).

Key tools
- Table 1: Minimum standards
- Annex 1: External resources for general statistics

Key outputs
- Annotated checklist (Annex 2) shared to the Data Unit in HQ for review and validation

Chapter 2 explains the process guiding the analysis, from preparation over execution to the interpretation of the statistical results.. Each step below is introduced with the relevant annexes and support materials.

Step 2.1. is the first step for any assessment using a probability sample. It includes verifying your population of interest and sampling strategy, and defining your hypothesis and analysis parameters. At the research design stage, most parameters for your analysis (including the sampling strategy, population of interest, and what indicators need to be analysed) should be clarified already, but often additional details need to be added prior to analysis. After data collection, the outline for your analysis will correspond to the lines in the extended data analysis plan.

Key tools
- Annex 3: Analysis preparation file
- Annex 4: Key Definitions
- Data Analysis Plan Guidance

Key outputs
- Filled out Analysis preparation file (Annex 3) shared with Data Unit in HQ *before the beginning of analysis.*

Step 2.2. explains **how to identify** the case for each indicator. Each indicator you intend to report on each case will correspond to one line in the data analysis plan. To find the case, you need to identify the following, if not already clarified in the ToR and the extended data analysis plan:

a. Hypothesis type: one of: **direct reporting**, **group difference**, **change**, **correlation,** or **limit**

b. Number and types of variables: one or more of: **numerical**, **categorical.**

Key tools

- Tables 2a and 2b in step 2.2
- Flowchart in Graph 1 for step 2.2
- R hypegrammaR mapping functions

Key outputs

- Filled out Analysis preparation file (Annex 3) shared with Data Unit in HQ *before the beginning of analysis.*

Step 2.3. is the step by step guide to find out exactly what you need to do for each analysis case. Based on the the Hypothesis type and the number and type of variables you identified the case in the previous section. There are 9 main cases that cover most of the types of analysis currently done in IMPACT assessments. The details of their implementation depend on the sampling strategy:

i. Case 1: Direct reporting for a numerical variable

ii. Case 2: Direct reporting for a categorical variable

iii. Case 3: Limit

iv. Case 4: Change in an indicator (for discrete time points)

v. Case 5: Group difference for a numerical variable

vi. Case 6: Group difference for a categorical variable (correlation of two categorical variables)

vii. Case 7: Correlation of two numerical variables

viii. Case 8: Correlation of a categorical variable with numerical variable(s)

ix. Case 9: Correlation of >2 variables.

Key tools

- R hypegrammaR tool
- Annex 5: Resources for statistical tests

Key outputs

- Analysis output for each line in the extended data analysis plan

Step 2.4. Interpreting results provides guidance on reporting, letting you tie the results of your analysis back to the research question.

Key tools

- Reporting guidelines
- Research cycle guidelines

Key outputs

- Draft report of results shared with the analysis.

1 Minimum Standards for Probability Sample Data Analysis

The requirements below must be fulfilled in all analysis done in all IMPACT probability samples analysis. For non-probability samples please refer to the qualitative data analysis. Should one of the standards not be achievable, the person in charge of the analysis must request an exception in advance, which will be granted in writing. Following the completion of the entire analysis, the annotated checklist in Annex 2 is shared to the Data Unit in HQ for review and validation.

Minimum standard	Current status (January 2019)
General	
Data and cleaning must be validated according to minimum standards for data cleaning	Always done
Appropriate aggregation method must be used (mean, median etc.)	Always done
Calculations must be correct and implemented according to Research Design	Always done
Analysis Plan / Research design While this is part of the research design, the requirements are necessary in order to perform meaningful analysis[1][2][3][4]	
Hypotheses must be formulated to best serve answering (sub)research questions	Sometimes done
Analysis that does not serve to directly answer a (sub)research question must be minimised. There must be a well defined prior hypothesis associated with each performed analysis. (Exceptions can be made for exploratory analysis if and only if the “exploratory analysis” requirements below are respected)	Sometimes done
Analysis that will not be directly or indirectly reported on must be minimised; All hypotheses that were tested must also be reported on (exceptions can be made for exploratory analysis if and only if the “exploratory analysis” requirements below are respected)	Sometimes done
Independent variable(s) must be used only as specified in the data analysis plan (exceptions can be made for exploratory analysis only if the “exploratory analysis” requirements below are fulfilled)	Sometimes done
Comparison between groups must be made only as specified in the data analysis plan and as directly relevant to the research question (exceptions can be made for exploratory analysis only if the “exploratory analysis” requirements below are fulfilled)	Sometimes done

[1] Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. PLoS biology, 13(3), e1002106.

[2] Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. Perspectives on Psychological Science, 9(5), 552-555.

[3] Chambers, C. (2014). Psychology’s ‘Registration Revolution.’. The Guardian.

[4] Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic medicine, 66(3), 411-421.

Unstructured exploratory analysis results must not be presented as statistically valid. There is a fundamental difference between *exploratory analysis* which *generates* hypothesis, and validatory analysis which *quantifies and tests hypothesis.* Doing both on the same data jeopardises your research as a whole unless major precautions are taken. [5,6,7,8,9,10,11] Exploratory Analysis can be performed and reported on only in the cases below: A) All Exploratory results are labeled as such and declared as speculative; they must be clearly separated from validatory analysis in the process as well as in the outputs. B) Alternatively, advanced statistical methods to mitigate alpha inflation and overfitting are applied and carefully executed throughout the process. Secondary effects of the mitigation measures (e.g. type II inflation) are transparently communicated.	Rarely done
Representativeness & Generalisability	
Unequal sampling probabilities must be corrected with weights (depending on sampling strategy; whenever samples with unequal sampling probability are aggregated together)[12]	Always done
Weights used for stratified probability sampling data must be adjusted for effective per-indicator sample size (depending on the question and skip-logic, the actual sample may deviate from the overall sample of the assessment).	sometimes
Specifying certainty	
General certainty must be declared in methodology / limitation section (e.g. findings are indicative; findings are generalisable with 95% level of confidence / 5% margin of error)[13]	Always done
Calculated certainty must be declared: Confidence intervals[14] **(error bars for graphs) and p-values**[15] **(hypothesis tests) reported.**	Done where capacity
Certainty calculations must be adjusted for sampling strategy. [16] Error margins, Confidence intervals and Hypothesis tests must make the Rao-Scott adjustments for probability stratified sampling and for Cluster probability sampling	Done where capacity
Hypothesis tests must be adjusted for multiple testing[17] If multiple hypothesis tests are deployed, they must be corrected with the False discovery rate or the Bonferroni correction	Rarely done
Comparability	

[5] Ioannidis, JPA (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124.

[6] Anderson, DR, Link, WA, Johnson, DH, and Burnham, KP (2001). Suggestions for Presenting the Results of Data Analysis. The Journal of Wildlife Management, 65(3)

[7] Michels, KB and Rosner, BA (1996). Data trawling: to fish or not to fish. Lancet, 348, 1152-1153.

[8] Lord, SJ, Gebski, VJ, and Keech, AC (2004). Multiple analyses in clinical trials: sound science or data dredging?. The Medical Journal of Australia, 181(8), 452-454.

[9] Smith, GD and Ebrahim, S (2002). Data dredging, bias, or confounding. BMJ, 325, 1437-1438.

[10] Afshartous, D and Wolf, M (2007). Avoiding 'data snooping' in multilevel and mixed effects models. Journal of the Royal Statistical Society A, 170(4), 1035–1059

[11] Anderson, DR, Burnham, KP, Gould, WR, and Cherry, S (2001). Concerns about finding effects that are actually spurious. Widlife Society Bulletin, 29(1), 311-316.

[12] Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, *9*(1), 1-19.

[13] Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and psychological measurement*, *30*(3), 607-610.

[14] Newcombe, R. G. (2012). *Confidence intervals for proportions and related measures of effect size*. CRC Press.

[15] Shi, N. Z., & Tao, J. (2008). *Statistical hypothesis testing: theory and methods*. World Scientific Publishing Company.

[16] Rao, J. N. K., & Scott, A. J. (1987). On simple adjustments to chi-square tests with sample survey data. *The Annals of Statistics*, 385-397.

[17] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.

Comparisons over time must only be done on comparable populations of interest (or advanced analysis for bias mitigation / estimation)	Always done
Comparisons between groups/locations must only be within comparable time frames (or advanced analysis for bias mitigation / estimation)	Always done
Implementation: Transparency, reproducibility and reusability	
All steps of the analysis must be well documented, explained and easily understandable.[18]	Sometimes done
Separation of concerns must be applied[19] Data, parameters, calculations and visualisations should be separated (in different sheets/files/sections)	Sometimes done
Calculations must be data agnostic[20] Formulas / calculations must be as unspecific to the data as possible (with specific input parameters defined separately, see ‘separation of concerns’)	Sometimes done
DRY principle: “Don’t repeat yourself” must be applied Analysis must be done with as little repetition as possible (for example avoiding many pivot tables where the same can be achieved with a single one; using functions instead of copy pasting code. Ideally each type of logic (formulas, pivot tables, data transformations,...) should be defined in a single place and reused[21]	Sometimes done

Table 1: Minimum standards for analysis of probability samples

[18] Stodden, V., Leisch, F., & Peng, R. D. (Eds.). (2014). *Implementing reproducible research*. CRC Press.

[19] Hürsch, W. L., & Lopes, C. V. (1995). Separation of concerns.

[20] Baldwin, C. Y., & Clark, K. B. (2000). *Design rules: The power of modularity* (Vol. 1). MIT press.

[21] Hunt, A., Cunningham, W., & Thomas, D. (1999). The pragmatic programmer: from journeyman to master.

2 Step by Step Guide

The steps below define the process for completing quantitative data analysis as part of the IMPACT research cycle. They should be carried out as early as possible, but latest after the data has been cleaned according to the Data Cleaning Guidelines. The first step (2.1) formalises the verification that all aspects of the assessment that affect the analysis are clarified. It requires sharing the analysis preparation sheet. The second step (2.2) is to identify which type of analysis is needed for each line in the extended data analysis plan. The third step (2.3) is to apply the appropriate analysis for that case. The final step (2.4) is to interpret the results of the analysis, leading over into the Reporting Guidelines.

2.1 Fill out the analysis preparation sheet

The first step in the preparation of your analysis is to review and determine properties of the assessment that affect the analysis. The final analysis parameters need to be cross checked against those defined in the ToR when filling out the analysis preparation sheet. This ensures that all properties of the analysis have been considered and specifies what questions the analysis aims to answer.

The data analysis plan and the ToR should contain all the information you need to get started:

- what research question each variable was collected for
- what the population of interest is
- what sampling strategy was used
- what type of hypothesis each indicator falls in (see 4.2.1. for details on hypothesis types)
- what independent variable(s) is/are relevant to answer the research question
- whether the research questions should be answered for multiple subsets of the data
- what data type each indicator corresponds to (see 4.2.2. for details on data types)

The cross check of these elements is achieved in the following three steps.

2.1.1 Clarify representativeness and the population of interest

The first step in determining the parameters of an assessment is to verify what exactly the *population of interest* (POI) is. To formalise the cross-check, the assessment summary tab in the analysis preparation file needs to be filled out with the results of step 1 to 3 below. Steps 4 and 5 verify that coverage and bias are mitigated where possible and always reported.

i. Verify who could have been in your sample and enter the information into the assessment summary tab:
    - General:
        - **POI:** Who was in your sampling frame: the initial list of individuals / households etc. you randomly picked from.

    - **Not POI:** Who could not have been sampled (e.g. no access, no phones etc.)?: not in the population of interest.
  - Stratified probability sampling
    - **POI**: The sampling frame - the people in the list of strata you pick your sample from - make up the entire population of interest
    - **Not POI:** Any strata that are excluded, e.g. area could not be reached, no sample could have been drawn from them, or no households could be assessed.
  - Stratified cluster or cluster sampling
    - **POI**: The sampling frame - the people in the list of All the clusters that you pick your locations ore individual samples from - make up the entire population of interest
    - **Not POI:** Anyone outside the list that you selected your clusters from e.g. when assessing four cities selected from the 16 cities in a region as part of you cluster sampling strategy: HHs in all cities are part of your PoI, but the rural population is not part of your PoI.

ii. Check the basic unit in your population of interest and enter the information into the assessment summary tab
  - Make sure the unit of analysis is clear and consistent (% of households is not the same as % of the population or % of individuals).

iii. Make an educated guess about bias directions (e.g. no access often worsens the situation) and enter the information into the assessment summary tab.
  - Whether your sample is **representative** of the population of interest depends on whether it was obtained without bias from that population.
    - Often, a sampling bias can be introduced inadvertently (e.g. in a phone survey, missing all the respondents who did not answer their phone).
  - Adjust your population of interest based on the sample you obtained so that you are able to confidently generalize to that group.

iv. See if you can mitigate/triangulate against biases
  - E.g. In a report on migrant children: those with the worst experiences could not be interviewed. To fill information gaps, the team decided to do FGD's with the accommodation center staff.

v. Clearly report population of interest (methodology text; maybe coverage map), who is included and who is excluded
  - Report the direction of biases
  - ensure reporting wording according to representativeness (see reporting guidelines)

Requirement for completion:

- Assessment summary tab must be filled out in the analysis preparation file.
- Coverage must be included in the report / factsheet / readMe tab.

2.1.2 Verify the sampling strategy and its implications

a) Identifying the sampling strategy

The second step in determining the parameters of an assessment is to formalise how the sample was obtained from the population of interest, i.e. what sampling strategy was used. To formalise the cross-check, the assessment summary tab in the analysis preparation file needs to contain the name of your sampling strategy as defined by Table 3 and the chapter below. The strata sampling frame and cluster sampling frame tabs need to be filled out where applicable. Your sampling strategy for a given population of interest is a combination of any methods and attributes below, which aim to capture information about the population of interest[22].

- **Check against impossible sampling strategies:**
    - If you did not use a probability sampling approach, it cannot be **stratified or cluster** sampling.
    - If you conducted a **census**, it by definition excludes all other sampling strategy options.
    - Your sample can not be **probability** and **purposive** at the same time
- **Check if complementary methods were used:**
    - An assessment may employ different sampling strategies for different population groups or geographical areas. For example, you may use stratified probability sampling of Households by district for most of the areas in the country, and purposive sampling of Key Informants for the hard to reach areas.
    - In that case, the analysis of the two separate samples needs to be done independently of each other. Treat them like two different assessments, all the way until results are produced. (It may be possible to combine them depending on the situation; this is complicated and should only be attempted if the implications are clear and the technical capacity is available in the team)

The table below summarises the necessary and sufficient conditions that define the different sampling strategies.

Condition	Examples	Case
Every individual in the population of interest has been assessed.	Every household in a camp was interviewed All infrastructure items in a defined area were assessed.	**census**
Every individual in the population of interest had the same probability to be sampled. We assume that no factors led us to over- or undersample a particular group. This category includes systematic random sampling	Households were selected randomly from a list of beneficiaries. Every fifth household was interviewed in a camp that is organised in a regular grid.	**simple probability**

[22] If you are using the hypegrammaR tool in R, that information will need to be entered into the "parameters" tab. The implications below are then handled automatically (weights and design effect)

Every individual in the population of interest had a clearly defined and known probability to be sampled. The population is split into groups (“strata”). Inside each group (“stratum”) with simple probability sampling was followed The population size in each stratum is known.	Within multiple separate population groups, 100 Households were randomly sampled	**stratified probability**
Every individual in the population of interest had a clearly defined and known probability to be sampled. The sample is drawn from a limited number of randomly selected locations. The population is divided into clusters(e.g. locations). These clusters are selected via simple probability sampling. . Individuals are randomly sampled in each of the selected clusters	Over a country, 20 different locations were selected. In each of the selected places, a random sample was drawn.	**cluster probability**
Like “stratified probability”, but within each stratum, “Cluster probability” is used rather than simple probability. Both “Cluster probability” and “stratified probability” guidelines must be applied	A nationwide sample is drawn, stratified by governorate. Within each governorate, a limited number of locations is randomly selected, and from these a simple probability sample is drawn in each.	**stratified cluster probability**
simple probability, Cluster probability or stratified cluster sampling are combined in any way other than listed above	A stratified probability cluster sample is drawn nationwide, except for two governorates that had a simple probability approach within.	**mixed probability**
The sample is taken from a group of the population that is well suited to answer our question	2 key informants: a nurse and a teacher, were selected for each district because they have specific sector knowledge. Respondents were selected because they have recently been or live in a hard to reach area.	**non-probability**

Table 3: Sampling strategy types

b) Verifying the sampling strategy implications

After identifying the sampling strategy in the previous step, the specific implications of the strategy must be considered. They are listed below for each case.

Sampling strategy 1: Census

You have full information about your population of interest and there is no uncertainty associated with statistical generalisation of your results (this does not include data cleaning issues and human error). None of the test statistics for significance and certainty are necessary: anything you observe can be stated as a fact about this population. You can skip all “Measuring Certainty” instructions in the subsequent chapters.

Sampling strategy 2: Simple probability

Before assuming this, make sure you are in fact in a case of simple probability sampling and have not used implicit bias, strata or clusters: e.g. if you picked random villages from a list, and then interviewed people at random in each of these villages, you have implicit clusters. A sample is not simple probability if some factor made a person more or less likely to be selected, for example if HH closer to a main road were more likely to be selected. If HH closest to uniformly random GPS points were interviewed but the population density is not uniform, the sampling probabilities for different households are not the same, and the sample should be treated as a stratified probability sample.

No sample is a perfect simple probability sample; there are always known and unknown biases. If biases can be accounted for with weighting, the sample should be treated as a stratified probability sample. If biases can not be accounted for with weighting, but acceptable[23] overall, the sample should be treated as a simple probability sample and the potential biases stated in the methodology limitations.

Sampling strategy 3: Stratified probability sampling

- If you are not aggregating between strata, do the same as for a simple probability sample.
    - Your population of interest is now only the population in each stratum
- If you are aggregating to anything that combines more than one stratum, use post stratification weights to correct for any sampling imbalance.
    - Adjust for nonresponse: your weights change when your sample size changes but your population of interest does not change proportionally.
    - Adjust your certainty measurements to account for stratification and weights (using software described in implementation)

Sampling strategy 4: Cluster probability sampling

- Use the design effect to correct for the greater variance in the population than in your clustered sample.
    - While you can obtain an estimate for the design effect in your sample by using the formula and existing data in the cluster sampling memo (on the resource centre), these estimates are for sample size estimation and should not be used during analysis.

- One stage cluster: after splitting your population of interest into clusters and randomly selecting a subset of clusters, you assess every unit within the selected clusters.
- Two stage cluster: after splitting your population of interest into clusters and randomly selecting a subset of clusters, you assess a random sample within each selected cluster.
- If individuals in different clusters did not have the same probability to be selected, the clusters need to be additionally treated as strata and weighted. The weights should be inversely proportional to each individuals sampling probability.

[23] acceptable in the sense that the results are not misleading; this needs to be decided on a case by case basis

Sampling strategy 5: Stratified cluster probability sampling

Both stratified probability and Cluster probability sampling requirements and implications must be respected. If the clusters are not sampled proportional to size, the sample needs to be weighted twice, once by cluster and once by stratum. To combine the weights, the cluster weights need to be calculated per stratum, and then combined with the strata weights.

Sampling strategy 6: Mixed probability sampling

If sampling strategies are mixed in any way other than the combinations described above, the correct procedure needs to be decided on a case by case basis. This can get quite complicated quickly, and we highly recommend to contact the data unit when facing a mixed probability sampling. However, the generalised approach to get to a solution is as follows:

- Identify parts of the population that may have been excluded from the population of interest.
- Trace the real samping strategy and identify the lowest level of strata/cluster that we have reliable population numbers for.
- Identify the real probability to be selected for all records for each stratum/cluster
- Calculate weights for each record based on the above, normalised so that the sum of weights matches the total number of records.
- Identify potential clusterings, and try to find a single “clustering level” for each stratum. Be careful that meaningful cluster variances can be calculated.
- Get overall weights and cluster ids.
- Follow **stratified probability cluster** instructions using the newly identified strata and cluster

Requirement for completion:

- Sampling frame and cluster sampling fame tabs must be filled out in the analysis preparation file.
- Any deviation from the sampling strategy defined in the ToR must be logged in the deviations tab.

2.1.3 Complete and share the extended data analysis plan

Once the sampling strategy and the population of interest is clear, we can use *the sample* to answer questions about the *entire population of interest* with a measurable level of certainty.
We achieve this by splitting our questions into quantifiable hypotheses, that can then be tested with statistical hypothesis tests. The instructions on filling out the analysis plan below will guide you step by step to gather all relevant information to uniquely identify meaningful hypotheses and collect all the information you need to apply meaningful analysis to them.

Hypothesis Tests & P-Values

When to test

Since hypothesis testing is about **measuring the uncertainty introduced when generalising** from a sample to a larger population, there is no point in testing when we can not generalise with a measurable level of precision (e.g. when using non-probability sampling such as purposive sampling was used), or when we do not need to generalise (census).

Null Hypothesis

At the heart of the scientific method is the idea to **form a hypothesis of what we believe to be true**, and to test it by *rigorously search for evidence against it*, rather than to look for evidence that confirms our prior belief. All (frequentist) statistical tests are therefore performed on the opposite of our initial hypothesis, called the “null hypothesis”.

Measuring uncertainty and interpreting p-values

When we generalise from a sample to a larger population, there will always be a possibility that, by pure chance, we picked a sample that leads us to the wrong conclusion concerning our hypothesis. While we can **not** know whether that is the case, we can calculate the probability of that having happened. That probability is called the p-value, and only if it is small enough (usually <5%) we reject the null hypothesis with the given confidence.

- A small *p*-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis. There is a ≤ 5 % chance that you would have observed what you did if the null hypothesis was true.
- A large *p*-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- *p*-values very close to the cutoff (0.05) are considered to be marginal (could go either way). Always report the *p*-value so your readers can draw their own conclusions.

It is important to also understand what the p-value does **not** do:

- The p-value does **not** tell you anything about the importance of your result
  - p-values should **never** be used to decide whether a result is reported on or not.

- The p-value does **not** tell you whether your null hypothesis **is** true or false: it merely provides you with the probability that you would have observed what you did in the sample, had the null hypothesis been true.
- The p-value does **not** tell you anything about how strong the effect is that you observed, it only tells you about the accuracy of your measurement. See "Effect size vs. Certainty" in the reporting section.
    - With proper sampling, the larger your sample, the more precise your estimations. With large enough samples you can therefore *always* find a significant effect, **no matter how tiny and irrelevant that effect may be.**
    - If you find a small effect with high certainty, that should not be confused with finding an important or large effect.
- Conversely, if you find a large effect with low certainty, that does not diminish the observed size of the effect; It just means that we do not have enough data to make a reliable judgement, and need to report the result more carefully and with less confidence. If, for example, reporting a difference between groups when that is not true in the population would have a harmful effect (like falsely diverting aid), it should be made very clear that we have low confidence in the observed difference.

All statistical tests result in a p-value, which always refers to the probability that your observed sample came from a state where the null hypothesis was true. What statistical test you need to calculate the p-value depends on what type of hypothesis you have as well as on the data type. (instructions on identifying these can be found below).

The final step in preparing your assessment for analysis is to fill out the extended analysis plan, thereby formalising the questions you are hoping each indicator will answer in a hypothesis. This section serves two purposes: to give a background on hypothesis testing in the context of the research cycle, and as a guide in filling out the extended data analysis plan.

Fill out the extended data analysis plan

It is decisive that the data analysis plan be made and up to standard before data collection begins. If that has not been done, you may still use this section to write/correct the data analysis plan accordingly before beginning the analysis. Apart from the research design, the analysis plan may require specific details to help perform good analysis. The sections below will guide you through the basic considerations based on your research questions that you will need in order to correctly choose your hypothesis type, disaggregation level and variable combinations. If your data analysis plan fulfills the minimum standard requirements in chapter 1, it will already contain all of this information. If you are using the hypegrammaR tool in R, the preparation steps are formalised in filling out the input sheet.

1. **Open the extended data analysis plan template where you find/document all results from the following steps**
2. **Identify whether all research questions should be answered for multiple subsets independently, or whether only a subset of the data is relevant for analysis.**
    a. Often we produce individual outputs/pages/chapters for different parts of the data (e.g. for different locations, different settlements etc.)
        i. In this case all analysis will be applied identically and independently for each subset (e.g. for each location or settlement). This does not affect the analysis itself: you prepare your analysis as if you were running it on one subset alone and then repeat for each group.
        ii. Comparative analysis between those subsets (e.g. between locations, between camps) as well as analysis that aggregates those subsets do not fall in this category!
        iii. **Note down by what variable(s) you need to split your data to repeat the analysis for each subset in the "repeat analysis for each.." column in the extended data analysis plan.**.
3. **Identify what hypotheses about the dependent variable answer the sub-research question**
    a. Each indicator that is analysed (each row in the data analysis plan) is the dependent variable in the analysis defined in that row.
    b. What you want to know about the dependent variable is called the hypothesis. A good hypothesis stated as a question asks for
        i. a number (What is the average ...? What percentage…? How many..?)
        ii. a quantifiable yes/no answer (did ... increase? Is … more prevalent in some part of the population? Do individuals with larger X also have larger Y? Is there a difference between …?
    c. You should be able to match this question to one of the hypothesis types in the table in the Hypothesis Type paragraph in section 2.2.1.
        i. If you are unsure which hypothesis your question falls under, contact the data unit.
    d. Note down your hypothesis in the corresponding column in the data analysis framework
4. **Identify the independent variables**
    a. Often your quantifiable question will look at your indicator Y (dependent variable) depending on some other information *X* (independent variable(s)). Note down the name, or the value, of that variable:
        i. group difference: Is there a difference in the average *Y* depending on what group *X* someone belongs to?
            1. E.g. If the quantitative question is: "Are IDPs or Returnees more likely to receive NFI's?", then you analyse if the dependent variable "received NFI" is affected by the independent variable specifying the IDP/returnee population group.
        ii. correlation: Is *Y* larger/smaller depending on the value of *X*?
        iii. limit: Is *Y* larger than a constant value *X*?
    b. If the independent variable is categorical, it is important to have a large enough sample in each category to be able to carry out statistical tests for the different groups. This should be identified ahead of time in the research design stage, and you should potentially stratify by those groups.

Note down the name of the independent variable(s) (if any) in the corresponding column(s) in the data analysis framework.

Requirement for completion:

- Extended analysis plan tab must be filled out and shared with Data Unit at the end of data cleaning.

Distinguishing **exploratory vs. confirmatory** analysis

The reason why the extended data analysis plan must be shared **before** the analysis is performed is because there is a fundamental difference between *exploratory analysis* which *generates* hypotheses, and *validatory analysis* which *quantifies and tests hypotheses.* Doing both on the same data is called “p-hacking”, “data dredging” or most accurately “data torturing” and jeopardises your research as a whole unless major precautions are taken.[5,6,7,8,9,10,11]. At this stage at IMPACT, we usually use qualitative research to form hypotheses, while the quantitative assessments are restricted to hypothesis evaluation. There is a near infinite number of ways each indicator could be analysed - for example, different levels of aggregation in different combinations, or comparisons between population subsets or indicators can easily lead to more aggregate statistics than values in the original data. It almost never makes sense to apply a blanket analysis to all indicators and multiple disaggregations equally[24]. Analysis is only valuable and statistically valid if:

Any decisions about how to analyse each indicator is made to directly serve the goal of answering one or several specific research questions.

- Hypotheses are not created from the same data as they are tested on.
- There is as little unreported analysis as possible.
    - “If you torture the data long enough, it will confess to anything” - If you run a large amount of analyses without a research question clearly in mind, you will always find interesting patterns eventually - whether they exist in the population or not. Looking at a large number of different combinations/aggregations/hypothesis tests destroys the statistical validity of the result in theory, and in practice puts you at risk of reporting on patterns that are pure random noise in your sample.

These requirements may seem very limiting with regards to exploratory analysis, especially to those familiar with data mining who are used to creating and testing hypotheses in an iterative process. IMPACT heavily relies on both processes, but stresses that **within** a single dataset, exploratory analysis requires a whole additional set of research practices and statistical methods to remain rigorous. However, the intention is not to make exploratory analysis outside the data analysis plan more difficult. Exploratory analysis can readily be conducted, but unless advanced workflows and methods are used for mitigation of alpha inflation, overfitting and other risks, all results must be treated as speculative **discovery of potential hypotheses.** Exploratory analysis should be conducted and the results reported as completely separate from our generally validatory research designs.

[24] *In cases like past MSNAs where the goal was to establish the difference among groups for every indicator possible, the answer to your research questions would have to include all of the variables where no significant differences were found. You then should closely follow 2.4* Multiple hypothesis: Interpretation in context

2.2 Identify the type of analysis for each line

Once you have cross checked the parameters of your assessment and shared the analysis preparation file, the next step in the research cycle is to determine which statistical methods need to be applied for each line in the extended data analysis plan. This follows directly from the hypothesis type (3.2.1.) and the data types (3.2.2).

2.2.1 Identify the Hypothesis type

Type of question	Example question	**Case**
What is the level of…?	What percentage of the population received assistance? How many people arrived? What is the median income?	**direct reporting**
Is there a difference between different parts of the population of interest?	Do IDP's and non IDP's have the same level of access to education? Are prices higher in besieged communities?	**group difference**
did X change?	Did prices increase in the past month? How much?	**change**
does variable A depend on variable B? *not a causal relation !!!*	Is there a connection between increased IDP arrivals and market prices?	**correlation**
Is X larger/smaller than a set value?	Do households on average consume at least X liters of water per day?	**limit**

Table 2a): Hypothesis types

If you followed all steps up to here correctly, filled out the data analysis plan, the sampling frames, the questionnaire and choices correctly, the rest of the steps 2.2. be done using the R hypegrammaR package for the cases "direct reporting" and "group difference" (including the implementation of statistical tests and basic visualisations).

2.2.2 Identify the Types and Combinations of variables

To determine the case, you will need to know what type of variable you are faced with. You can use the table below or extract this information from Kobo: select_one or select_multiple questions are **categorical** while integer and decimal are **numerical.** When you combine two variables to answer your question, the test will depend on the type of the dependent variable.

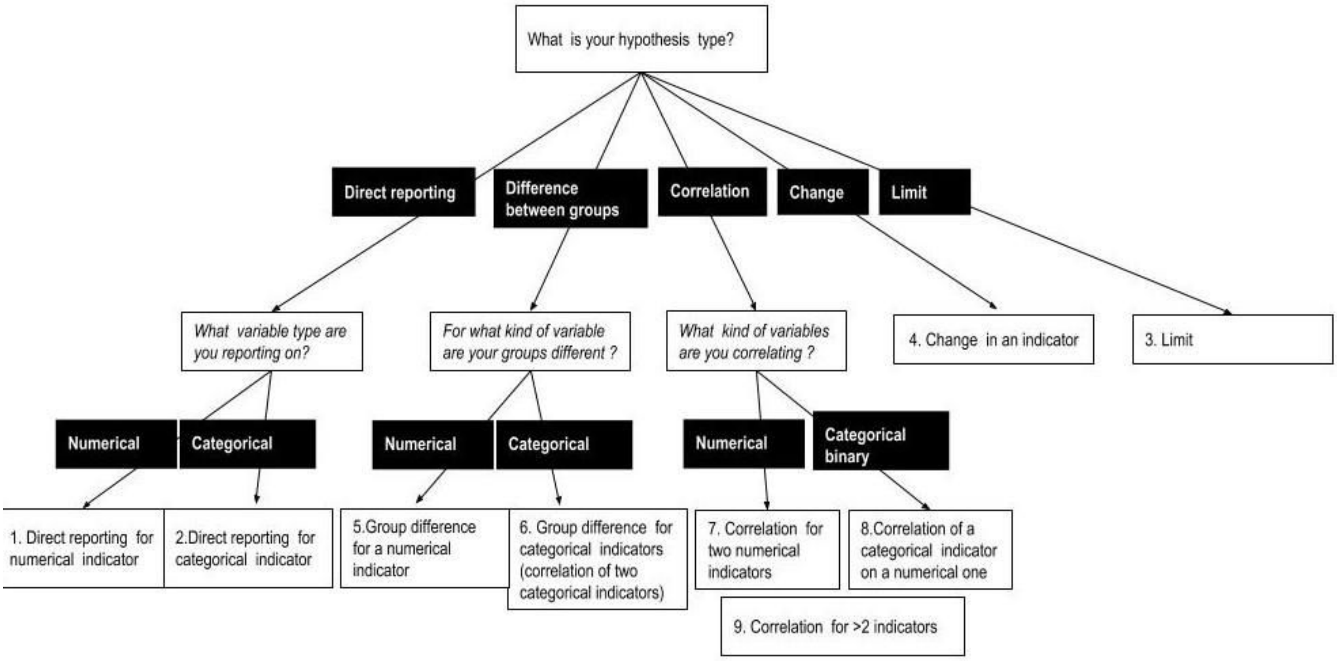
Data type	Examples	**Case**
decimal, integer	Normal: Number of children, Age of oldest household member, number of NFI items received, Household size	**Numerical**

	Log normal: Prices, income, expenses, debt	
select_one from a range select_one where options have a defined order	Quality: poor, normal, good	**Categorical ordinal**
select_one from options that have no well defined order; select_multiple	district codes, pull factors,	**Categorical nominal**
select_one from two logical (TRUE/FALSE) integer only 1 or 0; each answer option from a select_multiple when looked at individually	yes or no questions true or false questions	**Categorical binary**

Table 2b): Data types

2.2.3 Identify your analysis case

After clarifying representativeness, defining your hypothesis and analysis parameters, you perform a statistical test depending on the case determined by your data and hypothesis types. We have defined 9 basic cases most commonly used by IMPACT in probability samples. The exact details of your analysis further depends on the sampling strategy. **Use the following flowchart to identify the case for each line in your extended data analysis plan (each of the results you intend to report on):**


What is your hypothesis type?
Direct reporting
Difference between groups
Correlation
Change
Limit
What variable type are you reporting on?
For what kind of variable are your groups different ?
What kind of variables are you correlating ?
4. Change in an indicator
3. Limit
Numerical
Categorical
Numerical
Categorical
Numerical
Categorical binary
1. Direct reporting for numerical indicator
2.Direct reporting for categorical indicator
5.Group difference for a numerical indicator
6. Group difference for categorical indicators (correlation of two categorical indicators)
7. Correlation for two numerical indicators
8.Correlation of a categorical indicator on a numerical one
9. Correlation for >2 indicators

Graph 1: Flowchart of typical “cases” based on indicator types.

2.3 Apply appropriate analysis for the identified case

This chapters lays out what statistics need to be applied for each case and how to interpret the results. Instructions for the actual implementation: calculating the summary statistics, measuring uncertainty and producing appropriate visualisation can be found in Chapter 5: Implementation. Please note that cases 4, 7, 8 and 9 leave decisions open that may require advanced skills in implementation and interpretation.

Case 1: Direct reporting for a numerical variable
Case 2: Direct reporting for a categorical variable
Case 3: Limit for a numerical or categorical variable
Case 4: Change in a variable over time
Case 5: Group difference for a numerical variable
Case 6: Group differences for a categorical variable
Case 7: Correlation of two numerical variables
Case 8: Correlation of a categorical variable with numerical variable(s)
Case 9: Correlation of >2 variables

2.3.1 Case 1: Direct reporting for a numerical variable

Hypothesis

> The mean/median of variable X in the population is …
>
> **Examples:**
> The median income is ....
> The mean liters of water per person per day is...

Preparation

- If your sampling strategy is **non-probability** sampling, please refer to the non-probability guidelines.
- If your sampling strategy is a **census** you can not measure certainty, skip the "Measuring certainty" instructions.
- If you used a form of **probability sampling** and are aggregating SPSS or Excel, calculate weights and obtain a weighted aggregate finding.
- Find the observed value of your indicator for your sample as per the quantifiable question: usually mean, median for income. (see implementation for each sampling strategy)

Measuring certainty

- If your sampling strategy falls under **probability**: Calculate confidence intervals around your observed value:
    a. Choose your confidence Level (usually 95%)
    b. Sampling strategy:

i. **Simple probability:** simple confidence intervals
    1. If your sample for this question is ≥30:
        a. z-value based confidence intervals for simple probability sampling
    2. If your sample for this question is <30:
        a. t-value based confidence intervals for simple probability sampling

ii. **Stratified:** confidence intervals with stratification adjusted standard error

iii. **Cluster sampling:** confidence intervals adjusted for cluster sampling design effect

iv. **Stratified cluster sampling:** confidence intervals with stratification adjusted standard error and cluster sampling design effect

Interpreting results

- Depending on sampling strategy:
    - The confidence interval is the range of values for which you can say with 95% that it includes the population result.

Answer

The result of the assessed sample is X. The mean/median/etc. value in the population lies within the confidence interval with 95% certainty

Example:
The average amount of water per person per day is 15 liters in the assessed sample. The chance of observing this result with an average in the population below 13.4 or above 16.6 litres is less than 5%.

Reporting

- Add **error bars** to all plots if you used a probability sampling strategy
    a. The R tool produces these plots with error bars automatically
- For large confidence intervals it may be better to report ranges instead of / in addition to observed means to adequately represent the high level of uncertainty.
- State your confidence level (usually 95%) with all confidence intervals and ranges. In this example it would be “15 litres, 95% CI [13.4,16.6]”.

2.3.2 Case 2: Direct reporting for a categorical variable

Hypothesis

The different levels (or proportions) for each answer to a question in the population are ___
How many people fall into each category?

Examples:
…. % of households use a protected water source.
…. out of 1000 households use a protected water source.

Preparation

- If your sampling strategy is **non-probability** sampling, please refer to the non-probability guidelines.
- If your sampling strategy is a **census** you can not measure certainty, skip the “Measuring certainty” instructions.
- If you used a form of **probability sampling** and are aggregating SPSS or Excel, calculate weights and obtain a weighted aggregate finding.
- Find the frequency of each answer of your categorical variable in your sample (see implementation for each sampling strategy)

Measuring certainty

- If your sampling strategy falls under **probability**: Calculate confidence intervals around your observed value:
    a. Choose your confidence Level (usually 95%)
    b. Sampling strategy:
        i. **Simple probability:** simple confidence intervals
            1. If your sample for this question is ≥30
                a. Z based confidence interval for proportion
            2. If your sample for this question is <30
                a. One proportion t test, to define a CI using the sampling distribution
        ii. **Stratified**: confidence intervals with stratification adjusted standard error
        iii. **Cluster::** confidence intervals adjusted for design effect
        iv. **Sampling strategy**: **stratified cluster sampling:** confidence intervals with stratification adjusted standard error and cluster sampling design effect

Interpreting results

- Depending on sampling strategy:
    a. The confidence interval around your observed frequency is the range of values for which you can say with 95% confidence that it includes the population frequency for that answer

Answer

The proportion for each answer of the assessed sample is Y.
The proportion in the population for each answer lies within each answer ‘s confidence interval with 95% certainty.

Example:
50% of households in the sample use a protected water source. The chance of observing this result with an average in the population below 43% or above 57% is less than 5%.

Reporting

- Report each answer as a binary proportion (between 0 and 1) or as a percentage
    a. A percentage is **a ratio** (observed / total) expressed as a fraction of 100
    b. A percentile is the **value** below which a certain proportion of observations fall
- Add **error bars** to all plots, if you used a probability sampling strategy
- For large confidence intervals it may be better to report ranges instead of / in addition to observed frequencies to reflect the large uncertainty
- State your confidence level (usually 95%) with all results and ranges. In this example it would be “50% 95% CI [0.43,0.57]”.

2.3.3 Case 3: Limit for a numerical or categorical variable

Hypothesis

> The value of X in the population is less than / more than a fixed value Z
>
> Example:
> The average liters of water per person per day is more than 15.

Preparation

- If your sampling strategy is **non-probability** sampling, please refer to the non-probability guidelines.
- If your sampling strategy is a **census** you can not measure certainty, skip the “Measuring certainty” instructions.
- If you used a form of **probability sampling** and are aggregating SPSS or Excel, calculate weights and obtain a weighted aggregate finding.
- Find the observed value or proportion of your indicator for your sample as per the quantifiable question (usually mean, median for income)
- Calculate the (weighted) test statistic: the difference between X observed in your sample and target value Z.

Measuring certainty

- If your sampling strategy falls under **probability**: Calculate confidence intervals around your observed value:
    a. Choose your confidence Level (usually 95%)
    b. Sampling strategy
        i. **Simple probability**
            1. If your sample for this question is ≥30:
                a. z-value based confidence intervals
            2. If your sample for this question is <30:
                a. t-value based confidence intervals
        ii. **Stratified**: confidence intervals with stratification adjusted standard error
        iii. **Cluster:** confidence intervals adjusted for cluster sampling design effect

iv. **Stratified cluster sampling:** confidence intervals with stratification adjusted standard error and cluster sampling design effect

- If your sampling strategy falls under **probability**: test the hypothesis that there is a difference between X and Z
  a. Choose your confidence Level (usually 95%)
  b. Sampling strategy: **simple probability:** simple confidence intervals
     i. For a numerical variable: One sample, one sided t test
     ii. For a categorical variable:
        1. One proportion, one sided Z test (compare one observed proportion to a hypothesized one)
        2. Chi Squared goodness of fit test (compare the distribution of proportions in your sample to a hypothesized distribution over multiple groups)
  c. Sampling strategy: **stratified**: t test or z test, standard error adjusted for stratification
  d. Sampling strategy: **cluster sampling:** t test or z test, standard error adjusted for design-effect.
  e. Sampling strategy: **stratified cluster sampling:** t test or z test, standard error adjusted for stratification and cluster sampling design-effect.

Interpreting results

Your null hypothesis was that X was greater (or smaller) in the population than the fixed value Z.

- If your p value is too large to reject the null, then you have not found enough evidence to say that X was smaller (or greater) than Z.
- If your p value is small enough to reject the null, you can state with 95% confidence that the population X is in fact smaller (or greater) than Z.

Answer

The mean/median proportion of the assessed sample is X which is / is not statistically significantly different from Z ($p < 0.05$)

Example:
The mean liters of water per person per day is more than 15.

Reporting

- Report each answer and the difference from the fixed value
  - using color codes for achievement (e.g. <50%; 50-80%, 80-100%) or
  - showing the gap between recorded and fixed value
  - state whether this difference is statistically significant
- Include confidence interval for your observed results and p-values for your hypothesis tests.

2.3.4 Case 4: Change in a variable over time

Hypothesis

> Indicator X changed between time A and time B in the population
> Indicator X changed by ___ between time A and time B
>
> The average liters of water per person per day increased between baseline and endline.

Preparation

- If your sampling strategy is **non-probability** sampling, please refer to the non-probability guidelines.
- If your sampling strategy is a **census** you can not measure certainty, skip the "Measuring certainty" instructions.
- If you used a form of probability sampling and are aggregating in SPSS or Excel, calculate weights and obtain a weighted aggregate finding.
- Compute the difference between the mean/median etc at the two time points in the sample.
- What kind of change in time are you measuring:
    a. If indicator is numerical and time is two fixed points being compared (before and after):
        i. Choose if you need a paired or unpaired test (see below)
        ii. Follow instructions for a difference in groups for a numerical variable (Case 5)
    b. If indicator is categorical and time is two fixed points being compared:
        i. Choose if you need a paired or unpaired test (see below)
        ii. Follow instructions for a difference in groups for a categorical variable (Case 6)
    c. If indicator is numerical and time is continuous:
        i. Time series regression: there is no single way of doing this and interpreting it outlined here. If you would like to perform one, we would love to help.
    d. If indicator is categorical and time is continuous:
        i. Logistic time series:there is no generalizable single way of doing this and interpreting it outlined here. If you would like to perform one, we would love to help.

Measuring certainty

- If your sampling strategy falls under **probability:**
    a. To test the null-hypothesis that there is no difference in value between two fixed time points (that there was no change in the indicator):
        i. Choose your confidence Level (usually 95%)
        ii. Identify if you need a paired or an unpaired test:
            1. **paired:** If the exact same subset of the population was sampled at both points in time
            2. **unpaired**: If two different samples were taken
        iii. Follow the steps in case 5 or case 6, with time as the independent variable to determine whether there was a change, and provide a confidence level for that change if there was one.
    b. To test the null hypothesis that time did not influence the level of indicator X

i. Time series: there is no generalizable way of doing this and interpreting it. If you would like to perform one, we would love to help.

Interpreting Results

For interpreting results see linked case 5 or case 6

Answer

A. The change in X observed in the sample between time A and time B, was Y. At 95% confidence level that change
   a. was statistically significant. The confidence interval for the change that occurred in the population is ... at 95%
   b. was not statistically significant, can not be generalized to the population

B. The time series regression shows that X has the following trends over time: growth, decrease, cyclical, seasonal etc.

Example:
The average liters of water per person per day in the sample increased by 5l. The difference was statistically significant. The confidence interval around the difference is:

Reporting

- Depending on one or two samples:
  - we observed a change in levels of X for the sample population, or
  - we observed different levels of X at point 1 and point 2.
- Include a graphical representation of the change
  - Time series plot if appropriate (more than 4 values)
- Include confidence interval for the observed difference and p-values for your hypothesis tests.
- Things to avoid:
  - ***Do not*** plot regression lines for a change in time that exhibits a low p value and high r squared
  - ***Do not*** mistake a change over time for a causation by external events: values may change over time as part of a general trend. A difference between two time points (before and after an intervention for instance) do not necessarily mean that the change in value was caused by this event.

2.3.5 Case 5: Group difference for a numerical variable

Hypothesis

There is a difference in indicator X between groups A and B (and C…).
The difference in indicator X between groups A and B is ____

Example:
There is a difference in average liters of water per person per day between IDPs, refugees and host community.

Preparation

- If your sampling strategy is **non-probability** sampling, please refer to the non-probability guidelines.
- If your sampling strategy is a **census** you can not measure certainty, skip the “Measuring certainty” instructions.
- If you used a form of probability sampling and are aggregating in SPSS or Excel, calculate weights and obtain a weighted aggregate finding
- Find the observed value or proportion of your indicator in each group as per the quantifiable question (usually mean, median for income).
- Calculate the test statistic: the difference between X observed in your sample for the different groups.

Measuring certainty

- If your sampling strategy falls under **probability**: examine the hypothesis that there is a difference between X and Z
    a. Test the null-hypothesis that there is no difference between the groups (that the difference between the means/medians is 0) and build a confidence interval around your estimated value
        i. D Choose your confidence level and decide if you need a paired or unpaired test (usually 95%)
            1. **paired:**If the exact same subset of the population was sampled at both points in time; for example if the exact same households were interviewed in a baseline/endline study.
            2. **unpaired:** If two different samples were taken
        ii. Decide how many groups you want to compare
            1. **Difference between two independent groups**: use an unpaired two samples t test
            2. **Difference between two time points for the same group:** use a paired two sampled t test
            3. **Difference between more than two groups:** use ANOVA / F-test
                a. If a significant difference between groups is found, run a t-test between the two groups where the difference supposedly originates from.
    b. Sampling strategy
        i. **Simple probability:** unpaired two samples t test/ANOVA
        ii. **Stratified**: unpaired t test / ANOVA with stratification adjusted standard error
        iii. **Cluster:** unpaired t test / ANOVA adjusted for design effect
        iv. **Stratified cluster sampling:** unpaired t test / ANOVA with stratification adjusted standard error, adjusted for cluster sampling design effect.

- If you reject the null-hypothesis that there is no difference and you want to get an estimate for the difference in the population, calculate the confidence interval for the difference in means.

a. Use the test output from the previous step
b. Read (R, SPSS) or calculate (Excel) the confidence interval for difference in means
    i. Sampling strategy: **simple probability** use simple confidence intervals
    ii. Sampling strategy: **stratified**: confidence intervals with stratification adjusted standard error
    iii. Sampling strategy: **cluster sampling:** confidence intervals adjusted for cluster sampling design effect
    iv. Sampling strategy: **stratified cluster sampling:** confidence intervals with stratification adjusted standard error and cluster sampling design effect

Interpretation

- If the t-test is significant, there is evidence that there is at least some difference between two groups.
- if ANOVA is significant, there is evidence that there is at least some difference between any of the groups.Most of the time, a simple ANOVA will not be sufficient to prove a relation between two variables.
    a. The residuals of the anova test for each cell can give you an indication of which differences drove the significance of the test.
    b. Generate more specific hypothesis only connecting a single choice in both groups, then perform an exact test on those.

Answer

A. There is / is not a difference between groups A and B for variable X with 95 % confidence.
B. The average difference between groups A and B for variable X is between the bounds of this confidence interval

Example: There is a statistically significant difference in average liters of water per person per day between IDPs, refugees and host community in the population. The difference between IDPs and refugees was of 3L per person per day with a confidence interval of:

Reporting

- Add **error bars** to all plots if you used a probability sampling strategy
    - The R tool produces barcharts with error bars automatically.
- Include confidence interval for the observed difference and p-values for your hypothesis tests.
- State your confidence level (usually 95%) at least with an asterisk or footnote.

2.3.6 Case 6: Group differences for a categorical variable

Hypothesis

There is a difference in frequency of X between groups A and B

Example:
IDPs, refugees and returnees are not equally likely to use a protected water source

Preparation

- If your sampling strategy is **non-probability** sampling, please refer to the non-probability guidelines.
- If your sampling strategy is a **census** you can not measure certainty, skip the "Measuring certainty" instructions.
- If you used a form of probability sampling and are aggregating in SPSS or Excel, calculate weights and obtain a weighted aggregate finding
- Find the observed proportion of your indicator for your sample as per the quantifiable question
- Calculate the test statistic: the difference between X observed in your sample for the different groups.

Measuring certainty

- If your sampling strategy falls under **probability**: Calculate confidence intervals around your observed value:
    a. Choose your confidence Level (usually 95%)
    b. Sampling strategy:
        i. **simple probability** use simple confidence intervals
        ii. **stratified**: confidence intervals with stratification adjusted standard error
        iii. **cluster sampling:** confidence intervals adjusted for cluster sampling design effect
        iv. **stratified cluster sampling:** confidence intervals with stratification adjusted standard error and cluster sampling design effect
- If your sampling strategy falls under **probability**: test the null hypothesis that there is no difference between the groups (that the difference between the frequencies of responses is 0)
    a. Choose your confidence level and decide if you need a paired or unpaired test (usually 95%)
        i. **paired:**If the exact same subset of the population was sampled at both points in time; for example if the exact same households were interviewed in a baseline/endline study.
        ii. **unpaired**: If two different samples were taken
    b. Sampling Strategy:
        i. **simple probability sampling**: Use a Chi Squared test for paired or unpaired samples
        ii. **stratified,cluster or stratified cluster probability sampling**: Use a Chi Squared test, for paired or unpaired samples, correcting for stratification and cluster design effects.

Interpretation

- The chi-squared test takes all possible response combinations and tests their joint null “all of these variables are statistically independent! if ANY of them is not I’ll reject”.
- Most of the time, a simple chi squared will not be sufficient to prove a relation between two categorical variables.
  a. The residuals of the chi squared test for each cell can give you an indication of which differences drove the significance of the chi squared.
  b. Generate more specific hypothesis only connecting a single choice in both groups, then perform a chisq or fisher's exact test on those.

Answer

A. Significant results: There is a statistically significant difference in the frequencies for variable X between groups A and B in the population of interest. The probability to observe a difference larger or equal than our sample if there actually is no difference is < 5% (1/20 tests)
B. Insignificant results: There is no conclusive evidence from the sample.

Example:
IDPs, refugees and returnees are not equally likely to use a protected water source. Specifically, a higher % of IDPs uses a protected water source compared to refugees.

Reporting

- Add **error bars** to all plots if you used a probability sampling strategy.
  - The R tool produces barcharts with error bars automatically.
- Include confidence interval for the observed difference and p-values for your hypothesis tests.
- State your confidence level (usually 95%) at least with an asterisk or footnote.

2.3.7 Case 7: Correlation of two numerical variables

Hypothesis

There is a correlation between variables X and Y in the population
The strength and direction of that relationship are ____

Example:
There is a correlation between income and average liters of water per person per day.

Preparation

- If your sampling strategy is **non-probability** sampling, please refer to the non-probability guidelines.
- If you used a form of probability sampling calculate weights.
- Check your sampling strategy and choose the appropriate regression model

  a. **simple probability:** simple linear regression
  b. **Stratified**: linear regression with stratification adjusted standard error
  c. **Cluster sampling:** regression adjusted for design effect
- transform any variables that you believe will not have a linear relationship to the other variables (e.g. log transform for income)
- Run a **linear regression**
  a. Your dependent variable is the numerical variable that you think is influenced by the other indicators (e.g. price).
  b. Your independent variable(s) is what you think is influencing the dependent variable.

Measuring certainty

- If your sampling strategy falls under **probability:** Check the significance of the correlation in the regression output
  a. Choose your confidence Level (usually 95%)
  b. Your regression output should include a p-value or an R squared statistic (see below)
    i. Use the R squared to determine the strength of the correlation: how much of the variance in your dependent variable is explained by your independent variable
    ii. Use the p value as usual, as a measure of significance
  c. Check residuals for non normality
    i. If the residuals exhibit a pattern other than a normal distribution, the relation is not a linear correlation and we need to adapt the regression model (not covered by this document)

Interpreting results

The slope of the output is your observed effect size, whether it is significant depends on a combination of the R squares and p-value. Your regression can result in four alternative interpretations:

1. Low R-squared and low p-value (p-value <= 0.05): means that we have certainty that Y explains a very small part of the variation in X (the model is valid, but is missing some important aspects driving X)
2. Low R-squared and high p-value (p-value > 0.05): means that Y doesn't explain much of the variation in X and we have no certainty about it (no correlation is confirmed)
3. High R-squared and low p-value: means that Y explains a lot of variation within X and we have certainty about it(a great model: the two are highly correlated and their correlation is significant)
4. High R-squared and high p-value: means that your model appears to explain a lot of the variation within X but we have no certainty about it (your model is worthless because of insufficient data)

The Intercept is generally not important indicative of anything: ignore.

Answer

1. Significant + weak fit: A correlation between X and Y exists but Y is not sufficient to explain X (discard)
3. Significant + strong fit + big slope: A large, statistically significant correlation exists between X and Y.
3. Significant + strong fit + small slope: Y has a statistically significant effect on X
4. Not significant + strong fit: Y can explain X in your model, but the correlation is not significant. (discard)

Example:

1. Significant + weak fit: Income and water consumption are correlated, but income is not an important factor.
3. Significant + strong fit + big slope: Income is strongly correlated with water consumption. We are at least 95% confident that at least some correlation exists.
3. Significant + strong fit + small slope: Water consumption does not vary much, but the small differences correlate well with income. We are at least 95% confident that at least some correlation exists.
4. Not significant + strong fit: Income may affect water consumption but there is not enough data to make any conclusions.

Reporting

- Interpret / Report
    a. Report the **effect size** of the correlation: the slope estimate and **the certainty:** the p value and R squared
    b. Consider the following limitations of a regression
        i. Correlation does not equal causation
        ii. Do not add regression lines in graphs if you observe low significance or weak fit

2.3.8 Case 8: Correlation of a categorical variable with numerical variable(s)

Hypothesis

Indicator X depends on indicator(s) T (and Y)?
The strength and direction of that relationship are ____

Example:
HH with high income are more likely to use a protected drinking water source.

Preparation

- If your sampling strategy is **non-probability** sampling, please refer to the non-probability guidelines.
- Check your sampling strategy and choose the appropriate regression model
    a. **simple probability:** simple logistic regression
    b. **Stratified**: logistic regression with stratification adjusted standard error
    c. **Cluster sampling:** logistic regression adjusted for design effect

- Run a **logistic regression** of the categorical indicator (CV) on the numerical one(s).
    a. Your dependent variable is the categorical binary variable that you think is influenced by the other indicators (e.g. intention to move).
    b. Your independent variables are the variables that you think are influencing the numerical variable

Measuring certainty

-
- If your sampling strategy falls under census or **probability**: Check the significance of the correlation
    a. Choose your confidence Level (usually 95%)
    b. Your regression output should include a p-value or an F statistic. If this number is less than 0.05 (for 95% confidence), reject the null hypothesis that there is no correlation.

Interpretation

- Look at the output of your general linear model
    - Deviance of the model: Goodness of fit
    - Pr(>|Z|): Significance of the independent variable in predicting the dependent variable.
    - Estimate: strength of relationship. For a one unit increase in T we expect the log odds of X to change by estimate of T.
        - Exponentiate your estimate to find the odds ratio of being X given that you are T (if T is categorical) or given a one unit increase in T (if T is numerical), over the odds of not.
        - To understand the odd ratio, think of it as the amount of influence the independent variables T has on the frequency of “yes” responses for the categorical indicator X.
    - A more standardised interpretation falls outside of the scope of these guidelines. If you would like to perform a logistic regression please contact us.

Answer

Significant: The outcome of X depends on T (and Y)
    a. The average effect of T (and Y) on X is ...
Not significant: There is no conclusive evidence that X is influenced by T (and Y).

Example:
HH with higher income are more likely to use a protected drinking water source (x% more likely per $ in the sample); There is at least some relationship in the population with 95% confidence.

Reporting

- Interpret / Report
    a. Intercept is generally not important: ignore

b. For a logistic regression, do not include any mentions of causal directionality (no “increases the chances of”), and only point out the statistically significant correlation with wording like “A person of group A is z times more likely than a person of group B to X“

c. Correlation does not equal causation. Keep the language sufficiently vague (like “at least some relationship” in the example) to avoid leading people to make misleading conclusions. Contact the data unit if you would like to do a causal inference.

- Remember that the estimate refers to the log odds. It will likely be returned with a 95% CI. Both bounds of the CI need to be exponentiated to find the CI of the odds ratio.

2.3.9 Case 9: Correlation of >2 variables

Hypothesis

Does indicator X depend on indicators Y, Z, ... ?

Can the food consumption score be predicted from a combination of income, number of children and liters of water per person per day?

Preparation

- If your sampling strategy is **non-probability** sampling, please refer to the non-probability guidelines.
- Check your sampling strategy and choose the appropriate regression model
- **Census** or **simple probability:** simple regression
- **Stratified**: regression with stratification adjusted standard error
- **Cluster sampling:** regression adjusted for design effect
- Run a regression:
  a. Your dependent variable is the numerical variable that you think is influenced by the other indicators (e.g. price).
  b. Your independent variables are the variables that you think are influencing the numerical variable

Measuring certainty

- Check the significance of the prediction
  a. Choose your confidence Level (usually 95%)
  b. If the p value for an indicator is more than 0.05, fail to reject the null hypothesis that there is no correlation, ***and re-run the model without this predictor.***
  c. If your model does not appear to have a linear relationship
     i. Abandon
     ii. Control for interaction (correlation between predictor variables)

Interpreting results

- Your regression output should include an F statistic and a p value.
  a. In your resulting model, if the p value is less than 0.05 (for 95% confidence) reject the null hypothesis that there is no correlation.

The slope of the output is your observed effect size, whether it is significant depends on a combination of the R squares and p-value. Your regression can result in four alternative output combinations.

1. Low F and low p-value (p-value <= 0.05): means that your model doesn't explain much of variation in X but it is significant (there is a correlation but it is weak)
2. Low F and high p-value (p-value > 0.05): means that your model doesn't explain much of variation in X and it is not significant (worst scenario)
3. High F and low p-value: means that your model explains a lot of variation within X and is significant (this is a great model that helps predict the variables)
4. High F and high p-value: means that your model explains a lot of variation within X but is not significant (your model is worthless)

Answer

1. Significant + strong fit + big slope: A statistically significant correlation exists between the independent variables and X
2. Significant + weak fit + big slope: There is a statistically significant correlation between the independent variables and X, but the model is missing some important factors.
3. Significant + strong fit + small slope: There is not much variation in X, but the little variation is well explained by the independent variables.
4. Not significant & big slope: Not enough data
5. Not significant & small slope: No relationship or not enough data

Example:

1. Significant + strong fit + big slope: The food consumption score can be explained well from a combination of income, number of children and liters of water per person per day. At least some relation exists with 95% confidence
2. Significant + weak fit + big slope: A combination of income, number of children and liters of water per person per day is correlated with the food consumption score, but there are other important factors missing
3. There is not much variation in the food consumption score, but the little variation is well explained by the independent variables.
4. Not significant & big slope: Not enough data
5. Not significant & small slope: No relationship or not enough data

Reporting

- Interpret / Report
    a. No direction
    b. Correlation != causation
    c. Interaction effects

2.4 Interpretation

2.4.1 Effect size vs. Certainty

From a data analysis perspective, it is important to distinguish between effect size and certainty:
Effect size refers to the observed value: median, mean, or magnitude of difference between two values (**what you saw**). **Certainty** refers to how likely it is that the reported effect size that you saw in the sample is also true in the population. **(how sure you are of what you saw)**

We should always strive to communicate both in conjunction, but without confusing them. Examples:

- In a baseline endline study with a large sample, we could have a significant difference for a change of 0.01%. While we are very confident that there was a change (statistically significant, high certainty) we should not report this as if there had been any relevant change (it's only 0.01%)
- In a bar chart, the height of the bars show the effect size, the error bars show the uncertainty. Both are communicated clearly and in conjunction, in a way they can not be confused.

2.4.2 Multiple hypothesis: Interpretation in context

For each research question, you will have analysed multiple indicators, and thus, have multiple hypothesis test results; some of them may be statistically significant, some not, some may show large effect sizes and some small. It is essential to interpret and describe these results *in context of each other, including all large/small/significant/ insignificant results.*

- If the indicators for the same research question all **show a similar picture:**
    - **with high certainty**: we have a clear and confident answer to the research question
    - **with low certainty**: there is a clear answer, but the confidence is low; more data is needed to be certain
- if the indicators for the same research question give **conflicting answers:**
    - **with high certainty**: Interesting! We are clearly missing something here.
        - Maybe there are some "unknown unknowns" that would explain the situation.
        - Maybe checking back with the field team resolves some of the apparent differences
        - Maybe someone with a strong qualitative understanding of the local context can help; Likely the field team could help judge the situation
        - You could explore the questions raised by the conflicting answers in FGDs or other semi-structured interviews to seek explanations
        - It is a great piece of research, as long as we report the confident but conflicting results in context of each other, and explain or point at the unexplained contradictions.
    - **with low certainty**: There may be some dynamics we do not understand, but likely we simply do not have enough information to get a conclusive picture. Report the results but be clear about the limitations.