Technical Rapid Response Team

Link NCA
NUTRITION CAUSAL ANALYSIS

Quantitative Data Management and Analysis Session - R

Alexandra Humphreys

23 March 2020

USAID FROM THE AMERICAN PEOPLE
International Medical Corps
ACTION AGAINST HUNGER
Save the Children
Irish Aid Rialtas na hÉireann Government of Ireland
Sida SWEDISH INTERNATIONAL DEVELOPMENT COOPERATION AGENCY

Objectives of this Session

- Review current best practices for Link NCA Quantitative Data Management and Analysis
- Review descriptive statistics for samples
- Review analysis of statistical associations
- Review presentation of results

Note: this training does not cover the selection or operationalization of hypothesized risk factors, as this training is catered towards the handling of data post quantitative data collection.

A Note on Data Cleaning

Data cleaning is a critical step in quality results. The removal or modification of observations in the dataset during cleaning should be justified and documented. This serves to:

- Increase accountability of the analyst
- Ensure that results can be replicated (ensuring validity)

(Using R, for example, these changes can be recorded using one of several R notebooks. If changes are made in an Excel, they should be documented elsewhere)

Source: https://www.datacamp.com/community/blog/jupyter-notebook-r#jupyter

Missing and Unknown Data

Missing data should never be filled in without a strong justification. Empty variables should be left blank, and if a large proportion of the responses are missing (rule of thumb: >20%), this should be discussed because this may risk the representativeness of the data.

HOWEVER: having an “unknown” option for quantitative questions is very important, this avoids respondents/surveyors being forced to make a response fit into a “yes/no” answer.

For calculating statistical associations, “unknown” responses should be coded as missing as they do not contribute to the analysis.

Descriptive Statistics

Technical
Rapid
Response
Team

Analysing and Reporting Prevalence

When basing the quantitative data collection on the SMART Methodology, it is possible to analyze and report the prevalence of binary or categorical or indicators for the area/population of interest.

However:

- *The prevalence must be calculated in consideration of the sampling methodology (cluster or simple random sampling).*
- *The area/population for the prevalence must be clearly stated (i.e. if calculating the prevalence uniquely among households with children <5 yrs)*

Technical Rapid Response Team

Global Technical Assistance Mechanism *for* Nutrition

Adjusting for Complex Sampling

When analysing descriptive statistics, we must ensure that we've accounted for the **sampling design**

In R this is accomplished using the "survey" package, whereby the sampling (svydesign) can be established:

apipop – *in the case of an exhaustive survey*

apisrs – *when using simple random sampling*

apiclus2 – *for two stage cluster sampling (most common for SMART surveys)*

Source: https://rpubs.com/trjohns/survey-cluster

& https://cran.r-project.org/web/packages/survey/survey.pdf

Technical Rapid Response Team

Example: Analysing Prevalence

Prevalence R coding example:

Again, using the "survey" package, we can generate prevalence:

svyciprop – *generates proportions (prevalence) with confidence intervals*

level - *set the confidence level to 95%*

deff – *returns the design effect*

Source: https://www.rdocumentation.org/packages/survey/versions/3.37/topics/svyciprop

Technical Rapid Response Team

Example: Presenting Prevalence

Prevalence and 95% CI should be presented for each binary or categorical variable, with the population clearly noted in the report.

For this example, the prevalence is based only on households with children under five and was reported as such.

N=overall sample

n=affected sample subset

95% CI in accordance with sampling design

Risk Factor *Logistic Regression*			
Indicator	N	n	Prevalence [95% CI]
Male child	416	201	48.3% [43.6-53.1]
Female head of household	416	157	37.7% [29.9-46.3]
Male child *and female head of household*	201	73	36.3% [27.9-45.6]
Barriers to access of health center	414	281	67.9% [59.0-75.7]
Fever	414	189	45.7% [38.8-52.7]

Presenting Design Effect

Reporting the **design effect** (DEFF) allows us to assess the heterogeneity of the risk factor.

Generally speaking, ≤1.00 DEFF indicates homogeneity, around 1,50 some heterogeneity, ≥2.00 high heterogeneity.

DEFF

Risk Factor *Logistic Regression*				
Indicator	**N**	**n**	**Prevalence [95% CI]**	**Design Effect**
Male child	416	201	48.3% [43.6-53.1]	0.94
Female head of household	416	157	37.7% [29.9-46.3]	3.02

Example: Analysing Mean

Mean R coding example:

Again, using the "survey" package, we can generate the following:

svymean – *generates the mean adjusting for sampling*

confint() – *generates a confidence interval*

level = *set the confidence level to 95%*

deff – *returns the design effect*

Source: https://cran.r-project.org/web/packages/survey/survey.pdf

Technical Rapid Response Team

Example: Presenting the Mean

Mean and 95% CI should be presented for each continuous variable, with the population clearly noted in the report.

For this example, the mean is based only on households with children under five and was reported as such.

N=overall sample

Mean and 95% CI in accordance with sampling design

Standard deviation

Risk Factor *Linear Regression*			
Indicator	**N**	**Mean [95% CI]**	**Std. Dev.**
Distance to health center (hours)	416	1.68 [1.23-2.14]	1.45
Number of prenatal consultations	327	4.12 [3.94-4.30]	0.93
Birth spacing (months)	223	27.1 [24.7-29.4]	10.54

Statistical Associations

Technical
Rapid
Response
Team

Analyze One Risk Factor at a Time

Important note: **multivariate analysis** of statistical associations is not recommended by the Link NCA at this time. The independent variables (risk factors) should be examined one at a time against dependent (outcome) variables. For two reasons:

- Multivariate analysis is highly complex and requires robust consideration of confounding factors.
- We want to refrain from comparing strength of statistical significance between independent variables. We are interested in statistical significance ($p<0.05$ yes/no only), then these associations are mapped to demonstrate pathways.

Technical Rapid Response Team

Logistic Regression

Logistic regression is a method of demonstrating statistical significance between an independent variable (risk factor) and an outcome variable.

Requirements:

- The outcome and independent variable must both be binary (0/1)

With '1' being the condition of interest

Logistic Regression with R using the glm() function:

glm(outcome variable independent variable, dataset)

Source: https://www.datacamp.com/community/tutorials/logistic-regression-R

Technical Rapid Response Team

Global Technical Assistance Mechanism *for* Nutrition

Logistic Regression

For **logistic regression**, the sampling method is not considered because we are interested in the statistical association (p-value), not in representativeness.

P-value to demonstrate statistical significance (<0,05)

Outcome Variable			
GAM (MUAC) *Children 6-59 months*		Combined GAM* *Children 6-59 months*	
P-value	Odds Ratio [95% CI]	P-value	Odds Ratio [95% CI]
0.626	0.84 [0.41-1.71]	0.909	0.97 [0.54-1.72]
0.956	1.02 [0.57-1.80]	0.819	1.05 [0.68-1.62]
0.471	1.65 [0.42-6.38]	0.607	0.79 [0.32-1.93]

Odd ratio and 95% CI to show directionality and precision.

Technical Rapid Response Team

Linear Regression

Linear regression is a method of modelling the relationship between an independent variable (risk factor) and an outcome variable.

Requirements:

- The outcome variable must be continuous
- The risk factor should be continuous (*can* be categorical but requires special attention)

Linear Regression R using the lm() function:

lm(outcome variable independent variable, dataset)

Source: https://rstudio-pubs-static.s3.amazonaws.com/298538_5fe14a64496740d39650f78a0fa3ed91.html

Technical Rapid Response Team

Linear Regression

For **linear regression**, the sampling method is also not considered because we are interested in the statistical association (p-value), not in representativeness.

WHZ			MUAC		
P-value	Coeff	SE	P-value	Coeff	SE
0.384	0.03	0.04	0.184	-0.61	0.46
0.575	-0.04	0.07	0.136	1.13	0.75
0.346	-0.01	0.01	0.277	0.09	0.09

P-value to demonstrate statistical significance (<0,05)

Coefficient helps to infer directionality (interpret carefully)

Standard Error (SE) functions similarly to a standard deviation (SD)

Technical Rapid Response Team

Global Technical Assistance Mechanism *for* Nutrition

Interpreting Directionality

Although we do not attempt to compare the strength of statistical associations between risk factors (p-value <0.05 yes/no only) we do try to interpret ***directionality***.

From this, we can hypothesize if a risk factor is a risk factor or actually a protective factor.

Risk factor: increases likelihood of undernutrition

Protective factor: decreases likelihood of undernutrition

Interpreting Directionality

Logistic regression interpretation

Examples:

Diarrhea/wasting association ($p<0.05$) with an odds ratio >1 is a risk factor – the odds of being malnourished increase.

Measles vaccination/stunting association ($p<0.05$) with an odds ratio <1 is a protective factor – the odds of being malnourished decrease.

Interpreting Directionality

Linear regression interpretation (*is complicated, take your time to think through the results!*)

Examples (assuming p<0.05):

Each one unit increase in household size (person) decreases (negative coefficient) the child's MUAC (mm) – larger household size is a risk factor

Each one unit increase of child's age (months) increases (positive coefficient) the child's WHZ – child's older age is a protective factor

*Note: we do not try to quantify the increase or decrease, our aim is to understand **directionality***

Technical Rapid Response Team

Presentation of Results

Technical
Rapid
Response
Team

Risk Factor Color Codes

More recently, Link NCA has introduced color coding of regression results to ease interpretation.

For risk factors:

P<0.05 is **orange** to highlight statistical significance

P≥0.05 and <0.10 although not statistically significant, is coded as **lighter orange** to highlight a potential association for future research

For protective factors:

P<0.05 is **green** to highlight statistical significance

P≥0.05 and <0.10 also coded as **Lighter green** to highlight a potential association for future research

Annexing Analysis Tables

Example **logistic** regression results table

Risk factor *Logistic regression*					Outcome Variable: Wasting *Children 6-59 months*		GAM MUAC *Children 6-59 months*		cGAM *Children 6-59 months*		Stunting *Children 6-59 months*	
Indicator	N	n	Prevalence [95% CI]	Design effect	P-value	Odds Ratio [95% CI]	P-value	Odds Ratio [95% CI]	P-value	Odds Ratio [95% CI]	P-value	Odds Ratio [95% CI]
Male child	356	174	48.9% [43.7-54.1]	1.00	0.551	1.36 [0.49-3.76]	0.899	0.14 [0.02-1.18]	0.940	1.04 [0.40-2.69]	0.809	0.94 [0.58-1.53]
Female head of household	356	234	65.7% [62.5-68.8]	0.40	0.172	0.49 [0.18-1.36]	0.827	0.85 [0.20-3.64]	0.150	0.50 [0.19-1.29]	0.438	0.82 [0.49-1.36]
Mother currently <19 years old	356	194	67.8% [62.1-73.1]	1.02	0.409	1.92 [0.41-9.11]	0.615	0.63 [0.10-3.84]	0.722	1.27 [0.34-4.83]	0.231	1.41 [0.80-2.47]
Household >1 child under 5 years old	356	100	28.1% [25.1-31.3]	0.42	0.135	2.18 [0.79-6.08]	0.507	1.64 [0.38-7.01]	0.099	2.26 [0.86-5.94]	0.621	1.15 [0.67-1.96]
Household size > 5 members	356	85	23.9% [18.1-30.8]	2.00	0.120	0.20 [0.03-1.53]	0.950	1.05 [0.21-5.34]	0.205	0.38 [0.09-1.70]	0.966	0.99 [0.56-1.74]
Household size > 7 members	356	29	8.2% [5.0-12.9]	1.87	0.559	1.58 [0.34-7.40]	0.010	7.23 [1.62-32.3]	0.214	2.3 [0.62-8.56]	0.274	1.59 [0.69-3.64]
Measles vaccination Confirmed by card	341	216	60.7% [54.0-67.0]	1.64	0.032	0.53 [0.25-1.88]	0.225	0.41 [0.10-1.74]	0.423	0.68 [0.26-1.76]	0.089	0.75 [0.42-0.95]
Vitamin A supplementation	353	52	14.6% [9.5-21.8]	2.75	0.846	0.81 [0.10-6.75]	0.271	0.32 [0.04-2.45]	0.991	1.00 [0.51-1.97]	0.700	0.84 [0.35-2.01]
Fever	353	162	45.5% [38.7-52.5]	1.80	0.771	0.86 [0.31-2.37]	0.395	1.88 [0.44-8.00]	0.822	1.12 [0.43-2.89]	0.945	0.98 [0.61-1.59]
Diarrhea	353	242	68.0% [61.9-73.5]	1.43	0.041	1.51 [0.47-4.80]	0.007	2.48 [0.29-7.49]	0.033	1.76 [0.56-5.50]	0.096	1.32 [0.78-2.23]
Diarrhea *for unbathed child <24 months*	68	25	36.8% [32.5-43.8]	0.40	0.172	0.49 [0.18-1.36]	*Perfect collinearity**				0.438	0.82 [0.49-1.36]

Technical Rapid Response Team

Global Technical Assistance Mechanism *for* Nutrition

Annexing Analysis Tables

Example **linear** regression results table

Risk factor *Linear Regression*					**WHZ** *Children 6-59 months*			**MUAC** *Children 0-59 months*			**HAZ** *Children 6-59 months*		
Indicator	**N**	**Mean [95% CI]**	**SD**	**Design Effect**	**P-value**	**Coeff.**	**SE**	**P-value**	**Coeff.**	**SE**	**P-value**	**Coeff.**	**SE**
Child age (months)	356	30.8 [29.0-32.5]	0.90	0.79	0.000	0.02	0.00	0.000	0.05	0.00	0.509	0.00	0.01
Mother's age (years)	270	27.4 [26.4-28.4]	0.51	1.6	0.031	0.02	0.01	0.012	0.03	0.01	0.060	0.02	0.01
Mother's MUAC (mm)	266	290.8 [28.6-29.5]	2.34	1.4	0.991	0.00	0.02	0.509	0.02	0.02	0.010	0.06	0.02
Prenatal consultations (0-n)	270	5.7 [5.2-6.2]	0.24	2.1	0.087	-0.04	0.02	0.153	-0.04	0.03	0.735	-0.01	0.02
Number of people in the household (2-n)	356	7.1 [6.8-7.5]	0.18	2.2	0.902	-0.00	0.02	0.035	-0.05	0.06	0.559	-0.01	0.02
Distance to the clinic (minutes)	356	72.8 [60.0-85.7]	6.52	0.3	0.797	0.00	0.00	0.568	0.00	0.00	0.053	-0.05	0.02
Distance to the waterpoint (minutes)	286	13.6 [11.1-16.2]	1.28	0.92	0.306	-0.00	0.00	0.259	-0.01	0.00	0.709	0.00	0.00
IDDS Score (1-14)	159	2.1 [1.9-2.3]	0.09	1.0	0.335	0.084	0.09	0.148	0.15	0.10	0.564	0.06	0.11
Postpartum rest days (0-n)	139	29.6 [23.5-35.7]	3.08	2.2	0.050	0.01	0.00	0.110	0.00	0.00	0.818	0.00	0.00
Child caregiver checklist (1-8)	313	4.1 [3.9-4.4]	0.12	1.2	0.297	0.03	0.03	0.165	-0.05	0.04	0.500	-0.03	0.04
MAHFP (months)	356	10.3 [10.2-10.5]	0.07	2.0	0.031	-0.08	0.05	0.393	-0.05	0.06	0.642	-0.03	0.06

Concluding Thoughts

- The Link NCA Methodology has recently been updated to a more rigorous analytical process of analyzing the associations between risk factors and outcome variables in order to demonstrate pathways
- Data should be carefully managed and cleaned
- Descriptive statistics should be presented for every risk factor variable
- It is recommended that P-values be derived from simple (*not multivariate*) logistic and linear regressions
- All analytical results should be annexed in the final Link NCA report

Technical Rapid Response Team

Global Technical Assistance Mechanism *for* Nutrition

Your Questions are Welcome

Alexandra Humphreys

ahumphreys@actioncontrelafaim.ca

Check us out at TechRRT.org or Twitter: @TechRRT

Technical Rapid Response Team