Heighten Science PUBLICATIONS Corporation

ISSN
2639-6769

Review Article

A review of research process, data collection and analysis

Surya Raj Niraula*

Professor (Biostatistics), School of Public Health and Community Medicine, B.P. Koirala Institute of Health Sciences, Dharan, Nepal

***Address for Correspondence:** Dr. Surya Raj Niraula, Postdoc (USA), PhD (NPL), Professor (Biostatistics), School of Public Health and Community Medicine, B.P. Koirala Institute of Health Sciences, Dharan, Nepal, Tel: +977 9842035218; Email: surya.niraula@bpkihs.edu

Submitted: 10 December 2018
Approved: 10 January 2019
Published: 11 January 2019

Background

Research is the process of searching for knowledge. It is systematic search pertinent information on specific topic of interest. It is a careful investigation or inquiry especially through search for new facts in any branch of knowledge [1]. It is a scientific way of getting answers for research questions and testing hypothesis. The research question is based on uncertainty about something in the population. This can be formulated by searching different literatures from index and non index journals, books, internet, and different unpublished research work etc. A good research question should follow the FINER criteria i.e. Feasible, Interesting, Novel, Ethical and Relevant [2].

The complete research is the whole design which arises from defining research problems to the report writing (Figure 1). The research problems are determined on the basis of well known concept and theories or previous research findings. The assumptions in term of hypothesis are made. The process of inquiry is done by interviewing or observing or recording data and the collected data are analyzed with interpretation. Basically there are two approaches of data collection, quantitative and qualitative. The quantitative approach views human phenomena as being focused to study objective i.e. able to be measured. It has its roots in positivism. Quantitative approach to research involves data collection methods such as structured questionnaire, interviews and observations together with other tools. This approach helps investigators to quantify the information.

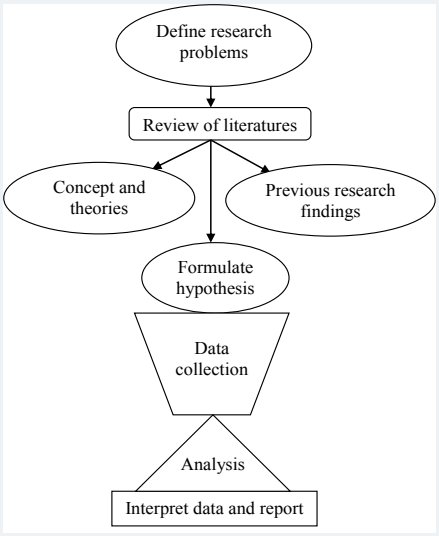

Figure 1: Flow chart-a complete research process.

How to cite this article: Niraula SR. A review of research process, data collection and analysis. Insights Biol Med. 2019; 3: 001-006. https://doi.org/10.29328/journal.ibm.1001014

On the other hand, in depth interviews and unstructured observations are associated with qualitative research. The socially stigmatized and hidden issues are understood and explored by the qualitative research approach. In fact, the purpose of quantitative research is to measure concepts or variables that are predetermined objectively and to examine the relationship between them numerically and statistically. Researches have to choose methods which are appropriate for answering their questions.

Where do data come from?

Basically there are two sources of data, primary and secondary. The secondary data, which are generally obtained from different departments of country like health, education, population and may be collected from different hospitals, clinics, and schools' records, can be utilized for our own research. The secondary sources may be private and foundation databases, city and county governments, surveillance data from the government programs, and federal agency statistics - Census, NIH, etc. The use of secondary data may save our survey cost, time, and may be accurate if the government agency has collected the information. However, there are several limitations over it. The secondary data may be out of date for what we want to analyze. It may not have been collected long enough for detecting trends, e.g. organism pattern registered in a hospital for 2 months. A major limitation is we should formulate the research objectives based on availability of variables in the data set. On the other hand there may be missing information on some observations. Unless such missing information is caught and corrected for, analysis will be biased. There may be many biases like sample selection bias, source choice bias, drop out etc.

If we look at primary source, it has more advantages than the secondary source of data. The data can be collected through surveys, focus groups, questionnaires, personal interviews, experiments and observational study. If we have time for designing our collection instrument, selecting our population or sample, pretesting/piloting the instrument to work out sources of bias, administration of the instrument, and collection/entry of data, using primary source of data collection, researcher may minimize the sampling bias, and other confounding bias.

Analysis

The analysis is an important part of research. The analysis of the data depends upon the types of variables and its' nature [3]. The first thing for the data analysis is to describe the characteristics of the variables. The analysis can be scrutinized as follows:

Summarizing data: Data are the bunch of values of one or more variables. A variable is a characteristic of samples that has different values for different subjects. Value can be numeric, counting, and category. The numeric values of continuous variables are those which have numeric meanings, a unit of measurement, and may be in fraction like – height, weight, blood pressure, monthly income etc. Another type of variables is discrete variables which are based on counting process like – number of student in different classes, number of patients visiting OPD in each day etc [4].

If the variables are numeric, they can be explored by plotting histogram, steam and leaf plot, Whisker box plot, and normal plots to visualize how well the values fit a normal distribution. When the variables are categorical, they can be visualized by pie chart or bar diagrams or just the frequencies and percentages.

A statistic is a number summarizing a bunch of values. Simple or univariate statistics summarize values of one variable. Effect or outcome statistics summarize the relationship between values of two or more variables. Simple statistics for numeric variables are

a) Mean: the average

b) Standard deviation: the typical variation

c) Standard error of the mean: the typical variation in the mean with repeated sampling divided by the root of (sample size).

Mean and standard deviation are most commonly used measure of central tendency and dispersion respectively in case of normally distributed data (Tables 1,2). Median (middle value or 50th percentile) and quartiles (25th and 75th percentiles) are used for grossly non-normally distributed data.

Common statistical tests

The table 1 describes how the different tests are applied for different purpose. Simple statistics for categorical variables are frequency, proportion or odds ratio. The effect size derived from statistical model (equation) of the form Y (dependent) Vs X (Predictor) depend on type of Y and X.

a) If the model is numeric vurses to numeric e.g. SBP and cholesterol; linear regression with correlation coefficient could be used to find the relationship between the variables, where effect statistics gives slope and intercept of the line of equation, called as parameters. The correlation coefficient is explained in terms of variance explained in the model. This provides measures of goodness of fit. Other statistics typical or standard error of the estimate provides the residual error and based measure of validity (with criterion variable on the Y axis).

b) But if the model is numerical versus categorical e.g. marks in medical exam versus sex, the model will be t-test for 2 groups and one way ANOVA for more than two groups (Table 2). Effects statistics will be difference between means, express as row difference, percent difference, or fraction of the root mean square error which is an average standard deviation of the two groups. The table 2 shows the result of ANOVA for academic performances.

c) If the model is numerical versus categorical (repeated measures in different time interval) eg. weight loss (kg) and each month, the model will be paired t-test (2 months) and repeated measures ANOVA with one within the factor (>2 month), where effect statistics will be change in mean expressed as row change, percentage change, or fraction of pre standard deviation.

d) If the model is categorical versus categorical e.g. smoking habit versus sex, the model test will Chi-square or Fisher exact tests where the effect statistics provides relative frequencies, expressed as a difference in frequencies, ratio of frequencies (relative risk) or odds ratio. The relative risk is appropriate for

Table 1: Test statistics based on types of variables.

Y (Response)	X (Predictor)	Model/Test	Effect Statistics
Numeric	Numeric	Regression	Slope, intercept, Correlation
Numeric	Categorical	t-test, ANOVA	Mean difference
Categorical	Categorical	Chi-square, Fisher exact	Frequency difference or ratio
Categorical	Numeric	Categorical	Frequency ratio

Table 2: Academic performance in different levels of the MBBS students during 1994 to 1996.

Batches (n)	Mean ± SD								
	SLCS	ISS	EES	MBBSI	MBBSII	MBBS III	MBBSIV	MBBS V	MBBS Total
1994 (29)	74.2±6.3	71.9±7.8	71.3±2.5	67.8±5.2	71.0±5.1	73.3±5.9	69.5±4.3	65.8±3.0	69.5±4.2
1995 (29)	75.2±5.1	69.98.7	52.1±4.0	67.3±5.4	68.04.0	65.3±3.5	65.3±3.5	62.317.6	65.6±5.6
1996 (28)	76.4±5.3	71.28.4	54.4±4.2	69.3±5.1	73.24.6	64.3±3.0	65.7±3.2	66.53.2	67.8±3.4
F value	1.1	0.4	241.2	1.1	9.2	42.7	11.1	1.3	5.2
P value	NS	NS	<0.0001	NS	<0.0001	<0.0001	<0.0001	NS	< 0.01

Source: Niraula et al, 2006[6].

cross-sectional or prospective designs. It is a risk of having a certain disease for one group relative to other group. Odds ratio is appropriate for case-control designs, a cross product of 2X2 contingency table.

e) If the model is nominal category versus >= 2 numeric e.g. heart disease versus age, sex and regular exercise, the model test will be categorical modeling where effect statistics will be relative risk or odds ratio. This can be analyzed using logistic regression or generalize liner modeling. The complex models will be most reducible to t tests, regression, or relative frequencies.

f) If the model is controlled trial (numeric versus 2 nominal categories) e.g. strength vs trial vs group, the model will be unpaired t test of change scores (2 trails, 2 groups) or repeated measures ANOVA with within-and between-subject factors (>2 trials or groups) where the effect statistics will be the difference in change in main expressed as raw difference, percent difference, or fraction of the pre standard deviation [5].

g) If the model is extra predictor variable to "control for something" (numeric versus >= 2 numeric) eg. Cholestrol vs physical activity vs age, multiple linear regression or analysis of covariance (ANCOVA) can be used. Another example of use of linear regression analysis to find the significant predictor for MBBS performance is demonstrated in table 3.

h) If we want to find out degree of association between two numeric variables, we can examine by the correlation coefficient which may take values from -1 to +1. A positive value of the coefficient indicates positive association, whereas negative coefficient indicates a negative association. Another example of use of correlation matrix to show the association between the two scores in different classes (Table 4).

Generalizing from a sample to a population

We study a sample to estimate the population parameter. The value of a statistic for a sample is only an estimate of the true (population) value which is expressed precision or uncertainty in true value using 95% confidence limits. The confidence limits represent likely range of the true value. There is a 5 % chance the true value is outside the 95% confidence interval, also called level of significance: the type I error rate [7,8].

Statistical significance is an old-fashioned way of generalizing, based on testing whether the true value could be zero or null.

Table 3: Stepwise linear regression for predicting MBBS performance.

Model	Coefficients			P value	Collinearity Statistics	
	Unstandardized Coefficients		Standardized Coefficients		Tolerance	VIF
	B	SE	Beta			
(Constant),	57.34	4.32		0.00		
Intermediate in Science Score	0.145	0.06	0.253	<0.02	1.0	1.0

R^2 = 0.064, Adjusted R^2 = 0.053; F(1,84)=5.77, P<0.02 Source: Niraula et al, 2006 [6].

Table 4: Correlation matrix of academic performance of MBBS students.

	SLCS	ISS	EES	MBBSI	MBBSII	MBBSIII	MBBSIV	MBBS V
ISS	0.290*3							
EES	-0.079	0.076						
MBBSI	0.177	0.247*2	-0.094					
MBBSII	0.167	0.208	0.025	0.770*4				
MBBSIII	0.075	0.245*2	0.647*5	0.299*3	0.442*5			
MBBSIV	0.151	0.197	0.362*4	0.632*5	0.712*5	0.755*5		
MBBS V	0.059	0.114	-0.055	0.544*5	0.404*5	0.224*1	0.512*5	Scores
MBBSS	0.145	0.242*2	0.179	0.806*5	0.789*5	0.631*5	0.872*5	0.7931*5

Source: Niraula et al, 2006 [6].

- Assume the null hypothesis: that the true value is zero (null).
- If we observed value falls in a region of extreme values that would occur only 5% of the time, we reject the null hypothesis.
- That is, we decide that the true value is unlikely to be zero; we can state that the result is statistically significant at the 5% level.
- If the observed value does not fall in the 5% unlikely region, most people mistakenly accept the null hypothesis: they conclude that the true value is zero or null!
- The p value helps us to decide whether our result falls in the unlikely region.

If $p<0.05$, our result is in the unlikely region.

One meaning of the p value: the probability of a more extreme observed value (positive or negative) when true value is zero. Better meaning of the p value: if we observe a positive effect, 1 - p/2 is the chance the true value is positive, and p/2 is the chance the true value is negative. For example: If we observe a 1.5% enhancement of performance (p=0.08). Therefore there is a 96% chance that the true effect is any "enhancement" and a 4% chance that the true effect is any "impairment". This interpretation does not take into account trivial enhancements and impairments. Therefore, if we must use p values as possible, show exact values, not $p<0.05$ or $p>0.05$. Meta-analysts also need the exact p value (or confidence limits).

If the true value is zero, there's a 5% chance of getting statistical significance: the Type I error rate, or rate of false positives or false alarms. There's also a chance that the smallest worthwhile true value will produce an observed value that is not statistically significant: the Type II error rate, or rate of false negatives or failed alarms. The type II error is related to the size of samples in the research. In the old-fashioned approach to research design, we are supposed to have enough subjects to make a Type II error rate of 20%: that is, our study is supposed to have a power of 80% to detect the smallest worthwhile effect. If we look at lots of effects in a study, there's an increased chance being wrong about at least one of them. Old-fashioned statisticians like to control this inflation of the Type I error rate within an ANOVA to make sure the increased chance is kept to 5%. This approach is misguided.

Summary

In summary, the research process begins with defining research problems and then review of literatures, formulation of hypothesis, data collection, analysis, interpretation and end in report writing. There are chances of occurrence of many biases in data collection. Importantly, the analysis of research data should be done with very caution. If a researcher use statistical test for significance, he/she should show exact p values. It is also better still, to show confidence limits instead. The standard error of the mean should be shown only in case of estimating population parameter. Usually between-subject standard deviation should be presented to convey the spread between subjects. In population studies, this standard deviation helps convey magnitude of differences or changes in the mean. In interventions, show also the within-subject standard deviation (the typical error) to convey precision of measurement. Standard deviation helps convey magnitude of differences or changes in mean performance.

Chi-square and fisher exact tests are used for categorical variables (category versus category). Two numerical variables are examined by correlation coefficient. For the model numeric versus two category, t test will be the suitable in case of normal data, ANOVA should be applied for the model numeric versus >=2 categorical variables. Multiple regression model is used to find out adjusted effects of all possible predictors (>=2) on a numeric response variable.

References

1. The Advanced Learner's Dictionary of Current English. Oxford. 1952; 1069. **Ref.:** https://goo.gl/K7pKvD
2. Farrugia P, Petrisor BA, Farrokhyar F, Bhandari M. Practical tips for surgical research: Research questions, hypothesis and objectives. J Can Surg. 2010; 53: 278-281. **Ref.:** https://goo.gl/Rf6DED
3. Niraula SR, Jha N. Review of common statistical tools for critical analysis of medical research. JNMA. 2003; 42:113-119.
4. Reddy M V. Organisation and collection of data. In Statistics for Mental Health Care Research. 2002; Edition 1: 13-23.
5. Lindman HR. Analysis of Variance in experimental design. New York: Springer-Verlag, 1992. **Ref.:** https://goo.gl/jXeec5
6. Niraula SR, Khanal SS. Critical analysis of performance of medical students. Education for Health. 2006; 19: 5-13. **Ref.:** https://goo.gl/5dFKUK
7. Indrayan A, Gupta P. Sampling techniques, confidence intervals and sample size. Natl Med Journal India. 2000; 13: 29-36. **Ref.:** https://goo.gl/1nbNpQ
8. Simon R. Confidence intervals for reporting results of clinical trails. Ann Int Med. 1986; 105: 429-435. **Ref.:** https://goo.gl/acDett