A purpose specific perspective on the statistical analysis for available data
On 17th October 2019, Minato Nakazawa

1. Data obtained by questionnaire

The data obtained by questionnaire is basically categorical data (factor variable). An important point is to clarify which of knowledge, attribute, attitude, behavior, or perception each question item addresses.
* To clarify the knowledge level, you should conduct a valid test. The test has to determine whether each answer is correct or wrong. If the test score obeys normal distribution, the score can be treated as continuous (numeric) variable.
* To clarify attribute, attitude or behavior, you have to carefully brush up the question sentence without any vagueness. Few exceptions such as age or sleeping hours can be regarded as continuous (numeric) variable, but usually most data of this kind is categorical variable.

For perception, Lickert scale (For the given sentence, how much the respondent agrees or matches is chosen from several levels of extent) is commonly used, because the perception cannot be assessed by any single question item, but by several interrelated question items. A common latent factor behind those question items can be regarded as the sum of scores in those question items. The levels of Lickert scale vary by the purpose, but usually 3 or 5 levels (3-points Lickert scale or 5-points Lickert scale) is used. For instance, in the case of 3-points, the answer should be chosen from (1) Disagree, (2) Neither disagree nor agree, (3) Agree.
Of course, the response for each item is ordered factor variable. However, the sum of the scores for several interrelated question items can be treated as numeric variable, if the scores are consistent enough (confirmed by Cronbach's alpha coefficient is larger than about 0.7, if not, the total score is not reliable as the scale and thus refactoring by factor analysis is needed).

For detailed explanation for Cronbach's alpha coefficient, please read
https://data.library.virginia.edu/using-and-interpreting-cronbachs-alpha/ or use alpha( ) function of psych package in R.

Example: http://minato.sip21c.org/advanced-statistics/cronbach.txt
This is the tab-delimited text data for 3 question items, each of which is 5-point Lickert scale with the same latent factor. To calculate Cronbach's alpha coefficient for these 3 variables, you can type as follows in R (">" is prompt, you don't need to type it, after "#" is comment).

```
> install.packages("psych", dep=TRUE) # Only once you need to install the psych package.
> library(psych)
> x <- read.delim("http://minato.sip21c.org/advanced-statistics/cronbach.txt")
> alpha(x)
```

Then you get the result of alpha coefficient for those 3 variables, which is 0.8 with 95% confidence intervals from 0.58 to 1.03. The fact that 0.8 is larger than 0.7 means those 3 questions' internal consistency.

Constructing questionnaire has to be conducted carefully, but you should refer the textbook for social research such as
https://nats-www.informatik.uni-hamburg.de/pub/User/InterculturalCommunication/top2.pdf
https://opentextbc.ca/researchmethods/chapter/constructing-survey-questionnaires/
http://www.fao.org/3/w3241e/w3241e05.htm

1-1. Statistical methods to analyze relationships between categorical variables

To analyze the relationships between 2 categorical variables, making cross table is usual.

In R, functions table( ) and xtabs( ) are commonly used.

The resulted cross table is a type of matrix object with table attribute. It's possible to test the independence between the 2 variables by Fisher's exact probability test (in R, fisher.test( ) function is available). To assess the strength of association, odds ratio or tetrachoric correlation coefficient can be used. Odds ratio can be calculated by the functions oddsratio( ) in fmsb package or oddsratio( ) in vcd package. Tetrachoric correlation coefficient can be calculated by the functions assocstats( ) in vcd package or polychor( ) function in polycor package.
If you have to assess the effects of 2 or more categorical variables (independent variables) on 1 categorical variable (dependent variable), there are 2 ways. One is, if you focus on 1 main factor, you can consider the strata composed by other independent variables and can check the consistent relationships between the main factor variable and the dependent variable by Cochran-Mantel-Haenszel pooled chi-square test and Mantel-Haenszel pooled odds ratio, the former by mantelhaen.test() function and the latter by epi.mh( ) function in epiR package. The other is, you can simultaneously assess the effects of independent variables on the dependent variable by logistic regression analysis. If the dependent variable is binary, usual logistic regression analysis is executable by glm( ) function with setting family=binomial(logit) option. If the dependent variable has three or more levels, multinomial logistic regression analysis can be applied using vglm( ) function in VGAM package with family=multinomial option. For logistic regression analysis, you have to show the indicator of goodness-of-fit, such as Nagelkerke's R^2 and AIC using the function NagelkerkeR2( ) in fmsb package and AIC( ), respectively.

To compare Lickert scales or other numeric scores between 2 groups, you can use t-test by t.test( ) function. If you compare such numeric scores among 3 or more groups, you can use one-way ANOVA (analysis of variance) by oneway.test( ) function or aov( ) function in R. The function oneway.test( ) doesn't assume equal variances over all groups, but aov( ) does.

In the final stage of the analysis, structural equation modeling is often used.

To confirm the reproducibility of the responses for a questionnaire (in other words, test-retest reliability) or the matches of evaluation by two or more different raters (in other words, inter-rater reliability), if the responses or evaluation is obtained as category, it's possible to make cross table. However, in this case, conducting Fisher's exact probability test doesn't make sense, because the 2 variables in concern are apparently **NOT** independent. Thus, instead of Fisher's test, whether the matches between 2 variables are significantly higher than randomly expected matches or not is in focus. For that purpose, Cohen's kappa coefficient can be used. If the 2 variables are completely matched, kappa=1 (if matches are same as random matches, kappa=0; if completely mismatch, kappa=-1). This can be done by Kappa.test( ) function in fmsb package.

If you want to know the significant changes of the responses for the same questionnaire by any kinds of intervention, whether the mismatches are asymmetric or not is in focus. For that purpose, McNemar's test (using mcnemar.test( ) for a matrix object in R) or Bhapkar's test (using bhapkar( ) function for 2 categorical variables in irr package) can be used.

2. For the data obtained by the experimental studies

In the experimental studies including clinical trials, categorical variables are usually given experimental condition. In toxicology test, binary variables like with/without the expression of toxic effect, with/without disease occurrence, life and death are used as outcome measure. Except for such variables, variables obtained in experimental studies are numeric measurements. For those, you have to pay attention to detection limit and precision of the measurement.

In the experimental study, analytical methods should be determined before getting result. Only if the adequacy of statistical analysis written in the experimental protocol (including the determination of sample size) is confirmed, the study plan can pass the investigation by the ethics committee. Frequently applied methodologies are mostly common such as:
* ANOVA or t-test in the efficacy of new drug
* Survival analysis (Kaplan-Meyer method, Log-rank test, Cox regression) for the time until the occurrence of outcome events, or probit/logit analysis for dose-response relationship to assess LD50/LC50/ED50 in toxicology
* Multiple regression analysis for the effects of 2 or more factors on quantitative outcome
* Repeated measures ANOVA for time-dependent effects of single exposure

Most of these analyses can be done by EZR, but dose-response relationship is not supported by EZR, and thus additional package such as drc (to calculate LD50, drm( ) and ED( ) functions are used) is needed.

3. For the data including both categorical and numeric variables obtained by health checkups or other field research

It's most difficult. Adequate statistical analyses largely vary.
At first, missing values are frequently included in the data obtained by fieldwork. After checking whether the missing occurred at random or biased, if at random, multiple imputation can be applied. For multiple imputation, mice and Amelia packages are readily available. If missing occurred in relation to other variables, the result of analysis is critically biased.

The most important thing is **to see the distribution of raw data before all other analyses**. Frequency bar chart can be drawn by barplot( ) function for categorical variables. Histogram can be drawn by hist( ) function for numeric variables. Normal QQ plot can also be drawn by qqnorm( ) function for numeric variables.

It's commonly seen in the analyses of health checkup data that is make the continuous quantitative raw data binary by fixed cutoff value such as categorizing blood pressure (mmHg) into hypertensive/normotensive, or categorizing waist circumference (cm) into with/without one factor of metabolic syndrome. However, if the distribution of raw data obeys normal distribution or other unimodal distribution, mechanical binarization may cause bias or decrease statistical power of analysis (in such case, the raw data should be analyzed as is, or categorize into 3 (low, middle, high) groups and compare the data only between low and high groups). If the distribution of raw data looks bimodal, such issue may not occur.

To draw the graph, there are many routine methods.
* Mosaic plot by mosaicplot( ) for cross table made from 2 categorical variables.
* Strip chart by stripchart( ) for the comparison of numeric data among categories (relationship between one numeric variable and one categorical variable).
* Box and whisker plot by boxplot( ) also for the comparison of numeric data among categories (relationship between one numeric variable and one categorical variable).
* Scatter plot by plot( ) for the relationship between 2 numeric variables. The effect of 3rd variable on the relationship between the 2 variables can be seen by specifying marker type or color.
* If you have to check the relationship among many numeric variables, pairs() function can be used to draw scatter plot matrix.

In health checkups, usually health status measurements are used as health outcome and the result of basic attributes and the questionnaire is used as factors affecting health outcome. Thus, multiple

regression analysis (lm( ) function) is commonly used. If you need to consider the effects of individual factors and group level factors simultaneously, mixed model (multilevel model) should be applied by lmer( ) function in lmerTest package.