Department of Statistics Malaysia

NEWSLETTER

www.dosm.gov.my

@StatsMalaysia

DOSM/BMP/5.2019/Series 14

HOW TO DEFINE DATA QUALITY

Quality measurement for statistical outputs is concerned with providing the user with sufficient information to judge whether or not the data are of sufficient quality for their intended use(s).

DIMENSION OF QUALITY

RELEVANCE

The degree to which the statistical product meets user needs for both coverage and content.

An assessment of relevance needs to consider :
- Who are the users of the statistics;
- What are their needs; and
- How well does the output meet these needs?

ACCURACY

The closeness between an estimated result and the (unknown) true value.

Accuracy can be split into sampling error and non-sampling error, where non-sampling error includes :
- Coverage error;
- Non-response error;
- Measurement error;
- Processing error; and
- Model assumption error.

TIMELESS & PUNCTUALITY

Timeliness refers to the lapse of time between publication and the period to which the data refer. Punctuality refers to the time lag between the actual and planned dates of publication.

An assessment of timeliness and punctuality should consider :
- The production time;
- Frequency of release; and
- Punctuality of release.

ACCESSIBILITY & CLARITY

Accessibility is the ease with which users are able to access the data. It also relates to the format in which the data are available and the availability of supporting information. Clarify refers to the quality and sufficiency of the metadata, illustrations and accompanying advice.

Specific areas where accessibility and clarity may be addressed include;
- Needs of analyst;
- Assistance to locate information; and
- Clarity and dissemination.

COMPARABILITY

The degree to which data can be compared over time and domain.

Comparability should be addressed in terms of:
- Comparabality over time;
- Spatial domains; and

(eg : sub-national,national,international)
- Domain or sub-population

(eg: industrial sector, household type)

COHERENCE

The degree to which data that are derived from different sources or methods, but which refer to the same phenomenon, are similar.

Coherence should be addressed in terms of coherence between data produced at different frequencies, other statistics in the same socio-economic domain and sources and outputs.

SOURCE : NATIONAL STATISTICS UK

#StatsMalaysia #mystatsday #myretailcensus2019 #hies2019 #isiwss2019 #mycensus2020 #LeaveNoOneBehind

HOW TO DETERMINE SAMPLE SIZE

1 Initial Sample Size :

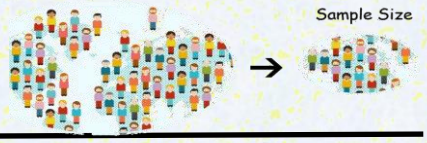
for estimating population mean:

$$n_0 = \frac{z^2_{\alpha/2} S^2}{d^2}$$

for estimating population proportion/ prevalence:

$$n_0 = \frac{z^2_{\alpha/2}\, p(1-p)}{d^2}$$

n_0 : initial sample size
$z_{\alpha/2}$: critical z score based on the desired confidence level
p : prevalence or proportion
s^2 : variance
d : desired margin of error

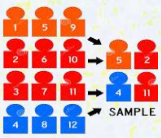


2 Adjust for the finite population correction factor:

$$n_1 = \frac{n_0}{1 + \frac{n_0}{N}}$$

n_1 : modified sample size after N been considered
N : population size

3 If the sample design is not a Simple Random Sampling (SRS), adjust the sample size for design effect :



$$n_2 = n_1 \times deff$$

where,

$$deff = \frac{Var_{COMPLEX}(\hat{y})}{Var_{SRS}(\hat{y})}$$

n_2 : sample size after taking into account design effect
Var_{COMPLX} : sampling variance of the complex sample
Var_{SRS} : sampling variance of the simple random sample

4 Finally, adjust for the response rate to determine the final sample:

$$n_{final} = \frac{n_2}{r}$$

n_{final} : final sample size after considered expected response rate
r : expected response rate

z score value

Confidence Level at 90% - 1.65
Confidence Level at 95% - 1.96
Confidence Level at 99% - 2.58

How many subjects do I need? "Neither 7 nor 30 nor any number is an all-purpose answer. A sample size of 30 is a "large sample" in some textbook discussions of "normal approximation"; yet 30,000 observations still may be too few to assess a rare event, but serious effect

— B. Jonathan

A statistical test provides a mechanism for making quantitative decisions about a process or processes. Sometimes it can be confusing even for a person with statistical background to understand the fundamental of statistical tests, and when to use which.

WHICH TEST SHOULD I USE?

TEST OF ASSOCIATION

Objective	Dependent variable	Independent variable	Parametric test	Non-parametric test
Relationship between 2 continuous variables	Scale	Scale	Pearson's Correlation Coefficient	Spearman's Correlation Coefficient
Predicting the value of one variable from the value of a predictor variable or looking for significant relationships	Scale	Any	Simple Linear Regression	Transform the data
	Nominal (Binary)	Any	Logistic Regression	–
Assessing the relationship between two categorical variables	Categorical	Categorical	–	Chi-squared test

COMMON SINGLE COMPARISON TEST

Comparing:	Independent variable	Independent variable	Parametric test	Non-parametric test
The averages of two independent groups	Scale	Nominal (Binary)	Independent t-test	Mann-Whitney test/ Wilcoxon rank sum
The averages of 3+ independent groups	Scale	Nominal	One-way ANOVA	Kruskal-Wallis test
The average difference between paired (matched) samples e.g. weight before and after a diet	Scale	Time/Condition variable	Paired t-test	Wilcoxon signed rank test
The 3+ measurement on the same subject	Scale	Time/condition variable	Repeated measures ANOVA	Friedman test

SOURCE : Chua, Y. (2006).Kaedah dan statistik penyelidikan. Kuala Lumpur: McGraw-Hill.