MATEMATIKA, 2008, Special Edition, Part II, 379 - 384
@ Department of Mathematics, UTM

THE ROLE OF SATISTICS IN BIOLOGICAL AND MEDICAL SCIENCES IN DEVELOPING COUNTRIES

[1]Asep Saefuddin, [2]Etih Sudarnika
[1]Vice Rector for Planning and Collaboration, IPB. Department of Statistics, Faculty of Mathematics and Science, IPB, 1660, Darmaga, Bogor, Indonesia
[2]Laboratory of Epidemiology, Faculty of Veterinary Medicine, IPB, 16680, Darmaga, Bogor, Indonesia
e-mail: [1]asaefuddin@gmail.com, [2]etih@ipb.ac.id

Abstract. Statistics is unavoidable in biology and medical research. It is due to the complexity of these areas involving large population, many factors related to outcome variables, measurement aspects, clustering, and other things that invite statistics to play its role in biology and medical sciences. This paper describes briefly statistical method in these areas encompassing basic/standard statistics, modeling application in estimating/predicting parameters, statistical conclusion, and recommendations. Due to biology and medical research characteristics, statistics play significant role in mixed effect models both linear and generalized linear models. Additionally, the experience of IPB (Bogor Agricultural University) in statistical education and research is presented. IPB is the oldest Department of Statistics in Indonesia has been contributing statistical concept originally in animal/plant breeding/quantitative genetics and now expanding to other biological sciences, epidemiology, and medical sciences. However, its contribution to molecular biology and geosciences is still limited. Hence, international collaboration in the area of geo and bioinformatics is required.

1 Preface

The importance of statistics in biology and medical science has been recognized since four centuries ago. John Graunt (1662) has recorded and counted the death rates in London and made an alert system at that time. Florence Nightingale has developed a mortality diagram called the "coxcomb", a pie chart that was used to describe the number of deaths in the military hospital that she run during the Crimean War. In the year 1858, she became the first woman chosen as a member of the Royal Statistic Society and later became an honorary member in the American Statistical Association. In the year 1854, John Snow has also used simple statistical analysis to outcome on the cholera pledge in London. (Wikipedia 2007)

Nowadays, statistical role cannot be separated from biological and medical research. Its role begins from research planning, such as research planning, sampling technique or experimental design, data analysis and research conclusions. As in the developed countries, the importance of statistics is also realized in developing countries. Although in a difference degree,. the use of statistical analysis has been implemented in biology and medical researches both in developed and developing nations. Nevertheless, there is still confusion in statistical procedures and interpretation. Even though statistics has already been introduced in every level of education (undergraduate/S1and graduate S2/S3) in every branch of sciences, many basic concepts are not well understood. To our knowledge, there is still confusion in understanding P - values, significance difference, confidence intervals, standard deviation and standard error. Hence, statistician is very important to be involved in the entire research activities from planning to

recommendation. In many cases, statisticians are only involved in analyzing the data affecting unreliable and insufficient data. In addition, many simple statistical procedures applied on a complex data.

This paper explores briefly the application of statistics on biological and medical science in developing countries. In this case, it includes basic statistical concepts, modeling application in estimating/predicting parameters and statistical conclusions. Also statistics role in research phase starting from study design, data collection and analysis, until interpreting results. In research activities, statistics must be treated as part of science, not only as tools to analyze data. However, although all statistical procedures applied correctly, statistical 'true' finding is considered temporary and spatially dependent, it is not the ultimate truth.

This paper also covers briefly the IPB's experience as the oldest Department of Statistics in Indonesian higher education. It originally was developed based on practical need to obtain conclusion of various treatment in agricultural experiments. It was not created from theoretical aspect. Indeed, Department of Statistics at IPB is the origin of Department of Mathematics and Department of Computer Science.

2 Statistics Basic Concept and Its Implementation

Knowledge in statistics is the key factor for a scientist in order to plan a research, i.e. Data collection, data analysis, and proper result interpretation. Statistics is not only applied in the beginning or ending of the research, but in each step of the whole research process. The following will describe the common problems found in biological or medical research in developing countries.

Study Design Phase

Researches must have a clear plan. Conceptualization of research plan is an important phase in the whole process. Statistics may contribute to formulation the research hypothesis, experimental design or sampling technique depending on the research type, data analysis, modeling application, hypothesis testing, and draw conclusion. In the whole process also includes data validity and reliability test, randomization, model diagnostics, outliers treatment and many other statistical techniques. In addition, besides many ideal statistical to obtain a desired research power, statistics provide some allowable exit to consider financial resource and time consideration. Hence, the research outcome still under statistical tolerance. In conclusion, it is very important for scientists to understand proper basic statistical concept to strengthen their research activities.

As an illustration, in a case-control study on the relation between physical activity and hypertension on women, the researcher must clearly define the type of the hypertension (pre hypertension, class I or class II) both in control and case groups. Mixture hypertension cluster will end up with unclear result. Considering the studies large range, the researcher must focus on the research and determine inclusive and exclusive criteria's. Also, if the research is about hypertension on women in productive age, then the sample must be limited to women in productive age (Saefuddin, 2007). However, to optimize time and financial resources, the cluster effect can still be accommodated in the statistic model with additional sample size to avoid overparameterisation (Soukri, 2000). The message is to correctly define the variables under study, measurement method and their units, and statistical relation among variables. In correlation test, when correlation value approaches 0, it is often concluded that there is no meaningful relation the two variables. Theoretically it must be translated that there is not enough evidence to prove the relation between the two variables.

Data Analysis Phase

In general data analysis includes simple and complex procedures. It is started by a simple one such as summary statistics to familiarize with the data. The commonly used charts are the block charts, histograms, polygons, line charts and pie charts. Then other more complex exploratory data analysis can

be implemented such as the t-test, z-test, the Mann-Whitney test, the simple ANOVA test, and the Kruskal-Wallis test (Sprent, 1998). To measure associations, the commonly used methods are the chi-square test and the Pearson and Spearmans coefficient correlation value. Simple linear regression model is also common in describing the correlations between variables.

In the application, these procedures are very important for researchers in order to understand and analyze the data as well as make some inferences for research conclusion. What needs to be noticed is a correspondence between the procedure used and the needs of the analysis. Simple statistical procedure usually assumes to have normal distribution with homogenous variance on each treatment, the effects of covariates (independent variables) are fixed, linear and additive. In reality, these assumptions are often not fulfilled, which results in the needs for other procedures likes transformation process.

The use of simple statistical procedures must fit the assumptions to provide answers to research questions. In a complex research, the usage of simple procedures is often not adequate. The type of outcome variables is very important as the model assumptions are heavily relies on error or residual reflecting the outcome. However, generalized models are available for non-normal distribution. For example, GLMs (Generalized Linear Models) are dedicated for larger class of distribution, which the exponential family distribution including the normal. Thanks to Nelder and Weddernburn (1972) who consolidated the concept of GLMs. Basically GLMs connects the outcome distribution and the predictors/covariates by a specific 'link' function. The approach is very useful in biology and medical sciences as their outcomes are usually not normally distributed.

Basic Result Interpretation

To avoid errors in interpreting analyzed data, researcher must understand the basic concepts of statistics, such as the significance values or P-values, significance difference and confidence intervals, and its link with the hypotheses. For example, refusing zero hypotheses in ANOVA and t-test are not the same. In an ANOVA, it means that at least one treatment that differs from the other treatments. Whereas in the t-tests, it means that both treatments are different. When zero hypotheses is accepted, it does not mean that "there is no difference between the treatments". The correct meaning is "there is not enough proof to show a difference in the treatments" which may be due to lack of data or high diversity.

Statistical power determines the quality of a research result. In statistics, there are two types of error, the type I error (α) and the type II error (β). Type I error is a case which rejects zero hypotheses in a condition where the zero hypotheses is right. Type II error is a case which accepts the zero hypotheses in a condition where the hypotheses is wrong. Statistical power shows the chance to reject the wrong zero hypotheses ($1-\beta$). The larger the statistical power, the higher the data validity.

Nowadays, many statistical software have been developed to help statistical calculations, both simple and complicated. But still, the existence of these soft wares cannot replace statisticians. It is still important for scientists to understand the statistical aspects in such reported research results.

More Advanced Statistics

The use of simple statistical procedures has been contributing significant roles in biology and medical research. Along with science development in general, biology and medical science are evolving into more complex sciences, which then require more sophisticated statistical procedures. For example, the logistic model is more frequent in epidemiology compared to the linear regression. Logistic regression is powerful to predict the probability of an incident in various conditions as well as the odds ratio (OR) of a risk factor. This model is a type of GLMs for the Binomial distribution with 'logit' as the link function. Although it is originally developed for dichotomous outcome, the multinomial logit is now available to analyze polytomous outcome (Hosmer and Lemeshow, 2000).

An illustration of a logistic regression application can be seen in the research of avian influenza in backyard farms (Sudarnika and Saefuddin, 2007). These researches were meant to find out the correlation between cultivating systems, farmer knowledge, access to education and bio-security acts and the infection of the H5N1 virus. By using logistic regression equations, it is possible to find out the risk

factors associated and how much they affect the infection of the virus. Hence, the outcomes were infected and uninfected individuals, while the covariates were cultivating system, farmer knowledge, access to education and bio-security.

A common problem found in the binomial outcome is overdispersion or extra-variation. Overdispersion is a condition where observed variance is higher than the expected variance (theoretical variance). If overdispersion is neglected, statistical test tends to reject the null hypotheses or some covariates have significant effect on the outcome. In other word, it will end up with a wrong conclusion. Overdispersion may be due to a random variation in the outcome probability, correlated outcome, and cluster effect. The simple method to overcome overdispersion is introduced by Williams (1982) by adopting the concept of weighted regression. Another technique is to introduce mixed model by combining fixed effect and random effect in the systematic factors (covariates). An excellent approach to handle overdispersion is presented by Collet (2002).

3 The Experience of IPB in Statistical Development in Indonesia

This section explores the summary of IPB's experiences in developing statistical education since its creation in 1965. The story is compacted based on my personal discussion with the founding father of Statistics Department at IPB, Prof. Nasoetion (1932-2002). In addition, several information obtained from classical statistical text book of Nasoetion, i.e. Matematika Mutakhir (1968), Aljabar Matriks (1972), Metode Statistika untuk Pengambilan Keputusan (1974), Landasan Matematika (1978), and Teori Statistika (1983).

IPB (Institut Pertanian Bogor) is an agricultural based university in Indonesia originally part of Univeritas Indonesia as Faculty of Agriculture and Faculty of Veterinary Medicine located in Bogor. In 1963 these faculties formed IPB with five faculties at the beginning (Agriculture, Veterinary Medicine, Fishery, Animal Husbandry, Forestry, and Agricultural Technology). The biometrics unit (created by Prof. Nasoetion, 1967) was under Department of Natural Sciences, Faculty of Agriculture). Prof. Nasoetion, the only statistician at IPB, obtained his PhD from North Caroline State University under Prof. Cockerham (1964).

In 1972 Prof. Nasoetion formed Department of Statistics and Computation under Faculty of Agriculture. The curriculum of statistics at IPB was originally developed to answer many questions related to agricultural treatment, such as experimental design (the analysis of variances) and dose-response analysis. Step by step it expanded to plant breeding for selecting the genetic merit, i.e. linear selection index model, regression analysis, variance component analysis. As agricultural education at IPB was getting broader including the concept of agribusiness, many statistics students took economy for their minor. Hence, late 1970 econometrics was developed followed by statistics for animal breeding in early 1980. In 1975 this department offered graduate program in applied statistics.

In 1980 Faculty of Mathematics and Natural Sciences was built with Department of Statistics and Computation is part of it. In 1981 it changed the name to be Department of Statistics (without Computation). The doctoral program was initiated in 1999. The students of graduate program came mainly from government research institutes and universities. Hence many statistiscian in Indonesia were the alumny of IPB, beside ITB (Bandung Institute of Technology), Universitas Padjadjaran and ITS (Surabaya Institute of Technology). In term of the application of statistics in biological and medical sciences, IPB is still dominating.

As the development of statistics at IPB was originally intended for designing research experiment and data analysis, the statistics students have to take some minor for its application. Initally, the minor was focused in biology, agronomy, or socio-economy depending on student interest. However, this year the minor may be taken from all departments at IPB. The alumni generally work in marketing research, plant/animal breeding research institutes, and educational institutions.

4 Future Agenda and Recommendation

Biological information can be arranged orderly from the smallest to the largest as the followings: DNA, mRNA, Protein, Pathways/Networks, Cells, Organs/Individual, and Population which related sciences respectively are genomics, functional genomics, proteomics, system biology, cellular biology, medicine, and public health sciences. Theoretically statistics can play role in all levels of biological information, although at the moment statistics involves strongly at the large level (organs, individual and population). Hence statistics in medicine and epidemiology are very advanced (Raftery et al, 2002). In developed countries, where the biotechnology is omnipresent, statistics is needed in interpreting genetic variability at molecular level. The combination among molecular biology, information technology and statistics produces bioinformatics (Pawitan, 2007).

In developing countries, many research issues still present at grouped individual or population level. Some important research agendas are listed below.

1. Geo-informatics

Statistics still plays an important role in analyzing time-space dependence data. Hot spot detection uses basic statistical concept on density, maximum likelihood, incident, regression models (fixed, random and mixed effect models), and many other statistical tools. In conjunction with information technology especially GIS (geographical information system) called geo-informatics, statistics is important in the area of poverty detection, environmental risk assessment, cause-effect and multifactor models, and bio-security.

2. Generalized Linear Models

Many outcome variables in biology and medical sciences are not in the continuous form and not-normal distribution. The generalized linear models (GLMs) of Nelder and Weddenburn (1972) will still be dominant in these areas. The basic idea is to create a link between the mean of outcome and the linear predictors (covariates) termed as link function. The common link function in biology and medical sciences are identity, binomial, and Poisson. Additionally, the technology in GLMs is getting well developed by incorporating GEE (generalized estimating equation) having working various types of working correlation matrix (Liang and Zeger,1986) to handle correlated data structures and overdispersion.

3. Statistical Evaluation in Modern Era

Many new technology in biology and medical science will certainly affect the statistical method. Reproduction technology in animal breeding such as embryo transfer from animals to other animals, embryo splitting, embryo sexing, and gene transfer may complicate statistical modeling. Mixed models equation of Henderson (1973) has to deal with other fixed effect due to certain gene transferred in combination with randomly additive gene effect. However, complex statistical computation is now handled easier due to the development in computer technology. Unfortunately, in developing countries, the main problem of data availability is still encountered. Hence, garbage in garbage out (GIGO) is the first thing to be solved.

As recommendations, statistical literacy must be introduced to all branch of science. Hence statistics is not treated as the tool per se, it is a required science to strengthen research methodology. Biology and medical sciences certainly need statistics as they deal with many data either as the responses or as the risk factors (covariates). Biology and medical sciences without statistics are incomplete. On the other hand, statistics students who will implement it biology and medical sciences need to comprehend modern biological concepts and terminologies. Statistics without understanding these concept (or other scientific domain) is dry.

References

Agresti A dan Finlay B. 1997. Statistical Methods for the Social Sciences 3rd edition. New Jersey: Prentice Hall.

Basuki B. 2000. Aplikasi Metode Kasus Kontrol. Jakarta: Bagian Kedokteran Komunitas FKUI

Collet D. 1991. Modelling Binary Data. London: Chapman & Hall/CRC

Dawson B and Trapp RG. 2004. Basic & Clinical Biostatistics 4th edition. Singapore: McGraw-Hill International

Henderson CR. 1973. Sire evaluation and genetic trend. Proc. Animal Breeding and Genetics Symposium in Honor of Dr.Jay L. Lush.10-41.

Liang KY and Zeger SL. 1986. Longitudinal data analysis using general linear models. Biometrika, 73(1), 13-22

Nelder, JA and Wedderenburn WWN. 1972. Generalized liner models. J. Roy. Statist. Soc., A. 135, 370-384

Pagano M and Gauvreu K. 1993. Principles of Biostatistics. California: Duxbury Press.

Pawitan Y. 2007. Statistics at the Frontier of Life Sciences. Annual Statistics Lecture. Departemen of Statistics Institut Pertanian Bogor, April 9th 2007.

Raftery AE, Martin AT and Martin TW. 2002. Statistics in the 21st Century. Chapman and Hall/CRC.

Saefuddin A. 2007. Socio-biostatistics: Historical Overview and Future Orientation. Proc. Second Joint Conference Indonesia Malaysia on mathematics and Statistics January 11-12, 2007 at ITS Surabaya, Indonesia.

Sprent P. 1998. Data Driven Statistical Methods. Chapman and Hall.

Sprent P. 2003. Statistics in Medical Research. Swiss Medical Weekly 133:522-529.

Sudarnika E, Saefuddin A, Zahid A. Basri C. 2007. the use of logistic regression model to identify the risk factor of H5N1 avian influenza virus of native chicken in Sumatera and Kalimantan Island, Indonesia. 2nd International Conference on Mathematical Sciences 2007 Universiti Teknologi Malaysia. May 27-28th 2007.

Wikipedia. 2007. Florence Nightingale. http://en.wikipedia.org/wiki/Florence_Nightingale. 20 April 2007.

----------. 2007. John Graunt. http://en.wikipedia.org/wiki/John_Graunt. 20 April 2007.

----------. 2007. John Snow. http://en.wikipedia.org/wiki/John_Snow. 20 April 2007.

Williams, DA. 1982. Extra-binomial variation in logistic linear models. Applied Stat., 31,144-148.