Journal of Science Research (2014) Vol. 13: 143-150

Statistics: A Quintessential Tool in Development of Scientific Theory

Shittu, Olanrewaju Ismail

Abstract
Science is a systematic enterprise that builds and organizes knowledge in the form of testable hypothesis, explanations and predictions about a phenomenon or the universe. The ultimate purpose of science is for growth and development of human conditions and through scientific research that employs the proper use and application of statistical methods. However, researchers usually take for granted a set of basic statistical assumptions and tests required to justify their method and results. This paper discusses the position of statistics in science, its relevance in the development of scientific theory. It provides brief explanations on correct use and appropriateness of some statistical techniques in research theory in order to achieve sustainable development. The concepts of the process of research and development of scientific theory were re-examined.

Key words: Science, Statistics, Development, Scientific theory, Hypothesis.

Introduction

In most developed nations, the issue of research and development ranks among the first five on the list of their priorities. In those countries, because the policy makers realize that no meaningful development in product or services, development has always been a product of intensive research. No meaningful project can take place without previous research into its feasibility or otherwise. It's no wonder then, that it took the Americans five years of intense research before they could dare to attack Osama bin Ladin [6]. Any firm that intends to control the market share in a highly competitive industry would have to invest heavily on research and development of its products. It is not mere coincidence that there now exist a department of research and statistics in most organized establishment, ministries and government parastatals (even though they do not allow experts to function in those capacities).

Literature Review

In the broadest sense research includes any gathering of data, information and facts for the advancement of knowledge. It may also refer to surfing the internet. Watching the news is also a type of research. Scientifically Research is defined as the act of performing a methodical study in order to prove a hypothesis or answer a specific question concerning a phenomenon. Finding a definitive answer is the central goal of any experimental process in research.

Research must be organized and systematic. It should follow a series of steps and a rigid standard protocol. These rules are broadly similar but may vary slightly between the different fields. Be it in scientific research, economic research, historical research or educational research. Scientific research must be organized and undergo some form of planning, including performing literature reviews of past research and evaluating what questions need to be answered.

Any type of research, whether scientific, economic or historical, requires some kind of organization, data collection, analysis, interpretation and an opinion from the researcher [13]. This opinion is the underlying principle, or question, that establishes the nature and type of experiment to be carried out in the course of the research.

Shittu, Olanrewaju Ismail
Department of Statistics, University of Ibadan, Ibadan
E-mail: oi.shittu@ui.edu.ng

ISSN 11179333

Development

Development is an indirect benefit of research. It is the use and application of research findings by government agencies, business men and individuals to bring about growth in the well-being of the people of a state or a nation and advancement of humanity socially, economically and technologically. The product of research can also bring about development of new product, improved production process and hence the growth of a manufacturing industry. Development can be said to be the product of intense research which may take many months and years to conclude. However, there is something between research and development - **STATISTICS**

Scientific Theory

While scientific law is a statistical proposition which has been found to be true through scientific hypothesis testing involving the application of statistics. Scientific hypothesis is a general proposition about all the things of a certain type. It is an empirical proposition in the sense that it is testable by experience. Scientific theory on the other hand is a well-substantiated explanation of some aspect of the natural world, based on a body of facts that have been repeatedly confirmed through observation and experiment. Scientists create scientific theories from hypotheses that have been corroborated through the scientific method, then gather evidence to test their accuracy. As with all forms of scientific knowledge, scientific theories are inductive in nature and do not make apodictic propositions; instead, they aim for predictive and explanatory force. A scientific theory is empirical, and is always open to falsification if new evidence is presented. No theory is ever considered strictly certain as science accepts the concept of fallibilism [16].

The strength of a scientific theory is related to the diversity of phenomena it can explain. It is measured by its ability to make falsifiable predictions with respect to those phenomena. Theories are improved as more evidence is gathered, so that accuracy in prediction improves over time. For instance, more reliable forecast are achieved if more recent information are observed in time series analysis. Scientists use theories as a foundation to gain further scientific knowledge, as well as to accomplish goals such as inventing new technology or curing a new disease such as Ebola. Development of scientific theory usually starts with working hypothesis, a provisionally accepted hypothesis proposed for further research. Such hypothesis have to be rigorously tested to yield scientific theories that are the most reliable, rigorous, and comprehensive. A theory that makes no observable predictions that can be sufficiently tested is not a useful theory and can hardly be applicable.

Development of Scientific Theory

A body of knowledge referred to as scientific theory should be capable of making falsifiable predictions with consistent accuracy across a broad area of scientific inquiry. It should be supported by many independent strands of evidence, rather than a single foundation to ensure good approximation, if not completely correct. It should be consistent with pre-existing theories and other experimental results. Its predictions may differ slightly from pre-existing theories in cases where they are more accurate than before. It should be adaptable and modifiable to account for new evidence as it is discovered, thus increasing its predictive capability of the theory over time. Development of scientific theories and practice are an axiomatic network of statements from which singular proposition can be derived and subsequently verified. The principles of induction and deduction and the process of acquiring data play prominent role in the development of modern scientific practice especially in experimental sciences. The combination of these three processes is referred to as the scientific learning process [9].

The three stages are interconnected such that one stage leads to another as (seen in Fig. 1).

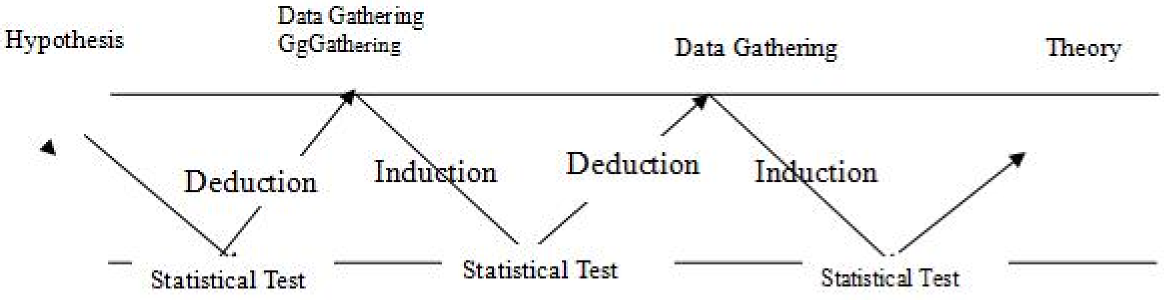

Fig. 1: Scientific learning process.

Induction and deduction are two different reasoning strategies in the development of scientific theory. They are important tools or instrument for prediction and explaining theories. Logical positivism otherwise known as inductive-hypothetico-deductive method is used to formulate general laws and theories. They provide a basis for formulating scientific hypothesis and establishing initial conditions for research designs. Thus, the ability to predict future conditions and formulate general laws and theories on the principles of deduction and induction is considered as the real power of scientific learning process. The process of acquiring data in scientific inquiry (Research) is by experimentation. It could be a walk into the library to unearth existing information; conduct a sample survey or carry out a laboratory experiment from which scientific theories can be developed.

Induction is the conscious mental process by which we pass from the perception of particular phenomenon (thing or event) to the knowledge of the general truth through experience (or data). Philosophically inductive reasoning can be described as a simple progression from particular/individual instances to broader generalizations. This is why it is referred to as “bottom-up” logic. It is a kind of reasoning that constructs or evaluates general propositions that are derived from specific examples. It derives its general principles from specific observations. Inductive reasoning consists of inferring general principles or rules from specific facts and is probabilistic in nature

Deduction is a method of drawing conclusions from one or more general statements (premises) to reach a logically certain conclusion. It is a process of deducing from something known or assumed to make inference or generalization about the particular. Deductive reasoning it is a "top-down" logical reasoning, is the process of reasoning [9]. Deductive reasoning links premises with conclusions. If all premises are true, the terms are clear, and the rules of deductive logic (hypothesis testing) are followed, then the conclusion reached is necessarily true.

The two reasoning strategies are synthetic reasoning used in scientific discovery, but the two do not share the same value or goals. While the theoreticians value induction, the empiricists value deduction, but the choice is based on taste rather than inherent superiority [9].

The three processes are interconnected such that one stage leads to another. Deduction stage is when hypothesis are developed with a view to testing with observed data. This leads to the process of data collection. The consequences of the hypothesis are compared with data. When consequences and data fail to agree, the discrepancy can lead by induction to modification of the hypothesis and a second cycle in the iteration is started again. The consequences of the modified hypothesis are worked out by deduction and again compared with data (old or newly acquired), by induction may lead to further modification and again of knowledge.

Our Experiences on Misuse of Statistical Methods

In the course of interaction or assisting colleagues in solving some rather serious statistical problems or helping to analyse their data, we have had some challenges.

There are occasions when supervisors insisted that the 'Null hypothesis' in the research question 'must' be rejected not minding the nature of the collected data or the measurement scale used in collecting the data. We have had irreconcilable data with the objectives of the research arising from ambiguous questions in the questionnaire prepared without proper consultation with the experts. For example, one may have two sample data from the same population where the researcher expects the analysis of variance table (ANOVA) to be constructed [12, 15] instead of students t-test. This is most common in the behavioural sciences and education.

In the natural sciences, we have colleagues confusing several 'iterations' with 'replication'. While iteration is a process of repeated computation of the parameters of a non-linear model each time applying the result of the previous stage, replication has to do with the repeated appearance of a treatment in a designed experiment. Replication helps to provide estimate of experimental error, improve precision of estimates and increase the scope of inference of the experiment. David L. Vaux, a Professor of cell biology at the Walter and Eliza Hall Institute of Medical Research, and at the University of Melbourne, Parkville, Victoria 3052, Australia, while commenting on an online paper [5], urged the experimental biologists, their reviewers and their publishers on the need to grasp the knowledge of basic statistics. He commented on the alarmingly high incidence of statistical mistakes in cell and molecular biology research papers—makes an important point—while also raising concerns about the credibility of some of the contents and said; "Consistent and well-articulated criteria for producing credible research are urgently needed".

Our Concern on the Use and Application of Statistical Methods

Statistics as a discipline has grown beyond what is known by many people. Many new statistical techniques capable of yielding more useful results have been developed. Some of them can be successfully applied in collaboration with academic statistician in experimental or physical sciences. For example, data modelling with its attendant drawbacks is now being replaced gradually with algorithmic modelling, while some software packages are now grossly inadequate for meaningful analysis (diagnosis and modelling) of assumed related variables. Researchers and colleagues in various fields should design, execute and validate research findings with appropriate statistical techniques. Non-professional statisticians teaching statistics in department other than statistics should teach well. There are cases of non-academic statistician writing books on statistical methods and application.

This is not to conclude that statisticians are averse to teaching and use of statistics and statistical methods by colleagues, we are only concerned about the correct use and application of statistics in solving research problems since our duty as academic statistician is to develop new tools and methodologies for use by other researchers in biological, physical sciences, engineering and medical fields. In some Universities, course system have never been strictly operated to the extent that statistics courses are taught by educationist or Social scientists in the Faculties of Education and the Social Sciences respectively simply because, they offered some statistics courses at 100 or 200 level during their projects at their undergraduate days. Statisticians are prepared to offer assistance to colleagues and researchers in forms of consultancy, collaboration in order to develop falsifiable, scientific theories for sustainable development.

Algorithms in Scientific Research

In response to the issue raised by Professor David Vaux above, here is the summary of the stages involved in statistical research which in my view is the most common statistical activities in applied research.

Irrespective of the kind of statistical model that is involved in research, one needs to go through the following major stages. The order and the specifics of how you go through each step will differ depending on the data and the type of model you use.

(a) Write out Research Questions in Theoretical and Operational Terms

Most times, researchers are confused about the right statistical method to use, but the real problem is how to define the research questions. They have a general idea of the relationship they want to establish, but have vague ideas about how to translate it to statistical questions. Researchers need to be very specific about the purpose of their research.

Define the Study Design

Whether one is collecting primary data or using secondary data, researchers need a clear idea of the design[1]. Design issues are about randomization and sampling:

(i) For survey data, there is need to
- Decide on the enumeration area
- Obtain the sampling frame
- Decide on the sampling technique (simple random sample or stratification or clustering)
- Determine the sample size
- Determine the analytical method

(ii) For an experiment (in the laboratory or on the field)
- Decode on the factors and factor levels
- Nested and Crossed Factors
- Potential confounders and control variables
- Longitudinal or repeated measurements on a study

(b) Choose variables and Determine their Level of Measurement

Every model takes into account both the design and the level of measurement of the variables, whether the variable is discrete or continuous, nominal, ordinal, or interval. It is absolutely vital to know the level of measurement of each response and predictor variable, because they determine both the type of information you can get from your model and the family of models that is appropriate.

(i) Write an analysis plan
Write the best guess for the statistical method that could answer the research question, taking into account the design and the type of data. It does not have to be the final plan but needs to be a reasonable approximation.

(ii) Calculate sample size estimations
There is need to obtain the sample sizes before collecting data and after the analysis plan.

(iii) Data Collection
Then the process of data gathering commences

(c) Prepare and Explore the Collected Data

After data collection, there is need to edit and clean the data, code and prepare the data for analysis

(i) Create new variables
It may be necessary to create new codes, categorize, reverse code and obtain new variables in their final form. This is especially useful in analysis involving principal components, factor analysis and time series analysis.

(ii) Run Univariate and Bivariate Statistics
Check the distributions of the variables to examine bivariate relationships among all variables that might go into the model. One might have to do some data manipulation or deal with missing data.

(iii) Run an initial model

At this stage one can run the model listed in the analysis plan. In most cases, this will not be the final model but it should be in the right family of models for the types of variables, the design, and to answer the research questions. There is need to have this initial model to have something to explore and refine.

(d) Refine Predictors and Check Model Fit

After data exploratory analysis and the initial model is available, one can use some sort of stepwise approach to determine the best predictors. If the analysis is to test hypotheses or answer theoretical research questions, there is need to

(i) Carry Out Assumptions and Validity Test

It is good to test the validity of the assumptions of the chosen model on the basis of which it may be possible to drop some variables, examine interactions and explore other types of analysis. Drop non significant control variables; do hierarchical modelling to see the effects of predictors added alone or in blocks; check for over dispersion and test the best specification of random effects, check for serial correlation with Durbin Watson test and Bartlet test[10].

(ii) Check for and resolve data issues

Data issues are about the data that occur within the context of the model. They are issues of multicollinearity, existence of outliers and influential observations, Missing data and Truncation and censoring. Once again, data issues do not appear until one have chosen variables and put them in the model.

(e) Interpret Results

Finally, interpret the results from the analysis. One may not notice any problem on data issues or misspecified predictors until the coefficients are interpreted [15]. It is possible to find something like a super high standard error or a coefficient with a sign opposite what you expected, sending you back to previous steps. Modelling process will be faster, easier, and make more sense if the above algorithm are properly implemented.

New Direction in Statistical Modelling

The goals in statistical modelling are to get information about the underlying mechanism that produced the data, extract information about the nature of relationship between the response variable and the input variables and to use data to predict what the response variable would be with future input variable. The current practice is to assume a stochastic model such as the linear or non-linear regression, logistic regression or the cox regression model that specifies the response variable as a function of response variables, random noise and the parameters [13]. This approach to modelling [4] is called the data model used by most academic statistician and modellers in other disciplines. The greatest plus of data modelling is that it produces a simple and understandable picture of the relationship between the input variables and responses. In the last two decade, there has been a shift from data model to another modelling culture [3].

The new culture is the algorithmic modelling which assumes that data generated are identically and independently distributed (i.i.d) from an unknown multivariate distribution. What is observed is a set of explanatory variables (x's) that go in, and a subsequent set of response variables (y's) that come out. The problem is to find an algorithm f(x) such that for future x in a test set, f(x) will be a good predictor of y. Algorithmic modelling has been used by industrial statisticians and psychometricians for decades [13]. Algorithmic models even though sometime complex and computationally more rigorous can give better predictive accuracy than data models. They also provide better information about the underlying mechanism. Most interesting and challenging statistical problems of both scientific and commercial nature in many fields have been solved with rapidly increasing ability of computers to store and manipulate large volume of data that are being generated even by minutes. Major advances in machine algorithmic modelling include machine learning, tree ensemble methods [2] neural networks, forest method e.t.c.

The problems with data modelling include its inability to handle large volume of data. [7] was quoted in his paper "_ _ _statistical experience suggests that it is unwise to fit a model that depends on 19 variables with only 155 data points available." Newer methods in machine learning thrive on large number of variables—the more the better. The model diagnostic tools being used (such as goodness – of- fit test and residual checks) in data modelling are now losing their powers. According to [10] goodness-of-fit tests have very little power unless the direction of the alternative hypothesis is precisely specified. The implication is that omnibus goodness-of-fit tests, which test in many directions simultaneously, have little power, and will not reject until the lack of fit is extreme. William Cleveland [11], one of the fathers of residual analysis, admitted that it could not uncover lack of fit in more than four to five dimensions.

There are occasions where two data modelling approaches were used on the same data set, confirm the model fit standard goodness-of-fit tests, residuals checks, coefficient of determination (R^2) etc., but yet the two models give different pictures of the mechanism that produced the data, different predictive power and led to different conclusions. A very good example is when deterministic model such as simple regression analysis is used for time structured data when it is most appropriate to use dynamic time series modes [5], [8]. Thus, misleading conclusions may follow even from data models that pass goodness-of-fit tests and residual checks. However, the only diagnostic tool for algorithmic models is the predictive accuracy. The issue of developing appropriate model diagnostic tool for algorithmic model is still an open research area for academic statisticians.

Finally, we are not against data models per se. In many situations they are the most appropriate way to solve the problem, but emphasis needs to be on finding alternative approach that overcomes the limitations that data models have posed.

Conclusions

This paper identifies the place of statistics as a discipline in the development of scientific theory and in the development of the society. The deductive and inductive scientific reasoning in relation to interconnected stages of scientific learning process were discussed. Our experiences and concern as academic statistician on issues of abuse of statistics and statistical techniques were also discussed. Statistics as a discipline has grown beyond what is known by many people. Many new statistical techniques capable of yielding more useful results have been developed.

Some of them can be successfully applied in collaboration with academic statistician. Researchers and colleague in various fields should design, execute and validate research findings with appropriate statistical techniques. Researchers and colleagues in various fields should design, execute and validate research findings with appropriate statistical techniques to formulate useful and implementable scientific theories for meaningful and sustainable development

References

[1] Adewoye, G.O. and Shittu, O.I. 1999. "Introduction to Socio-Economic Statistics (Survey Methods and Indicators)" *Victory Ventures, Lagos, Nigeria. Chapters (1-6,10-12) pgs. (1-94,124-151).* ISBN 978-33867-1-9.

[2] Breiman, L. 2000. Some infinity theory for tree ensembles. (Available at *www.stat. berkeley.edu/technical reports).*

[3] Breiman, L. 2001. Random forests. *Machine Learning J.* 45: 5-32.

[4] Breiman, L. 2001b. Statistical Modelling: The two Cultures, *Statistical Science* 16(3); 199-231.

[5] David, L.V. 2012. Research methods: Know when your numbers are significant, Comment on Nature 492: 180-181.

[6] Dedman, Bill. 2011. “How the US tracked couriers to elaborate bin Laden compound”. *msnbc.com*. *http://en.wikipedia.org/wiki/Search_for_Osama_bin_Laden*

[7] Diaconis, P. and Chatterjee, S. 2013. Estimating and Understanding Exponential Random Graph Models. *Ann. Statist.* 41(5): 2428-2461.

[8] FrancQ, C. and Zakoian, J. 2010. GARCH Model structure, Statistical Inference and Financial Applications, Willy-Interscience Publications, ISBN 978-0-470-68391-0.

[9] Johnson, T.L. 2000. Logical Background of Statistics and Decision Theory, Unpublished Lecture note.

[10] Tsay, R.S. 2006. Analysis of Financial Time Series, Second edition, Willy-Interscience Publications, ISBN-13-978-0-471-69074-0.

[11] William, S. Cleveland. 1994. The Elements of Graphing Data, Hobart Press.

[12] __________. 2008. The Social Indicator planning and evaluation (SIPES) for Nonpoint source management. Projects in USEFPA Region 5 Version 2.

[13] __________. 2011. Science, - *en.m. wikipedia.org/wiki/Science*

[14] __________. 2009. Building a statistical model, *cp.literature.agilent.com/litweb/pdf/iccal2006/icstat/icstat012.html*

[15] __________. 2013. The analysis – *www.theanalysisfactor.com/13-step-regression-anova/*

[16] __________. 2013. *http://www.iep.utm. edu/ fallibil/*