

Statistics: Its Utility, Its Risks

John Y. Wu

Pathology Department, Visalia Pathology Medical Group, Visalia, USA

Email: wutulare@yahoo.com

Received 18 August 2014; revised 16 September 2014; accepted 6 October 2014

Copyright © 2014 by author and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Since natural science is our most precise and careful method of investigation of reality, based on our most precise thinking, mathematics, statistics as the mathematical point of contact with reality is the spearhead of scientific investigations—points of reality-contact—such as physics and engineering. This essay briefs technical calculations of statistics, as it focuses on the vast significance of statistics by One, citing two concrete examples to show its utilitarian power penetrating the concrete situations, and then Two, our cautions on its use, some quite serious.

Keywords

Statistics, Utility, Risks

1. ONE: Two Concrete Cases of Statistical Investigations

In order to portray what statistics is and how it operates, here are two concrete cases, sampling from a population and making a difference healthwise.

2. Case One: Making Inferences with Caution: Sampling from Populations as a First Step

Drawing conclusions from data, making generalizations from specific instances, and making the seemingly unpredictable predictable is made possible by statistics—but only when it is used with great caution. Critical thinking regarding the proper and valid steps necessary in making inferences from data is indispensable; judicious sampling from a population is the crucial first step. I will define the terms population, sample, and discuss how they are related. Careful sampling and its implications for research are then discussed.

2.1. Population and Sampling

First, some definitions must be made. The statistician must define the population to be studied, for instance all

the fish in a pond, in a state, or in the world. A population is the complete collection of all elements (scores, people, measurements, and so on) to be studied (Triola, M. M. & Triola, M. F., 2006). The collection is “complete” when *all* subjects to be studied are included. A population is the set of all measurements (or record of some qualitative trait) in the entire collection of units about which information is sought (Johnson & Tsui, 1998). A unit is a single entity, usually a person or an object, whose characteristics are of interest. The population of units is the complete collection of units. A census is the collection of data from every member of the population. A measurement of a characteristic of a population is called a parameter (Triola, M. M. & Triola, M. F., 2006).

A sample is a subcollection of members selected from part of a population. A statistical population is the subset of measurements collected in the course of an investigation (Johnson & Tsui 1998). A measurement of a characteristic of a sample is a statistic (Triola, M. M. & Triola, M. F., 2006). The goal in obtaining a sample is to make inferences about the population; the better the sample the more closely related it is to the target population. For example, a study on the effects on the bodies of 10 mice eating refined sugar (sample) extrapolated to the effects on the bodies of all humans eating refined sugar is less correlated than a study on a sample of 10 persons who eat refined sugar compared to the entire world population who eat sugar. In other words, mice are not the same as humans, so the former study is less representative of the human population than the latter study.

2.2. Types of Sampling

Five methods are used to obtain samples from a population: voluntary response, stratified sampling, cluster sampling, random sampling and simple random sampling. The latter are methods of increasing the chance that a sample is representative of a chosen population. An explanation of each method follows.

“Voluntary response” or self-selected sample is one in which the respondents themselves decide whether to be included (Triola, M. M. & Triola, M. F., 2006). But this method is not representative of a population. For instance, a host of a radio music show announced that she wanted to know which singer is the favorite among city residents. Listeners were asked to call in and name their favorite singer. Those who listen to the particular radio station are already a special subgroup with a preference for the same type of music. Additionally, those listeners calling in are those who usually feel strongest about their opinions.

“Stratified sampling” separates the population into subgroups based on same characteristics (e.g. sex or age), and then samples from all subgroups. “Cluster sampling” separates the population not based on characteristics and then chooses all the members from randomly selected clusters.

“Random Sampling” and “Simple Random Sampling.” Rather than the previous sampling methods mentioned above, statisticians usually choose random sampling and simple random sampling in order to increase the chance that a sample is representative of a chosen population. “Random sampling” from a population is done when each member from the population has the same chance of being selected. For example, an urn with ten ping pong balls numbered 0 through 9 and shaken thoroughly. Then one is drawn and the digit recorded. It is then replaced, the balls shuffled, another one drawn, and the digit recorded. A computer can closely simulate this procedure. In a “simple random sample” a size of n subjects is selected in a way that every possible outcome of size n has the same chance of being chosen.

Furthermore, there are two types of randomized design in research: completely randomized design and randomized block design. In “a completely randomized design” the subjects are chosen entirely randomly to receive treatments. An example is an experiment to see the effect of diet on cholesterol levels. Persons are chosen to undergo a vegetarian diet or a meat diet on the basis of their social security number being an even or odd number. The outcome of an even number results in a vegetarian diet and the outcome of odd number results in a meat diet.

In “a randomized block design”, before subjects are randomly chosen to receive treatments, they are separated into blocks or groups. For instance, before persons are assigned to a diet they are separated into two groups, males in one group and females in another. Then within each group subjects are chosen to eat either a vegetarian diet or meat diet based on their social security number being even or odd. If diet does effect cholesterol levels we can see the results in the randomized design. If diet has an effect and if sex also has an effect on cholesterol levels we can see this in the randomized block design.

2.3. Conclusion

Sampling from populations is a powerful tool to help us begin to understand the unknown in our everyday lives.

But doing so requires caution in defining the population and selecting the sample. In any case, done with caution, sampling from populations is a powerful tool to make inferences and understand the unknown population in our everyday lives. Defining the population to be considered and the manner in selecting the sample are the critical first steps in the statistical approach to understanding reality.

3. Case Two: Making a Difference Healthwise: An Evaluation of Two Investigations

Two articles investigate the association if not causal relation of two factors on health, environmental and increased drug dosage. The present paper evaluates the methods of doing so. First, we discuss the population researched, data collected, dependent and independent variables, in both articles. Secondly, we explore the connection between probability and dependent variables. Probability is multidimensional. Tools used to express probability in each study will be discussed: we discuss relative risk, odds ratio and fetal death rates in the article “Fetal Deaths and Proximity to Hazardous Waste Sites in Washington State” (Mueller, Kuehn, & Shapiro-Mendoza, 2007); and absolute risk reduction and number needed to treat in the article “Intensive Lipid Lowering with Atorvastatin in Patients with Coronary Artery Disease” (Shepherd, Kastelein, & Bitfner, 2008). These statistic tools support their hypotheses that the presence of environmental and dosage factors affect the probability of subsequently selecting a healthy subject from the population sampled, i.e., these factors are, in fact, dependent variables.

3.1. Populations Researched, Data Collected, and Dependent and Independent Variables

The population or complete set of all measurements from which information is sought (Johnson & Tsui, 1998) is the Washington State vital records from 1987 to 2001 in the case control study “Fetal Deaths and Proximity to Hazardous Waste Sites in Washington State.” Cases were women with fetal deaths at 20 weeks gestation ($n = 7054$). Ten controls per case were randomly selected from live births. Other data collected in this study were the locations of 939 hazardous waste sites identified from the Department of Ecology registry. Distances were measured from maternal residence at delivery to the nearest hazardous waste sites. “Independent variable” occurs when the occurrence of one event does not affect the probability or outcome of a subsequent event. If a variable is not an independent variable it is a dependent variable (Triola, M. M. & Triola, M. F., 2006). In this fetal deaths article no associations were found (independent variables) for any proximity categories (0.5 miles relative to >5 miles) from sites with contaminated air, soil, water, solvents or metals, but fetal death risk increased among women residing = 1 mile from pesticide-containing sites. These sites were the dependent variable, while independent variables were contaminated air, soil, solvents, and metals. Both conclusions will be discussed.

The population sampled in the randomized double blind study, “Intensive Lipid Lowering With Atorvastatin in Patients With Coronary Artery Disease, Diabetes, and Chronic Kidney Disease,” were 10,001 patients with coronary artery disease that were treated with 8 weeks’ open-label therapy with atorvastatin (10 mg/d). They were randomly selected to receive either high dose (80 mg/d) or low dose (10 mg/d) of atorvastatin between July 1, 1998 and December 31, 1999; neither the investigators nor the patients knew who received either of the two dosages until the end of the study (double blind randomization).

Data collected were of 1501 patients with diabetes, 1431 had renal data available. Patients with chronic kidney disease (CKD) were defined as having a baseline glomerular filtration rate (GFR) of <60 mL/min per 1.73 sq meters using the Modification of Diet in Renal Disease Equation. Comparison of patients with CKD, stable coronary artery disease, and diabetes, who received low dose atorvastatin (10 mg/d), with patients of similar medical histories who received high dose of atorvastatin (80 mg/dl), indicated that patients with high dose of atorvastatin experienced marked reduction (35% reduction in relative risk) in cardiovascular events. In contrast, patients with no CKD, stable coronary artery disease, and diabetes, had a much lower reduction (only 10% reduction in relative risk) in cardiovascular events with high dose of atorvastatin, compared with low dose of atorvastatin administered.

The patients had a median follow-up of 4.8 years. The dependent variables among this group of patients were high dose of atorvastatin and the presence of CKD. The independent variables were low dose of atorvastatin and lack of CKD. How these variables were categorized follows.

3.2. Expressions of Probability and Dependent Variable

Certain expressions of probability were used in the two articles cited. These expressions are used in support of

the claim identifying dependent variables investigated in those studies.

Probability and Comparing Probabilities

Let us define how to find the probability of event A, $P(A)$, in more detail (Triola, M. M. & Triola, M. F., 2006), first in three approaches to find single probability value, and then three more approaches to compare probability values.

1) Relative Frequency Approximation of Probability. Conduct (or observe) a procedure, and count the number of times that event A actually occurs. Based on these actual results, $P(A)$ is *estimated* as follows: $P(A) = \text{number of times A occurred} / \text{number of times trial was repeated}$, e.g., trying to determine whether a thrown tack lands up $P(A)$ or down. The tack landing point up does not have the same chance as landing on its side. When trying to determine the probability of the tack landing point up we must repeat the procedure of tossing the tack many times and then find the ratio of the number of times the tack lands with the point up to the number of tosses.

2) Classical Approach to Finding Probability (Requires Equally Likely Outcomes). Assuming a procedure has n different simple events each with the same chance of occurring, $P(A)$ equals number of outcomes A can occur divided by the total number of possible outcomes. For example, rolling a 1 on a fair die, each of the six faces having an equal chance of occurring, $P(A) = 1/6$.

3) Subjective Approach to Finding Probabilities. $P(A)$ is estimated by using knowledge of the relevant circumstances, e.g., trying to estimate the probability of rain tomorrow.

The above approaches to finding probabilities are for a single probability value. However, a single probability value is less meaningful than comparing probability values. Comparing probabilities gives us some sense of whether there is an apparent difference, whether a substantial difference exists between the two proportions¹.

We have three additional ways of *comparing* probability. They are, one, odds ratio, two, absolute risk reduction, and, three, number needed to treat.

One, since data obtained cannot always be used to determine relative risk, another method is used, the “odds ratio,” the probability of an event to occur divided by the probability of its complement (the probability of all outcomes in which A does not occur, or $1 - P(A)$). The odds in favor of an event A is the probability of event A divided by the probability of its complement \bar{A} . The odds against event A to occur is the reciprocal of (1 divided by) the odds in favor for event A to occur.

In the article on fetal deaths the odds ratio for fetal deaths for women residing 0.5 miles relative to >5 miles from a hazardous waste site was 1.06, whereas the odds ratio for fetal deaths increased for women residing =1 mile from pesticide-containing sites to 1.28. An odds ratio close to 1 means the exposure does not affect the odds of outcome. An odds ratio of greater than 1 means the exposure is associated with higher odds of outcome (Szumilas, 2010). An additional tool in the article was fetal death rates, calculated as the number of deaths per 1000 people per year, but in the article it was given as $n = 7054$ fetal deaths at 20 weeks (the Washington State vital records from 1987 to 2001).

Two, the “absolute risk reduction” is the probability of an event A in the treatment group minus the probability of the event A in the control group. The “number needed to treat” is related to the “absolute risk reduction.” *Three*, the number needed to treat is the number of subjects needed that must be treated to prevent one event, such as a disease or adverse reaction (Triola, M. M. & Triola, M. F., 2006). It is calculated by dividing 1 by the absolute risk reduction.

In the article on intensive lipid lowering, the number needed to treat was 14 to prevent 1 major cardiovascular event over 4.8 years in patients with diabetes, stable coronary artery disease, and mild to moderate CKD. Absolute risk reduction equals 1 divided by the number needed to treat. The absolute risk reduction was therefore $1/14$. It is a large number highly suggestive of the factor of high dose of atorvastatin being a dependent factor on cardiovascular health.

3.3. Conclusion

Statistical expressions of probability can effectively be used to link variables to effects on health, to convince

¹There are more effective methods for determining whether the apparent difference is actually **significant**, such as the P value method. **Relative risk** (or risk ratio) is the proportion (or incidence rate) of some characteristic in a *treatment group* divided by the proportion (or incidence rate) of some characteristic in a *control group*.

the reader that these variables are dependent factors. The two articles cited show the relationship between environment and mortality, and an increased dosage on cardiovascular events. In other words, proximity to pesticide-containing sites was associated with an increase in fetal deaths, and high dose atorvastatin was associated with fewer cardiovascular events than low dose atorvastatin in patients with diabetes and chronic kidney disease. Given the appropriate selection of population and careful selection of data, expressions of probability and dependent variables can show a strong correlation. However, caution is advised on stating causality in such correlations.

4. TWO: Risks on the Use of Statistics

This conclusion continues the two conclusions of the above two concrete cases of statistical investigations. Here we continue the evaluation of the whole situation of statistical forays of science into reality. As the conclusion to concrete case No. 1 says, sampling from populations is a powerful tool to help us understand the unknown in everyday life, but we must be careful in defining the population and selecting the sample. As the conclusion to concrete case No. 2 says, probability can be seen to link variables as dependent factor on health, such as pesticide environment and fetal mortality and increased dosage of atorvastatin on cardiovascular events, though probability may or may not be causal.

Now, after summing up conclusions above this way, we further see five warnings on the risks of using statistics quite powerful, all quite sober, fundamental, and radically serious, one, human errors that cannot be stamped out infests science and statistics, two, mathematics our apodictic certainty that sires statistics is bankrupt at the base, three, science crafted by human hands and brains is as fragile as humanity, four, contingency the unpleasant surprises stays against statistical understanding of science, and five, reality science and statistics try to reach is never reachable at all.

One, we being human, errors, inappropriate sampling, and/or miscalculations do occur and can never be totally rooted out. There is yet to be statistics on statistics to study the pesky situation, and then immediately we get caught in statistics on statistics on statistics, and so on, ad infinitum. This infinite regress indicates a total impossibility of obtaining an ideal statistics.

Two, science is through and through infested with paradigm revolutions throughout its history (Kuhn, 1996). Paradigm includes standard of validity and definitions of validity and its proofs; when paradigm is revolutionized, the whole science is revolutionized. Kuhn carefully describes this radical whirlwind *inside* science throughout its history, but is oddly, and sadly, silent on its vast devastating significance for the whole scientific investigations. He simply, optimistically, assumes scientific “progress” akin to, even as part of, biological evolution. Scientific revolution is life-evolution for him; he offers no convincing reason for his optimism, on whose assumption the whole book proceeds.

Each alleged satisfaction obtained with a new paradigm is only to be overrun later by a new dissatisfaction to be satisfied by another “new” paradigm, and so on, and this series of dissatisfactions followed by satisfactions may or may not be a “progress”; it is humanly impossible to produce the decisive super-rationale to warrant this series—if this is a continuous series, not many haphazard stop-and-gos—as a real “progress.”

Three, this radically instability of science is matched by the radical bankruptcy in mathematics. In Gödel’s eternally other-dependence of demonstration (Dawson, 1997) (at the base of science), Gödel himself gets caught in the liar’s paradox (“I am a liar” can not be believed or not-believed), since he independently proves the total dependence of all proofs. Let me explain.

Someone says, “I am a liar”. If we take what he says as true, then he says truthfully, and what he says is false, for he is not a liar as he says he is. If we take what he says is false, then he says falsely, and what he says is true; he is a liar saying about himself falsely, as he claims he is a liar. If his saying is true, it is false; if it is false, it is true. In this manner, then, his saying is senseless, and the liar’s paradox destroys the self-liar.

Gödel proves, alone, that every proof must be proved by other proof than itself. If we take his proof as valid, then his proof done alone is invalid, not proved by other proof as its conclusion says; if we take his proof as invalid, then his proof is valid because it is not proved by other proof as his conclusion requires. If his proof is valid, it is invalid; if it is invalid, it is valid. Gödel is caught in a liar’s paradox; his incompleteness theorem demolishes itself. Beautiful precision of mathematics calculates to self-futility. The precision of mathematics and science based on it is bankrupt at the root.

In Gödel’s eternally other-dependence of demonstration (Dawson, 1997) (at the base of science), Gödel him-

self gets caught in the liar's paradox ("I am a liar" can not be believed or not-believed), since he independently proves the total dependence of all proofs. The precision of mathematics and science based on it is bankrupt at the root.

Four, the world out there always has unpleasant surprises in concrete happenings that violate our neat and clean "general picture" that statistics carefully crafted of the situation. We call these surprises "contingency". Contingency is what statistics purports to handle, and contingency is what wrecks statistics.

Five, science and mathematics are created by human hands and brains, while reality is *not* created by human hands and brains. Therefore, science and mathematics can never be sure that what they have calculated and discovered are indeed "real" or not. We need, yet can never have, the third party no human not real to compare human discoveries here and reality out there, and check on their correspondence, and then, again, we need (and can never have) the independent fourth party to check on the third party, and then the fifth on the fourth, and so on, to check on the checking on the checking, ad nauseam.

Actually, the situation here is even much more hopeless, much more desperately complex, in that each "party" has two connections on both ends, each of which requires an outside "party" to check, and the more numerous "parties" we get, the twice more as many "parties" we need to check, ad infinitum. The n th checking requires $2n$ outside "parties" to check, none of which can be obtained. Clearly, this is an alarming increase of impossibilities quite endless. Again, such infinite regress of impossibilities indicates the ugly ditch ever gaping between human science here, on one hand, and reality out there, on the other.

Now, these five alarming cautions do not, however, counsel us to abandon statistics, for it is our sole effective scientific tool with which to study the "reality" we see voluptuously proliferating around us. Statistics gathers, analyzes, and makes inferences from gathered data. Such operations apply to all the scientific investigations. Furthermore, statistical tools not only summarize past data by indicators of the mean, median, and mode, and the standard deviation, but serve to predict future events by using frequency distribution functions.

In this way, statistics supplies designs of efficient experiments to save time-consuming trial and error. As above concrete cases show, double-blind tests for polls, intelligence and aptitude tests, and medical, biological, and industrial experiments all benefit from statistical methods and theories. The results of all of them serve as predictors of future performance, though reliability varies. All such operations are related to estimation, hypothesis testing, least square method, probability, regression. The above cited cases just scratch the surface of the vast iceberg of statistics.

Thus without statistics, we would have been totally helpless as we are confronted with the data overwhelmingly abundant. Statistics in this way remains our sole sharpest tool science has on hand, so sharp and effective in fact that we are often induced to wholly rely on it as if it were our god-given favorite. In this situation, we must remember. A good servant is a disastrous mast. Statistics is an aid to science, never its aim. The above five dire disastrous warnings counsel us the scientists to use statistics with extremely sober caution, without illusion whatever.

References

- Dawson Jr., J. W. (1997). *Logical Dilemmas: The Life and Work of Kurt Gödel*. Wellesbey, Massachusetts: A K Peters.
- Johnson, R. A., & Tsui, K. (1998). *Statistical Reasoning and Methods*. New York, New York: John Wiley & Sons, Inc.
- Kuhn, T. S. (1996). *The Structure of Scientific Revolutions*. Chicago, IL: The University of Chicago Press.
<http://dx.doi.org/10.7208/chicago/9780226458106.001.0001>
- Mueller, B. A., Kuehn, C. M., & Shapiro-Mendoza, C. K. (2007). Fetal Deaths and Proximity to Hazardous Waste Sites in Washington State. *Environmental Health Perspectives*, 115, 776-780. <http://dx.doi.org/10.1289/ehp.9750>
- Shepherd, J., Kastelein, J. J., & Bitfner, V. A. (2008). Intensive Lipid Lowering with Atorvastatin in Patients with Coronary Artery Disease, Diabetes, and Chronic Kidney Disease. *Mayo Clinic Proceedings*, 83, 870-879.
[http://dx.doi.org/10.1016/S0025-6196\(11\)60763-5](http://dx.doi.org/10.1016/S0025-6196(11)60763-5)
- Szumilas, M. (2010). Explaining Odds Ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 19, 227.
- Triola, M. M., & Triola, M. F. (2006). *Biostatistics for the Biological and Health Sciences*. Boston, MA: Pearson.

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or [Online Submission Portal](#).

