Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Big data and a bewildered lay analyst

Limsoon Wong

School of Computing, National University of Singapore, 13 Computing Drive, Singapore 117417, Singapore

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Hypothesis testing
Hypothesis exploration
Data analysis tactics
Exception
Trend reversal
Trend enhancement

ABSTRACT

Lay analysts often test hypotheses incorrectly. They also need help to find interesting hypotheses. They usually do not know what to do next after testing an initial hypothesis. We discuss their common mistakes, and also suggest practical tactics for their problems.

1. I am a bewildered lay analyst

Many lay analysts are involved in analyzing data, to hopefully produce actionable insights. Unlike professional statisticians who have the benefit of many years of rigorous training and more years of practising and perfecting the art of data analysis, lay analysts – like me, a computer scientist – have rather ad hoc training. Our training in statistics, if any, tends to consist of learning the mechanical application of e.g. a statistical hypothesis testing method, like how to invoke it from a statistical software package or how to program it up in some software we are developing.

Lay analysts can be cavalier about doing hypothesis testing. After all, the statistics text books – the practical kind that we read any way – tend to just give a plain description defining a test statistic and the associated nominal null distribution of the statistical test, and show some straightforward examples in which it is used. There is hardly any discussion on the conditions that must be met in order for the test to be valid. There is hardly any discussion on how statistical experiments should be designed so that those conditions are met. In any case, we have never seen the required conditions ever being checked in the examples shown in these statistics text books.

So we do little checking when we run our statistical tests. Sometimes, we even state our null and alternative hypothesis incorrectly. Consequently, when a null hypothesis is rejected, we often accept the alternative hypothesis without realizing that it is conditioned on all assumptions being satisfied. And, in practice, when a null hypothesis is rejected according to the test, it is rejected for a variety of reasons other than the alternative hypothesis we have in mind, leaving us with an incorrect conclusion.

Often, a lay analyst does not even have a hypothesis to start with. A hypothesis is usually inspired by some frequent patterns observed in some datasets. This has motivated computer scientists like me to develop many data mining methods (Han et al., 2012) for identifying frequent patterns from large datasets. However, we have mostly focused on scaling issues (Jagadish et al., 2014), and much less attention on analyzing the thousands of patterns returned by our data mining methods. Thus, a lay analyst looking at the output of a data mining system usually gets little support in selecting patterns to analyze.

The lay analyst also gets limited help from the data mining system in investigating the selected patterns in greater depth. Moreover, statistics text books tend to illustrate statistical hypothesis testing with straightforward examples. These text

E-mail address: wongls@comp.nus.edu.sg.
URL: http://www.comp.nus.edu.sg/~wongls.

https://doi.org/10.1016/j.spl.2018.02.033
0167-7152/© 2018 Elsevier B.V. All rights reserved.

books seldom discuss analysis tactics that experienced statisticians and analysts have accumulated over the years. The lack of exposure to a rich body of analysis tactics is another obstacle to getting insight from data by a lay analyst.

In this article, some common mistakes made by lay analysts are discussed, and some practical tactics to help them come up with interesting hypotheses and derive deeper insight from these hypotheses are suggested.

2. Am I testing this hypothesis correctly?

Statistical hypothesis testing is central to data analysis. However, it is not straightforward to carry out statistical hypothesis testing correctly. Three types of common mistakes made by lay analysts are described below.

2.1. Not ensuring that the samples are fidel to real-world populations

Consider this toy genotyping dataset of a mutation site rs123, with two alleles A and G, in patients suffering a disease X and healthy control subjects: 1, 38, and 69 control subjects have the AA, AG, and GG genotypes respectively, while 0, 79, and 2 patients have the AA, AG, and GG genotypes respectively. A χ^2-test is highly significant on this dataset. Thus, a lay analyst concludes that rs123 is associated with disease X, and that a subject who has the AG genotype is likely to get the disease.

Here the lay analyst probably has in mind the null hypothesis "The distribution of the rs123 AA/AG/GG genotypes in the disease X *population* is identical to that in the healthy *population*". Thus upon its rejection, he accepts the alternative hypothesis "The distribution of the rs123 genotypes in the disease X *population* is different from that in the healthy *population*".

Actually, in the χ^2 test above, only a *sample* of the disease X population is compared to a *sample* of the healthy population. Hence any conclusion at the population level is subject to the absence of sampling bias. In other words, the significance of the χ^2 test on this dataset could be due to the distribution of rs123 genotypes in the disease X population being different from the healthy population, or it could be due to the distribution of rs123 genotypes in the observed disease X sample being different from the disease X population, or it could be due to the distribution of rs123 genotypes in the observed control subjects being different from the healthy population.

In order to ensure that the significance of the χ^2 test is due to the distribution of rs123 alleles in the disease X population being different from the healthy population, a careful statistician would validate the two assumptions on the absence of sampling bias. For this specific example, the laws of human genetics can be used to check the absence of sampling bias, as follows. In the given dataset, 62% of the subjects has the AG genotype. Under the assumptions on the absence of sampling bias, 62% of the entire population also has the AG genotype. By the laws of human genetics, when both parents of a person have the AG genotype, there is a 25% chance of that person having the AA genotype. It follows that at least $62\% * 62\% * 25\% = 9.6\%$ of the combined population has the AA genotype. In the given dataset, less than 1% of the subjects has the AA genotype, far less than 9.6%. Therefore, the dataset is likely biased, unless the AA genotype is lethal. In other words, the association of rs123 with disease X is quite likely a red herring.

In this example, the laws of human genetics are available for checking sampling bias. Similar laws may not be available on other types of datasets. Fortunately, in the big data era, this can be accomplished in other ways, e.g. by retrieving published genotyping works on rs123, and comparing the reported rs123 genotype distributions with the respective observed distributions in the present dataset. If there is a big difference, we should be suspicious of the significant outcome observed in the present dataset.

2.2. Not ensuring that the null distribution is appropriate

While commonly used statistical tests all have their associated nominal null distributions, such null distributions are not necessarily appropriate for the analysis at hand. Venet et al. (2011) gave a demonstration of this situation. They used Cox's proportional hazards model to analyze whether some given multi-gene biomarkers – these are called signatures – correlate well with breast cancer survival. They noticed that signatures reported in the literature are no better than randomly generated signatures. In particular, large fractions of randomly generated signatures achieved statistical significance according to Cox's model. Yet, by definition of statistical significance at $p < 0.05$, no more than 5% of null samples – viz. the random signatures – can achieve statistical significance. Obviously, there is a problem.

Ordinarily, the null hypothesis for the Cox's model would be "There is no difference between the survival curves induced by the signature in question" (H0). Herein lies a problem: Null samples must be exchangeable under the null hypothesis, whereas random signatures are not exchangeable under H0. To permit random signatures as null samples, the null hypothesis needed would be "The difference between the survival curves induced by the signature in question is no different from that between the survival curves induced by random signatures" (H0'). It is not obviously wrong to use the usual test statistic associated with the Cox's model as the test statistic for H0'. However, as shown by Venet et al., the usual null distribution associated with the Cox's model is inappropriate for H0'. And those reported breast cancer survival signatures that are significant according to this null distribution are apparently no more meaningful than random ones.