Statistics and Probability Letters 136 (2018) 73–77

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 23 February 2018

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/

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.

Context	Occupation	Income>50K	Income<50K
Race = White	Adm-clerical	439 (14%)	2,645 (86%)
	Craft-repair	844 (23%)	2,850 (77%)

Fig. 1. Contingency table for the hypothesis "Among white Australians, those in the Craft-repair occupation have higher income than those in the Adm-clerical occupation".

2.3. Not ensuring that the null hypothesis is sensible

Over-representation analysis (Lee et al., 2005) is popular for analyzing gene expression profiling data. Typically, it first identifies – using t-test – a set of differentially expressed genes, and then compares this set of genes to biological pathways to identify – using hypergeometric test – those pathways that have large overlaps with this set of genes. The dysregulation of these identified pathways are thereby implicated as explanation for the phenotypes studied.

In spite of over-representation analysis' popularity, it exhibits poor reproducibility over independent datasets of the phenotypes studied. That is, when this procedure is applied independently on two datasets studying the same set of phenotypes, there is poor agreement between the two resulting lists of pathways. Lim et al. (2015) discussed the pertinent explanation below.

Recall that the hypergeometric test is used in this procedure to assess whether a pathway has a large overlap with the set of differentially expressed genes. The underlying null hypothesis is "The overlap of the set of differentially expressed genes with the pathway in question is no different from that with a random set of genes the same size as the pathway". This null hypothesis implies that the pathway in question is exchangeable with same-sized random sets of genes. However, this is biologically unreasonable because, by definition of a biological pathway, the genes therein must behave in a coordinated manner to deliver the function of that pathway, whereas a random set of genes does not need to be coordinated in any way. Thus this null hypothesis has a tendency to be rejected by real biological data regardless of the relevance of the pathway to the phenotypes studied.

3. Are there tactics I can use to get deeper insight from data?

Discovery of frequent patterns is central to revealing insight on a dataset. Many data mining systems have been developed for this (Han et al., 2012), and they produce thousands of frequent patterns from a dataset. Unfortunately, bewildered by a deluge of frequent patterns, a lay analyst cannot easily identify the interesting ones. Even after he has identified an interesting pattern, he does not know what to do next. These problems are compounded when analyzing big data. Two tactics for dealing with these obstacles are outlined below.

3.1. Think contingency tables

Consider the Australian adult dataset from the UCI machine learning repository at http://archive.ics.uci.edu/ml. It contains the demographic data of 32,561 adults. If a frequent pattern mining method is run on this dataset, thousands of patterns are produced. A lay analyst wades through all these patterns, and perhaps sees some of the following four patterns (the number at the end of each pattern is the occurrence count of that pattern in the adult dataset): {Race = White, Occupation = Adm-clerical, Income>50 K}: 439, {Race = White, Occupation = Adm-clerical, Income<50 K}: 2645, {Race = White, Occupation = Craft-repair, Income>50 K}: 844, and {Race = White, Occupation = Craft-repair, Income<50 K}: 2850. Maybe this inspires him to hypothesize that among white Australians, the Adm-clerical occupation is better paid than the Craft-repair occupation.

The contingency table in Fig. 1 can be constructed from these four patterns. A χ^2 test on this table confirms a significant difference in income between these two occupations among white Australians, and that the Craft-repair occupation is nearly 1.6 times more likely to have high income.

It is important to note that a typical data mining system produces patterns, but does not automatically put these patterns together in a way – e.g. a contingency table – that leads to a more interesting conclusion. The lay analyst has to discover this himself. But in practice, unless the lay analyst already has in mind some questions on income distribution in relation to occupation, he is rather unlikely find this significant difference in income between the two occupations from the output of the data mining system.

Thinking in terms of contingency tables can be an effective tactic to efficiently organize the thousands of patterns returned by typical data mining systems. Fig. 1 is an example of a contingency table. It is clearly more palatable to a lay analyst than the four constituting frequent patterns. Also, a traditional practice in statistical analysis is to analyze and compare one variable at a time, while controlling the other variables, against a target variable. Following this practice, given a target variable, it

Context	Occupation	Income>50K	Income<50K
Race = White,	Adm-clerical	16 (35%)	30 (65%)
Workclass = Self-emp-not-inc	Craft-repair	90 (18%)	409 (82%)

Fig. 2. Contingency table for the hypothesis "Among white Australians in the Self-emp-not-inc workclass, those in the Craft-repair occupation have lower income than those in the Adm-clerical occupation".

Context	Occupation	Income>50K	Income<50K
Race = White,	Adm-clerical	251 (24%)	787 (76%)
Sex = Male	Craft-repair	829 (24%)	2,695 (76%)

Context	Occupation	Income>50K	Income<50K
Race = White,	Adm-clerical	188 (9%)	1,858 (91%)
Sex = Female	Craft-repair	15 (9%)	155 (91%)

Fig. 3. Contingency tables for the hypothesis "Among male white Australians, those in the Craft-repair occupation have higher income than those in the Adm-clerical occupation" (top), and "Among female white Australians, those in the Craft-repair occupation have higher income than those in the Adm-clerical occupation" (bottom).

seems possible to organize and summarize the thousands of patterns produced by a data mining system into a few tens of contingency tables, one table for each of the other variables, rendering the output of the data mining system less bewildering.

3.2. Look for exception, trend reversal, and trend enhancement

Continuing with our white Australian example, when a lay analyst specifically looks to check whether "white Australians in the Craft-repair occupation are more likely, than those in the Adm-clerical occupation, to have income exceeding 50 K" (H1), he is likely to simply carry out the χ^2 test, report the relative risk, and not dig much deeper. Consequently, he does not know that the four patterns represented in the contingency table in Fig. 2 also exist, as these get buried among thousands of other frequent patterns.

A χ^2 test based on the information in these four patterns reveals the significant result that, among white Australians in the Self-emp-not-inc workclass, the Adm-clerical occupation is 1.9 times more likely than the Craft-repair occupation to have income exceeding 50 K. This result is important because it identifies a clear exception – in fact, a complete reversal – of the earlier conclusion, H1, that white Australians in the Craft-repair occupation earn more than those in the Adm-clerical occupation.

In fact, there are many other natural subpopulations of white Australians for whom H1 does not hold. In particular, the eight patterns summarized in the two contingency tables in Fig. 3 are also among the thousands of patterns returned by the system. A pair of χ^2 tests on these two contingency tables reveals that the hypothesis H1 is not significant among female white Australians, and is also not significant among male white Australians. That is, both females and males are exception to this hypothesis, though not a reversal: There is no difference in terms of income between the two occupations, when male and female white Australians are considered separately. Thus, this hypothesis is likely an artifact of a more fundamental phenomenon. In particular, the two tables reveal what the more fundamental phenomenon is: Male white Australians earn more than female white Australians (24% males versus 9% of females has high income), and male white Australians are mostly in the Craft-repair occupation while female white Australians are mostly in the Adm-clerical occupation.

Noticing exceptions or reversals to a pattern or trend can be an effective tactic to deriving important insight from data. A large population that poses an exception may suggest the pattern and trend is incorrect, needs refinement, or requires more thorough investigation. A lay analyst, who is just mechanically testing a given hypothesis, is likely to terminate his inquiry upon running a statistical test on his hypothesis. He is unlikely to check for exceptions, and likely ends up with a superficial conclusion that may not even be correct. A large subpopulation can also exhibit an increased frequency of a pattern or trend. This suggests this subpopulation is more susceptible to the trend. This is also worth a more thorough investigation, e.g. this

subpopulation may be an easier target for a marketing campaign. Thus, noticing increased frequency of a pattern or trend in subpopulations is another effective tactic to deriving important insight from data.

Recall in the previous section, I highlighted three common mistakes made by lay analysts when they do statistical hypothesis testing. Even when a hypothesis (e.g. H1 above) is significant, it may still be an artifact due to correlation of its variables (viz. Occupation) with some confounding factor (viz. Sex). An experienced statistician accounts for this via the process of blocking or stratification, such that only subsets of the samples that are similar in all regards, except the variable being measured, are compared. This makes the significance of the hypothesis independent of other explanation. The tactic of looking for exceptions or reversals to a pattern or trend can be regarded as a generalization of the blocking process, as the absence of such exceptions and reversals implies that the hypothesis has remained significant in all natural subpopulations.

4. Can I have a self-diagnosing, self-correcting, helpful and convenient analytics system?

As remarked earlier, a lay analyst can make many types of mistakes when doing hypothesis testing, and he can get bewildered by the thousands of frequent patterns returned by typical data mining systems. A hypothesis exploration system that supports the two tactics described earlier can improve this situation. I also hope the experienced statisticians and analysts among the readers of this article will suggest more tactics that can be incorporated into such a system. This will be a good step toward an analytic system that is self-diagnosing (i.e. it tries to detect whether the analyst is doing a valid statistical test), self-correcting (i.e. it tries to propose and make corrections to the analyst's statistical test), helpful (i.e. it searches for promising or interesting hypotheses related to some initial specified hypothesis or target variable), and convenient (i.e. it is easy to use). This is a vision articulated but not achieved by the Redhyte system (Toh et al., 2017).

Acknowledgment

I was supported in part by a Kwan-Im-Thong-Hood-Cho-Temple chair professorship during this work.

References

Han, J., Pei, J., Kamber, M., 2012. Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco, CA.
Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., Shahabi, C., 2014. Big data and its technical challenges. Commun. ACM 57 (7), 86–94.
Lee, H.K., Braynen, W., Keshav, K., Pavlidis, P., 2005. ErmineJ: Tool for functional analysis of gene expression data sets. BMC Bioinformatics 6, 269.
Lim, K., Li, Z., Choi, K.P., Wong, L., 2015. A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small. J. Bioinform. Comput. Biol. 13 (4), 1550018.
Toh, W.Z., Choi, K.P., Wong, L., 2017. Redhyte: A self-diagnosing, self-correcting, and helpful hypothesis analysis platform. J. Inf. Telecommun. 1 (3), 241–258.
Venet, D., Dumont, J.E., Detours, V., 2011. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput. Biol. 7 (10), e1002240.