Hypothesis Testing and Statistically-sound Pattern Mining*

Leonardo Pellegrina† Matteo Riondato‡ Fabio Vandin†

Abstract

The availability of massive datasets has highlighted the need of computationally efficient and statistically-sound methods to extracts patterns while providing rigorous guarantees on the quality of the results, in particular with respect to false discoveries. In this tutorial we survey recent methods that properly combine computational and statistical considerations to efficiently mine statistically reliable patterns from large datasets. We start by introducing the fundamental concepts in statistical hypothesis testing, including conditional and unconditional tests, which may not be familiar to everyone in the data mining community. We then explain how the computational and statistical challenges in pattern mining have been tackled in different ways. Finally, we describe the application of these methods in areas such as market basket analysis, subgraph mining, social networks analysis, and cancer genomics.

1 Tutorial Outline

We start with an introduction to the fundamental concepts behind statistical hypothesis testing, and the key questions that will be answered in the rest of the tutorial. In particular, we first introduce the framework of testing a single hypothesis (defining, e.g., what a null hypothesis is) and example applications where testing hypothesis is crucial, such as in biomedical research and in the study of social networks. We then discuss fundamental tests such as Fisher's exact test [6] and the related χ^2 and Barnard's test [1]. The final part of the introduction covers issues arising from testing multiple hypotheses on the same data and how to address these issues: we outline how and why the probability of discovering false positives grows in such scenarios, and how to control for this growth by bounding different metrics, such as the Family-Wise Error Rate (FWER) [4, 12] and the False Discovery Rate (FDR) [2, 3].

In the central part, we focus on mining statistically-sound patterns. We first define the problem and highlight its computational and statistical challenges arising from the combinatorial explosion of the number of hypotheses being tested and from the sheer size of data [10, 22, 28]. We then tackle these challenges one by one. We discuss how to make the process of finding statistically significant patterns efficient from a *computational* point of view [8, 16, 18, 23]. Specifically, we discuss efficient permutation testing [8, 16], the groundbreaking LAMP method [23] which allows to apply Tarone [22]'s method to combinatorial patterns, TopKWY [18], which efficiently extracts the k most statistically significant patterns while preserving guarantees on the FWER, and SPuManTE [17], which enables significant pattern mining with unconditional tests. The *statistical* efficiency is covered next: the works presented here [14, 20, 26, 27] introduce different methods to increase the statistical power of methods to extract significant patterns while controlling the FWER, and to deal with different inferential aspects of pattern mining. This part is the core of our tutorial: statistics and data mining come together to obtain fast and statistically-sound methods for pattern mining.

We then overview other interestingness measures for patterns which, although not based on hypothesis testing, are grounded in statistics and therefore relevant to this tutorial, such as emerging [15] and discriminative patterns [11], significant association rules [9]. All these patterns are interesting on their own, and their presentation allows us to perform a comparison of different approaches. A discussion of applications of the presented methods, from mining of significant subgraphs and motifs from large graphs [21], to biomedicine [24] and computational biology [7], will be provided.

In the final third part, we focus on more advanced material. Specifically, we show how to remove the assumptions on the data generating process [5], which have classically been used to make the problem more tractable. We also discuss how to weight hypotheses in a data-dependent way, with the goal of increasing the statistical power [13]. The materials covered here are recent developments that should interest the attending researchers, as will the potential future directions that complete the tutorial.

The outline of the tutorial is the following.

1. **Introduction and Theoretical Foundations**
   (a) Testing a single hypothesis: setting, basic concepts, and applications [25, Ch. 10]
   (b) Fundamental tests: Fisher's test [6], χ^2 test [25, Sect. 10.3], Barnard's test [1]
   (c) Testing multiple hypotheses: Family-Wise Error

*Tutorial website: `http://rionda.to/statdmtut/`

†Università di Padova

‡Amherst College

Rate [4] and False Discovery Rate [2]
(d) Bonferroni-Holm and Benjamini-Yekutieli corrections [3, 12]

2. **Mining Statistically-Sound Patterns**

(a) Computational and statistical challenges in pattern mining [10, 22, 28]
(b) Computational aspects: LAMP [23], permutation testing [8, 16], TopKWY [18], SPuManTE [17]
(c) Statistical aspects: hold-out approach and layered critical values [26, 27], a threshold for significant pattern mining [14], true frequent itemsets [20]
(d) Other measures: emerging patterns [15], discriminative patterns [11], significant association rules [9]
(e) Applications: subgraph mining [21], cancer genomics [24], computational biology [7], and survival analysis [19]

3. **Recent developments and advanced topics**

(a) Removing assumptions [5]
(b) Data-dependent hypothesis weighting [13]
(c) Conclusions, future directions, and discussion

References

[1] G. A. Barnard. A new test for 2×2 tables. *Nature*, 156:177, 1945.

[2] Y. Benjamini, Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. of the R. Stat. Soc. B*, pages 289–300, 1995.

[3] Y. Benjamini, D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, 29(4):1165–1188, 2001.

[4] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pub. R. Ist. Sup. Sci. Econ. Commerc. Firenze*, 8: 3–62, 1936.

[5] L. Choi et al. Elucidating the foundations of statistical inference with 2×2 tables. *PloS ONE*, 10(4): e0121263, 2015.

[6] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.*, 85(1):87–94, 1922.

[7] Y. Fukasawa et al. Mitofates: improved prediction of mitochondrial targeting sequences and their cleavage sites. *Molec. & Cell. Proteom.*, 14(4): 1113–1126, 2015.

[8] A. Gionis et al. Assessing data mining results via swap randomization. *ACM Trans. Knowl. Disc. Data*, 1(3):14, 2007.

[9] W. Hämäläinen, M. Nykänen. Efficient discovery of statistically significant association rules. In *ICDM'08*, 2008.

[10] W. Hämäläinen and G. I. Webb. A tutorial on statistically sound pattern discovery. *Data Min. Knowl. Disc.*, 33(2):325–377, 2019.

[11] Z. He et al. Significance-based discriminative sequential pattern mining. *Exp. Sys. Appl.*, 2018.

[12] S. Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, pages 65–70, 1979.

[13] N. Ignatiadis et al. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature meth.*, 13(7):577, 2016.

[14] A. Kirsch et al. An efficient rigorous approach for identifying statistically significant frequent itemsets. *J. ACM*, 59(3):12, 2012.

[15] J. Komiyama et al.S. Minato. Statistical emerging pattern mining with multiple testing correction. In *KDD'17*, 2017.

[16] F. Llinares-López et al. Fast and memory-efficient significant pattern mining via permutation testing. In *KDD'15*, 2015.

[17] L. Pellegrina et al. SPuManTE: Significant Pattern Mining with Unconditional Testing. In *KDD'19*, 2019.

[18] L. Pellegrina, F. Vandin. Efficient mining of the most significant patterns with permutation testing. In *KDD'18*, 2018.

[19] R. T. Relator et al. Identifying statistically significant combinatorial markers for survival analysis. *BMC Med. Genom.*, 11(2):31, 2018.

[20] M. Riondato, F. Vandin. Finding the true frequent itemsets. In *SDM'14*, 2014.

[21] M. Sugiyama et al. Significant subgraph mining with multiple testing correction. In *SDM'15*, 2015.

[22] R. E. Tarone. A modified Bonferroni method for discrete data. *Biometrics*, pages 515–522, 1990.

[23] A. Terada et al. Statistical significance of combinatorial regulations. *Proc. Nat. Acad. Sci.*, 110(32): 12996–13001, 2013.

[24] F. Vandin et al. Algorithms for detecting significantly mutated pathways in cancer. *J. Comp. Biol.*, 18(3):507–522, 2011.

[25] L. Wasserman. *All of Statistics: A concise course in statistical inference*. Springer, 2013.

[26] G. I. Webb. Discovering significant patterns. *Mach. Learn.*, 68(1):1–33, 2007.

[27] G. I. Webb. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. *Mach. Learn.*, 71(2-3):307–323, 2008.

[28] P. H. Westfall, S. S. Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley-Interscience, 1993.