הכנס השנתי של האיגוד הישראלי לסטטיסטיקה

אוניברסיטת בן-גוריון בנגב, 25.5.2017

תקצירי פוסטרים

Discovering replicated findings across several studies of high dimension

Gili Baumer, Marina Bogomolov, The Technion

The aim of replicability analysis is to identify the findings that replicate across independent studies that examine the same features, and quantify the strength of replication. These features can be single-nucleotide polymorphisms (SNPs) examined for associations with disease, genes examined for differential expression, etc. The importance of replicability analysis is well recognized in many fields, where the intention is to show that the result is consistent over different studies and is not unique to a specific study or setting. We introduce a powerful replicability analysis procedure that extends the partial conjunction approach by incorporating an assumed lower bound for the fraction of hypotheses that are null in all the studies. We show the performance of our method in simulations as well as on real data from independent studies of different psychiatric disorders.

Scalable Non-Parametric Tests of Independence

Barak Brill, Ruth Heller, Tel Aviv University

In modern applications it can be of interest to identify complex relationships between pairs of random variables. Powerful distribution-free tests for this purpose have been suggested in Heller et al. (2016). However, the computational complexity of their tests are polynomial in sample size, thus limiting their potential use to studies with small enough sample size.

In this work we present a modification of the tests in Heller et al. (2016), which can be applied for studies of any sample size. The resulting tests retain the properties of omnibus consistency and distribution-freeness of the original tests in Heller et al. (2016). We show in simulations that the proposed tests have excellent power in comparison with classic and state-of-the-art alternative tests.

We also provide a real data example, were the tests provide identification of pairs of psychological disorders with shared genetics. In this example, the sample size is $N = 10^4$.

All methods presented are available from CRAN in the R package 'HHG'.

מידול היעדרות של עובדים במשרד הבריאות בישראל

רבקה בריד, מכון גרטנר

רקע ומטרות. משרד הבריאות מעוניין לקדם את בריאות העובדים במשק הישראלי, לשפר את איכות חייהם ולהגביר את פריון עבודתם. תנאי חשוב להצלחה הוא הבנת אורחות החיים ותנאי עבודה הקשורים להערכה נמוכה של העובד לגבי מצב בריאותו, להיעדרות מהעבודה ולנוכחותיות (נוכחות בעבודה, אך ברמת תפקוד תת-מיטבית).

איסוף הנתונים. נאספו נתוני 625 עובדים בסקר בריאות העובד שנערך בקרב עובדי מטה משרד הבריאות. העובדים נשאלו על הרגלי הבריאות שלהם (רמת הפעילות הגופנית, סוג התזונה, הרגלי עישון, הרגלי שינה וכו'), מצב בריאותם ותנאי עבודתם. כן נשאלו על התוצאים הבאים: (א)מס' שעות ההיעדרות מהעבודה במהלך השבוע הקודם לסקר (ב)דירוג הנוכחותיות בעבודה במהלך השבוע הקודם לסקר (ג) דירוג אישי של מצב הבריאות. ישנן 103 תצפיות חסרות במידע על היעדרות ונוכחותיות במקום העבודה ו-49 תצפיות חסרות במידע על מצב בריאות.

שיטות העבודה עבור תוצא ההיעדרות

- משתנים מסבירים הוגדרו כקטגוריאליים. כשמעל אחוז מהנתונים במשתנים אלה היו חסרים, הם הושמו בקטגוריה נפרדת.
- תוצא ההיעדרות חולק ל-3 קטגוריות סדורות.
- לצורך התחשבות בערכים החסרים בתוצא, ביצענו שקלול הסתברות משלימה (inverse probability weighting), לפי ההסתברות שידווחו.
- לחישוב משקל לכל תצפית נעזרנו במודל לוגיסטי (התוצא: היעדרות מדווח/לא מדווח).
- נבחר מודל רגרסיה לוגיסטית סודרת (ordinal) משוקללת לכל תוצא בנפרד כדי למצוא קשרים בין הגורמים המסבירים לבין התוצא. הכנסת המשתנים המסבירים נעשתה בשיטת stepwise וכללה 4 שלבים: א. משתנים דמוגרפים ב. משתנים הידועים מהספרות כמשפיעים על ממדי התוצאים ג. התוצאים האחרים: נוכחותיות ומצב בריאות ד. שאר המשתנים המסבירים. בכל שלב נשמרו המשתנים שנבחרו בשלב קודם.

תוצאות ראשוניות: משתנים הקשורים לתוצא היעדרות. גיל, מין (כולל קטגוריה נפרדת לנשים בהריון), השכלה, BMI בעיות תפקוד עקב חוסר שינה, שחיקה בעבודה, לחץ בעבודה ,נוכחותיות.

בשיתוף עם: מירב מרום[1], אלי רוזנברג[1], איליה נוביקוב[2], לורנס פרידמן[2].

1: משרד הבריאות, ישראל. 2: מכון גרטנר לחקר אפידמיולוגיה ומדיניות בריאות

Effective data visualization in 5 simple rules demonstrated with heatmaps in R

Tal Galili, Yoav Benjamini, Tel Aviv University

A cluster heatmap is a popular graphical method for visualizing high dimensional data, in which a table of numbers are encoded as a grid of colored cells (Wilkinson and Friendly 2009, Weinstein (2008)). The rows and columns of the matrix are ordered to highlight patterns and are often accompanied by dendrograms and extra columns of categorical annotation. Heatmaps are used in many fields for visualizing observations, correlations, and missing values patterns. There are many R packages and functions for creating static heatmap figures (the most famous one is probably `gplots::heatmap.2`).

The **heatmaply** R package allows the creation of interactive cluster heatmaps, enabling tooltip hover text and zoom-in capabilities (from either the grid or the dendrograms), while supporting sidebar annotation. The package brings together many well known packages such as **ggplot2** (Wickham 2016), **plotly**, **viridis**, **seriation** (Hahsler, Hornik, and Buchta 2008), **dendextend** (Galili 2015), and others. Also, it is now supported by the **shinyHeatmaply shiny** app.

You can play with a simple interactive example by running:

```
install.packages('heatmaply'); library('heatmaply')
```

```
heatmaply(percentize(mtcars), k_row = 4, k_col = 2, margins = c(40,120,40,20))
```

This poster will provide an overview of design principles for creating a useful, and beautiful, cluster heatmap. Attention will be given to data preprocessing, choosing a color palette, and careful dendrograms creation.

This work was made possible thanks to the essential contribution of Jonathan Sidi, Alan O'Callaghan, Carson Sievert, and Yoav Benjamini. As well as the joint work of Joe Cheng and myself on the **d3heatmap** package (which laid the foundation for **heatmaply**). The speaker is the creator of the R packages **installr**, **dendextend**, and **heatmaply**, and blogs at: www.r-statistics.com.

References

Galili, Tal. 2015. "Dendextend: An R Package for Visualizing, Adjusting and Comparing Trees of Hierarchical Clustering." *Bioinformatics*. Oxford Univ Press, btv428.

Hahsler, Michael, Kurt Hornik, and Christian Buchta. 2008. "Getting Things in Order: An Introduction to the R Package Seriation." *Journal of Statistical Software* 25 (3). American Statistical Association: 1–34.

Weinstein, John N. 2008. "A Postgenomic Visual Icon." *Science* 319 (5871). American Association for the Advancement of Science: 1772–3.

Wickham, Hadley. 2016. Ggplot2: Elegant Graphics for Data Analysis. Springer.

Wilkinson, Leland, and Michael Friendly. 2009. "The History of the Cluster Heat Map." *The American Statistician* 63 (2). Taylor & Francis: 179–84.

Optimization of Monte-Carlo simulation for Surface Danger Zones

Michael Gringauz, David Steinberg, Tel Aviv University

Field exercises that include small arms pose a potential risk to personnel, civilians and property. Our concern is how to establish a surface danger zone (SDZ) for such exercises. It should be safe outside that zone. Here "safe" usually means the probability of a hit outside the SDZ is at most 10-6. We want to provide a statistical confidence level for our SDZ. Many factors can affect the calculation of an SDZ. Some reflect specific conditions, others physics, others are random, i.e. weapon and ammunition characteristics, errors in aiming, distance to target, local topography, projectile and flight dynamics, ricochet dynamics, meteorological conditions.

The traditional method was to compute SDZ's deterministically for the worst case scenario, as a pie wedge from the point of fire with a fixed azimuth. This approach creates unnecessarily large and overestimated danger zones, placing heavy constraints on live fire training. In the last 20 years probabilistic approaches have been developed by NATO and the US Army. Computer code mimics the physical system to reflect the ballistic flight and ricochet dynamics. Some elements are derived from basic physics; others result from lab experiments. Multi-stage Monte Carlo simulation is used to generate a hit probability map, which in turn can be translated to a danger zone with the required probability and confidence coefficient.

The major goal of our current research is to drastically reduce the simulation burden. The primary tool will be to model the simulator so as to better focus the simulation budget using importance sampling. Conformal prediction is used to compute the SDZ.

The simulation includes numerous hidden random variables. Each final hit point may occur with different realizations of these intermediate variables. The idea is to find a better way of sampling from these intermediate variables in order to reduce the variance for estimating the shoulders of the final distribution, while reducing the total number of Monte Carlo runs.

The consequent calculation of the SDZ with required confidence/credibility level is performed using the method developed by Lei [1], with appropriate adjustments so that it can be used together with importance sampling and with the binned data in our application (as opposed to the point data in Lei's work).

Preliminary results show that significant reduction in the number of Monte Carlo runs is achieved using our method – about 20%, without significantly affecting the shape and the size of the resulting SDZ. More work is to be performed, mainly choosing the optimal importance sampling functions and optimizing the kernel functions of the credibility region calculation. The influence of reducing the total number of runs on the variance of the SDZ bounding contour will also be investigated.

1. Jing Lei, James Robins & Larry Wasserman, "Distribution-Free Prediction Sets", Journal of the American Statistical Association, pp. 278-287, 2012.

The Limit Distribution of Linear Combinations of Partial Maxima, with Application to Asymptotic Efficiency of Ranking and Selection Procedures

Royi Jacobovic, Or Zuk, The Hebrew University of Jerusalem

We derive the asymptotic properties of weighted sums of partial maxima of a sequence of i.i.d. random variables, using methods and insights from extreme-value theory. Such sums arise naturally in the study of ranking and selection procedures. We use our approach to analyze one-stage and two-stage selection procedures for the case of selection from Gaussian populations under Bechhofer's indifference-zone formulation: (1) We generalize previous results of Siegmund and Robbins by finding first order approximation of the sample size required to maintain exogenous level of probability for correct selection of k(n) items out of a list of n when $n \to \infty$. (2) We derive the asymptotic relative efficiency guaranteed by Dalal & Dudewicz and Rinnot's two-stage selection procedures, showing a constant multiplicative gap in sample size in favour of Dalal & Dudewicz.

On the Implications of Prior Distribution and Asymptotic Independence Assumptions over the True and Sample Correlations under Bayesian Modeling.

Royi Jacobovic, The Hebrew University of Jerusalem

The prediction of cancer prognosis and metastatic potential immediately after the initial diagnoses is a major challenge in current clinical research. The relevance of such a signature is clear, as it will free many patients from the agony and toxic side-effects associated with the adjuvant chemotherapy automatically and sometimes carelessly subscribed to them. Motivated by this issue, Ein-Dor, Zuk and Domany [Proceedings of the National Academy of Sciences, 15(103):5923-5928 (2006)] presented a statistical model which states that thousands of samples are needed to generate a robust gene-list for predicting outcome. Their model assumptions are done over the joint distribution of the sample correlations of the genes with the survival status. This work is devoted to the implications of these assumptions on the joint distribution of the data. In particular it is proved that sparsity or Gaussianity assumptions over the joint distribution of the data are inconsistent with the assumptions of Ein-Dor et al. To see the practical contribution of these results observe that as mentioned by Ein-Dor et al., the performance of their methodology depends on the existence of their preliminary assumptions. Finally, with regard to the sensitivity of this methodology to the model assumptions, the issue of testing these assumptions by empirical data is revisited by conduction of simulation analysis.

Co-evolution of genes and phenotypes by gene embedding

Iyar Lin, The Hebrew University of Jerusalem

We propose a new method for the analysis of co-evolution of genes and phenotypes. Our method embeds genes into Euclidean space, which allows us to use a generative model for the continuous embedded coordinates, based on a Brownian motion along the phylogenetic gene tree. We formulate co-evolution of genes and phenotypes as a hypothesis testing problem for the covariance matrix of the Brownian motion, and derive a finite-sample Bartlett-type correction for the log-likelihood ratio test statistics, which allows us to compute reliable p-values analytically.

While most current methods compare only genes' presence/absence in different species which may be useful for studying distant species, our method utilizes quantitative sequence distance information and is best suited for the study of closely related species where the set of encoding genes barely changes, and phenotypic variance is mostly due to changes in gene sequence. We perform extensive simulations showing that (i) our method is robust to deviations from generative model's assumptions, keeping the type-1-error under control, and (ii) our method outperforms the mirrorTree and tolMirror algorithms in detection accuracy of co-evolving gene-gene and gene-phenotype pairs.

Assessing the Accuracy of a Single Point on an ROC Curve: Sensitivity at Fixed Specificity

Yael Magen Embon, Tel Aviv University

We investigate methods for testing whether a diagnostic test, with results measured on a continuous scale, achieves a certain level of sensitivity for a given specificity. This problem was approached by different methods in the context of statistical hypothesis testing while taking into account the need to estimate the critical value for discriminating between diseased and healthy subjects. This work deals mainly with the performance of five tests addressing this issue and consists of an investigation and comparison of these via extensive simulations.

This work was done under the supervision of Professor David Steinberg

The Busy-Nurse Effect in Clinical Trials

Liran Mendel, David Steinberg, Tel Aviv University

Introduction. We consider an experimental design for estimating the parameters of a probit regression model when the actual experimental conditions that are used at the experimentation stage may randomly and independently fluctuate around those that are specified at the design stage. However, the values used in the experimentation are observed, and they can be used for estimating the parameters of the model. A direct extension of D-optimal design is to consider the expected value of the determinant of the information matrix with respect to the possible fluctuations of the experimental variables.

We assume that the experiment is designed off-line (nonsequentially), but when the experimentation takes place, the actual experiment that is used may differ from the one that was designed. For instance, when the time at which observations are made is one of the design variables, and human operation is required, the experimenter may be available only at a random time around that specified by the design.

Pharmacokinetic experiments, with observations corresponding to blood samples taken from individuals, provide examples of such a situation.

Goal. Evaluate the effect of limited control over the design factor, for the scenario of single factor sensitivity experiments with a binary outcome, and with good or poor prior information. Would that condition require modified design for optimality?

Evaluation (by Simulation). For each set of guessed parameters, a grid of 2-point symmetric designs was tested using no fluctuation, as well as mild, moderate and significant fluctuations. Each setting was tested with 500 simulated sets of 80 points, and the D-criteria (determinant of the information matrix) was calculated to find D-optimality: 1. When simulating stimulus only (simulate X) 2. When simulating both stimulus (X) and response (Y) based on X (simulate XY).

Results

1. With <u>good</u> prior information about the parameters: Given reasonable level of fluctuations, keep the D-optimal design.
2. With <u>poor</u> prior information about the parameters: Given uncertainty regarding your curve parameter, and reasonable level of fluctuations, it is better to underestimate the scale than overestimate it. A reasonable deviation in the guess of the location is of less importance.

מודל סטטיסטי להשוואת טביעות נעליים

אורית מורדוב, האוניברסיטה העברית בירושלים

מומחים פרונזים העוסקים בטביעות נעליים משווים טביעת נעל שנמצאה בזירת פשע לטביעת נעליו של חשוד וקובעים את מידת התאמתם ללא שימוש בכלים סטטיסטים. שיטות אלו זכו לאחרונה לביקורת קשה עד כדי אפשרות ממשית לפסילת השימוש בראייה של טביעת נעל בבתי המשפט. בחינת הטביעות והשוואתן מתבססת על פגמים בסוליית הנעל המופיעים גם בטביעות נעליים, אך לא בצורה מושלמת. מטרת המחקר הנוכחי היא לפתח שיטה סטטיסטית להשוואת טביעות נעליים. השיטה אותה אנו מציעים מתבססת על חישוב המרחק בין אוסף פגמים של שתי הטביעות, זו מזירת הפשע וזו שנלקחה מנעלו של החשוד, ובדיקת הסבירות למרחק שהתקבל תחת ההנחה שהטביעות לא נוצרו מאותה הנעל. הרעיון המרכזי הוא להגדיר מרחק בין אוסף פגמים בשתי נעליים המסתמך על כמות הפגמים, על מיקומם, ועל טעויות שונות באיסוף הנתונים. בעזרת אוסף נעליים הנמצא בידי המעבדה לזיהוי פלילי ניתן ללמוד על התפלגות המרחק ולחשב הסתברויות שונות.

בשיתוף עם מיכה מנדל, יורם יקותיאלי, נעמי קפלן, שרינה וייזנר וירון שור.

Bcd-Dependent Enhancer Sequences Reveal Robust Regulatory Rules for Anterior-Posterior Body Patterning in the Early *Drosophila* Embryo

Shazman Shula, The Open University of Israel

The “morphogen” hypothesis states that gradients of morphogenic molecules provide positional information that establishes different cell fates along the axes of the developing embryo. In the first steps of embryogenesis in D. melanogaster, the maternal transcription factor Bicoid (Bcd) is expressed as a gradient that is densest in the anterior pole of the cell. Binding of Bcd to specific sites in enhancers throughout the genome is critical for establishing the body plan and starting the remainder of embryonic development.

More than 60 Bcd-dependent enhancers have been identified, but the combinatorial regulatory rules that determine that vast scope of gene expression in these elements remain a mystery. We used machine learning and statistical analysis to explore the hypothesis that Bcd and six other transcription factors determine gene expression boundary through a combinatorial mechanism. We found that this system is robust enough that regulatory rules can be revealed by making predictions based on binding site numbers and **PBP - Posterior Boundary Position** alone. Furthermore, we computationally demonstrated Run's strong contribution to PBP determination, as well as discovered facets of Bcd and Hunchback\Krüppel binding that were previously unknown. In the future, this model can be used to predict the PBPs of suspect Bcd-dependent enhancers that have not yet been identified in vivo. Broadly, the process of this investigation also provides a framework for predicting gene expression in other systems outside of development and D. melanogaster.

Testing Independence with Biased sampling

Yaniv Tenzer, Or Zuk, Micha Mandel, The Hebrew University of Jerusalem

We study the problem of testing (quasi)-independence of two random variables under general known biased sampling, generalizing previous studies on independence under different truncation operations into a unified framework. We develop a weighted permutation testing framework for general test statistics and demonstrate its application using a modified Hoeffding's test statistic. We define an appropriate distribution over permutations, determined by the biased sampling operator, and develop an MCMC algorithm for sampling from this distribution, thus guaranteeing that we control the type-1-error at a desired level. We compare the permutation-based framework to a bootstrap approach via simulations, highlighting the strengths and weaknesses of the two methods.

Weighted averages of natural images

Jonathan Zouari, Yuval Benjamini, The Hebrew University of Jerusalem

Natural images - photos of natural scenes - are one of the most commonly found objects on the web. In our work, we present a new way to summarise a cohort of natural images using a weighted Euclidean mean that approximated the Frechet mean. Our algorithm consists of three major steps: (1) Alignment using a graph alignment algorithm; (2) measuring image distance by moving image representation to an invariant feature space (HOG, AlexNet) (3) taking a weighted average of the images based on the distances with an additional curvature correction. The algorithm can also recover a missing image from a proximity vector to images in the cohort. We apply our method to small cohorts of images taken from ImageNet Large-Scale Visual Recognition Challenge, and show the algorithm creates a high-quality recoveries the missing images. We discuss the choice of success metric.