הכנס השנתי של האיגוד הישראלי לסטטיסטיקה

אוניברסיטת בן-גוריון בנגב, 25.5.2017

תקצירי הרצאות

הרצאת מליאה

Design and Initial Analysis of Largish Omics Studies

Prof. Terry Speed, Bioinformatics division at WEHI, Melbourne, Australia

Many groups are carrying out large-scale genomic, transcriptomic, proteomic or related studies. An example is NIH's LINCS program. With such studies, there is an inevitable need for some kind of normalization. Biologists have long made use of controls of different kinds to *monitor* unwanted variation, such as negative control probes, spike-ins and replicate reference samples. However, we can also use these controls to *normalize*. I shall explain how, illustrating with gene expression data on microarrays, RNA-seq and Nanostring. As soon as we have effective tools for the analysis of data, including controls, from a given kind of study, we should turn our attention to the design of these studies. I shall say a little about that too.

This talk is based on joint work with Johann Gagnon---Bartsch of the University of Michigan, and Laurent Jacob of the CNRS, Lyon.

מושב I – סטטיסטיקה במערכות דינמיות

יו"ר: איתי דטנר, אוניברסיטת חיפה

What can be observed in real time PCR and when does it show?

Pavel Chigansky, Department of Statistics, The Hebrew University

Real time, or quantitative, PCR typically starts from a very low concentration of initial DNA strands. During iterations the numbers increase, first essentially by doubling, later predominantly in a linear way. Observation of the number of DNA molecules in the experiment becomes possible only when it is substantially larger than initial numbers, and then possibly affected by the randomness in individual replication. Can the initial copy number still be determined? We will approach this question using some new result from approximation of the density dependent branching processes.

Joint work with P. Jagers and F. Klebaner

A two-stage approach for estimating the parameters of an age-group epidemic model from incidence data

Rami Yaari, Department of Statistics, University of Haifa and Bio-statistical Unit, The Gertner Institute for Epidemiology and Health Policy Research

Age-dependent dynamics is an important characteristic of many infectious diseases. Age-group epidemic models describe the infection dynamics in different age-groups by allowing to set distinct parameter values for each. However, such models are highly nonlinear and may have a large number of unknown parameters. Thus, parameter estimation of age-group models, while becoming a fundamental issue for

both the scientific study and policy making in infectious diseases, is not a trivial task in practice. In this talk, we examine the estimation of the so called next-generation matrix using incidence data of a single entire outbreak, and extend the approach to deal with recurring outbreaks. Unlike previous studies, we do not assume any constraints regarding the structure of the matrix. A novel two-stage approach is developed, which allows for efficient parameter estimation from both statistical and computational perspectives. Simulation studies corroborate the ability to estimate accurately the parameters of the model for several realistic scenarios. The model and estimation method are applied to real data of influenza-like-illness in Israel. The parameter estimates of the key relevant epidemiological parameters and the recovered structure of the estimated next-generation matrix are in line with results obtained in previous studies.

Feature Selection for Accurate Time Prediction in Congested Healthcare Systems

Arik Senderovich, Faculty of Industrial Engineering and Management, Technion

Time prediction in healthcare systems such as outpatient clinics, hospital wards, and emergency departments is essential for decision making. Predictions are used for effective and efficient resource allocation, optimized ambulance routing, and accurate delay announcements. In this talk, we focus on time prediction in congested healthcare systems, where patients share scarce resources such as nurses, physicians, and MRI machines. To predict times, we employ the framework of supervised machine learning, using event logs, which are data recordings readily available in today's healthcare information systems.

The accuracy of supervised learning methods depends on the predictive power of the selected features. In congested systems, patient’s length-of-stay not only depends on clinical information (e.g., patient’s age, gender, and medical history), but also on cross-patient information (e.g., the number of patients in the hospital, and the length-of-stay of recently discharged patients). To properly select features that capture these cross-patient dependencies, we propose the model of congestion graphs, which are grounded in queueing theory and are mined from event logs. We show an evaluation of the approach based on two real-world healthcare systems, namely an Israeli emergency department and an outpatient cancer hospital in the United States.

Respondent Driven Sampling as a Counting Process

Yakir Berchenko, Department of Industrial Engineering and Management, Ben-Gurion University

Respondent driven sampling (RDS) is an approach to sample human populations that utilizes their social networks. Although RDS has become a widespread method for recruiting individuals within hidden populations, estimation using RDS is problematic due to biased sampling (e.g., over-sampling participants with many acquaintances). Most RDS studies attempt to adjust for this bias using inverse-degree weighting, assuming that the probability an individual is sampled is proportional to the number of their acquaintances (degree). However, this seemingly necessary assumption is unlikely to hold in practice. We propose a novel approach that relaxes this assumption, by using a source of information that is usually ignored, the precise timing of recruitment. Our new approach, adapting methods developed for inference in epidemic processes, allows us also to test the assumption of recruitment proportional to degree, as well as to generate estimates of the total population size.

We analyse these estimators and find them asymptotically consistent and normally distributed; applying them to empirical RDS data-sets studied, we show that the probability an individual was sampled was not proportional to their degree, and in two datasets where the size and degree distribution can be characterised, we show that our new maximum likelihood estimator outperforms the standard inverse-degree weighting estimator.

אסטרטגיה לתזמון שיחות הטלפון במערכת ה- CATI

נועם כהן, הלשכה המרכזית לסטטיסטיקה

בחלק מסקרי הלמ"ס נחקרים הנדגמים באמצעות שיחות טלפון כשההתקשרות נתמכת באמצעות מערכת מחשב CATI-computer assisted telephone interviewing. בכל חודש מבוצעות עשרות אלפי שיחות אל משקי בית שעלו במדגם כשחלקן נעשות בזמן לא אופטימלי מבחינת משק הבית ולכן אינן נענות או שנענות בסירוב. המטרה היא להגדיר אסטרטגיה שתקבע את זמן ההתקשרות המיטבי למשק בית מסוים על פי מאפייניו. הזמן המיטבי הוא כזה שממכסם את הסיכוי לחקירה של נדגם בהינתן מאפייניו. כתוצאה מכך קטן מספר ההתקשרויות הכולל תוך שמירה על אותו מספר משיבים. כמו כן על האסטרטגיה להיות מאוזנת בצורה כזו שתשמור על אותה תוחלת מספר התקשרויות לכל משק בית שנדגם. היינו בכל שעה משעות היום יהיה קיים תור בו לכל משק בית יש קדימות אחרת הנקבעת גם על פי הסיכוי שלו להשיב באותה שעה אך גם על פי תוחלת מספר השיחות הכולל שיעשה אליו במשך היום.

יעילות הענישה כגורם הרתעתי בישראל

ד"ר שחר יונאי*, ד"ר יורי גובמן**, מיכל מילר-כהן*, *משטרת ישראל, **נס טכנולוגיות

השפעה הרתעתית של ענישת העבריינים מהווה נושא חשוב במחקר קרימינולוגי וסטטיסטי. פרט לחשיבותם המדעית, ממצאי המחקרים בנושא משמשים את מקבלי ההחלטות בבואם לקבוע מדיניות הענישה אשר יש בכוחה להרתיע את העבריינים מביצוע עבירות חוזרות (רצידיוויזם).

ניתן להצביע על מספר מטרות של ענישה - הרחקה מן הציבור (שלילת יכולת), שיקום והרתעה. בספרות מציינת שני היבטים של הרתעה: הרתעת הרבים, בדגש על האוכלוסייה העבריינית, והרתעת היחיד, המכוונת לעבריין הבודד.

החוקרים חלוקים הן בנוגע להשפעה הרתעתית של ענישה והן בנוגע לשיטות מחקר היכולות לענות על שאלה זו. מרבית המחקרים נשענים על ניתוח איכותני, ניתוחי מקרה מיוחד או זעזוע חברתי (case study) או ניתוח סקרים שרובם מבוססים על מספר מועט של תצפיות וסובלים מהטיה עקב שיעורים גבוהים של אי-השבה מפאת רגישות הנושא.

במחקר הנוכחי פותחה מתודולוגיה סדורה לניתוח התופעה, אמידת ההשפעה הרתעית של הענישה וניבוי ההשפעה הזאת באמצעות כלים סטטיסטיים מתקדמים. המחקר התבסס על נתונים ייחודיים הנמצאים ברשות משטרת ישראל ומכסה את אוכלוסיית העבריינים, לרבות מאפייני פרט ותיאור מפורט של הקריירה הפלילית שלהם.

בוצע ניתוח הכולל מעקב אחרי "קבוצות טיפול" שונות לאחר תום ריצוי העונש. בבניית קבוצות הטיפול והביקורת נעשה שימוש בהגדרות שונות של עבריינות חוזרת הנדונות בספרות הקרימינולוגית.

ניתוח מנת יחס הסיכוים איפשר לאמוד את האפקט ההרתעי של סוגים שונים של ענישה, ולזהות גורמים המשפיעים באופן משמעותי ומובהק סטטיסטית על התופעה הנחקרת. העשרת בסיס הנתונים במספר רב של מאפייני הפרט איפשר פיקוח על מגוון רחב של תכונות דמוגרפיות וחברתיות של העבריין. ניתוח זמני עד האירוע באמצעות רגרסיה של Cox איפשר ניבוי של תווך הזמן אחריו העבריין הספציפי כלול לחזור לפעילות פלילית.

נמצא, כי הטלת עונש מאסר בפועל מגדילה את סיכויי העבריין לחזור ולבצע עבירות פליליות. יתירה מזאת, סיכוי זה גדל עם מספר המאסרים: מאסר בפועל נוסף מגדיל סיכוי לרצידיוויזם במאות אחוזים, ביחס לקבוצת שכללה עבריינים שתיקם הפלילי נסגר ללא הגשת כתב אישום.

המחקר ממחיש כי המשתנים המנבאים העיקריים של רצידיביזם הם משתנים המתארים את ההיסטוריה הפלילית והמודיעינית של העבריין. ניתוח קבוצות טיפול שונות מצביע על כך שהטלת קנס מרתיעה עבריינים, ובמיוחד קנסות בגובה משמעותי.

מודל צמיחה היררכי לחקירת דפוס ההיבחנויות החוזרות במכפ"ל

דביר קלפר, מרכז ארצי לבחינות ולהערכה

מטרת המחקר. הקבלה ללימודים של מועמדים לאוניברסיטאות בישראל מתבססת בעיקר על הישגיהם בתעודת הבגרות ובבחינת הכניסה הפסיכומטרית לאוניברסיטאות (מכפ"ל). לאור זאת, רבים מהנבחנים במכפ"ל נבחנים יותר מפעם אחת וזאת על-מנת לשפר את סיכויי קבלתם לחוג אותו הם רוצים ללמוד.

בנוסף, מחקרים מצאו שקיים קשר בין נתוני רקע דמוגרפיים (מין, גיל, השכלת הורים ועוד) להישגים לימודיים וציוני בחינות (למשל Liu et. al. 2012), כך גם באשר לבחינה הפסיכומטרית (סער ואורן 2014; קלפר ואחרים 2015).

בהתאם לכך, המטרה המרכזית של המחקר הנוכחי היא לבדוק ולאפיין את דפוס ההיבחנויות החוזרות אצל אנשים שנבחנים מספר פעמים במכפ"ל בהקשר של נתוני רקע דמוגרפי, ובפרט לענות על שתי השאלות הבאות:

- מהו השינוי הממוצע בציון הבחינה הפסיכומטרית עבור אנשים שנבחנים מספר פעמים, ומהם הגורמים שמשפיעים עליו?
- האם ניתן לנבא את הסיכוי שנבחן יבחן יותר מפעם אחת בבחינה הפסיכומטרית בעזרת משתני רקע? ואם כן, באיזה כיוון הם משפיעים?

אוכלוסייה ושיטת ניתוח הנתונים. בסיס הנתונים מכיל את כל הנבחנים בבחינה הפסיכומטרית בשנים 2001-2010 ואשר נבחנו בשפה העברית, כאשר הוחלט להוציא מהניתוח נבחנים שנבחנו יותר מחמש פעמים (ששיעורם נמוך מ-1% מהנבחנים), וזאת משום שעבורם התקבל דפוס שונה.

בסה"כ התקבלו 436,784 רשומות (היבחנויות) בבסיס הנתונים כאשר מתוכם ישנם 336,193 אנשים שונים.

בסיס הנתונים הכיל מספר רב של נתונים חסרים לגבי משתני הרקע הדמוגרפיים: השכלת אב, השכלת אם ומצב סוציואקונומי. נתונים אלו הושלמו ראשית בשיטה בסיסית (היבחנויות אחרות) ואז בשיטה מתקדמת המוכרת בשם "זקיפות מרובות" (multiple imputation).

הותאמו שני מודלים שונים וזאת על מנת לענות על כל אחת משתי שאלות המחקר:

- כדי לענות על השאלה הראשונה לגבי השינוי הממוצע הותאם מודל היררכי דו-רמתי (multilevel) למידול נתוני אורך (longitudinal data) הנקרא גם מודל של צמיחה (growth model) כאשר הרמה הראשונה היא המדידות החוזרות והרמה השנייה היא האנשים. נמצא כי הצמיחה מקוטעת ולכן הותאם מודל ספלייני עם נקודת שבירה אחת בהיבחנות השנייה
- כדי לענות על השאלה השנייה לגבי האפשרות לנבא סיכוי של נבחן להיבחן יותר מפעם אחת בבחינה הפסיכומטרית בעזרת משתני רקע, הותאמה רגרסיה לוגיסטית כאשר משתנה המטרה הוא נבחן פעם אחת בלבד = 0 לעומת נבחן יותר מפעם אחת = 1

תוצאות ומסקנות. הגידול הממוצע בציון הבחינה בין הבחינה הראשונה לשנייה הוא 45, והגידול הממוצע בין הבחינה השנייה לרביעית הינו 26.

משתנים דמוגרפים משפיעים הן על השינוי הממוצע בציון הבחינה הפסיכומטרית עבור אנשים שנבחנים מספר פעמים והן על הסיכוי שנבחן יבחן יותר מפעם אחת בצורות שונות.

חשיבות המחקר וחידושים. מחקר זה הינו מחקר ראשון המשתמש בשיטת הניתוח הרב-רמתי למידול הצמיחה בציוני המכפ"ל בהיבחנויות חוזרות, מודל סטטיסטי מתקדם שמאפשר להתמודד עם התלות הקיימת בציונים השונים של אותו נבחן, עם השונות שקיימת בין נבחנים שונים ועם העובדה שמערך בסיס הנתונים אינו מאוזן (קיימים דפוסים שונים של היבחנויות בנתונים שברשותנו). בנוסף, מחקר זה מראה כיצד משפיעים נתוני הרקע הדמוגרפי על דפוסי ההתנהגות של נבחנים בהיבחנויות חוזרות, ובכך למעשה משלים מחקר אחר שנעשה ע"י קלפר ואחרים (2015).

מדד מחירי הדירות בישראל: עבר, הווה ועתיד

דורון סייג, הלשכה המרכזית לסטטיסטיקה

מדד מחירי הדירות מיועד למדוד את השינוי הנקי במחירי מצבת הדירות לאורך זמן. המדד מחושב בלמ"ס בישראל החל משנות ה-50' על בסיס חודשי, בהתבסס על נתונים מנהליים מרשות המסים ועל מתודולוגיה סדורה שעודכנה במרוצת השנים. בשנים האחרונות בשל העלייה החדה במחירי הדירות וההתערבות הממשלתית, עומדים מדדי הדיור במרכז ההתעניינות הציבורית. ברמה העקרונית, בשל הטרוגניות גדולה הקיימת במאפייני הדירות וכן כתוצאה ממחזוריות נמוכה של מכירות בשוק הדירות, חישוב מדד מחירי הדירות מחייב מודל סטטיסטי המנכה הבדלי איכות בין הדירות שנמכרו, ושימוש במשקלות המתקנות את ההבדלים בין זרם הדירות שנמכרו ובין מלאי הדירות המהווה את אוכלוסיית המטרה של המדד. אחד האתגרים העומדים לאחרונה בפני הלמ"ס הינו קביעת מתודולוגיה מתאימה, בין השאר, המשקפת בצורה נכונה את משקלן של העסקאות שבוצעו במסגרת תכניות ממשלתיות ובהן: "מחיר למשתכן", "מחיר מטרה", וכן תכניות להתחדשות עירונית לסוגיהן. צורך בטיפול שונה בעסקאות אלו נבע מהייחודיות של אותן עסקאות (הגרלה, אי סחירות ועוד) בהשוואה לדירות הנמכרות בשוק החופשי. אתגר נוסף בנושא זה, הינו הרחבת התוצרים הסטטיסטיים על מחירי הדירות, כך שהמידע המתפרסם למקבלי ההחלטות ולציבור הרחב יכלול סטטיסטיקות מפורטות לפי סוגי דירות (למשל: דירות חדשות ביחס לדירות קיימות) ולפי פילוחים לשווקים מקומיים.

בשונה ממרבית מדדי המחירים האחרים, מדד מחירי הדירות אינו מחושב בעולם על סמך מתודולוגיה אחידה ומוסכמת, ושיטות חישוב שונות אינן מובילות בהכרח לאותם ממצאים. בחירת מתודולוגיית החישוב מתבססת בעיקר על שיקולים של איכות וזמינות הנתונים הנדרשים לצורך יישם שיטה זו או אחרת. לפיכך, אחד הסוגיות הראשוניות שיש לדון בהן בעת פיתוח מדד מחירי דירות, קשורה לבחירה מושכלת של מתודולוגיה מקובלת, המתאימה לנתונים הזמינים ולמגבלותיהם. בישראל, המידע על מחירי ומאפייני הדירות מפוזר על פני מספר מערכות ממשלתיות, והאחדת הנתונים המיועדת להקמת תשתית מידע מלא על ענף הנדל"ן למגורים נמצאת עדיין בשלבי הקמה ופיתוח. במקביל, נבדקות חלופות נוספות לטיוב והעשרת המידע אודות מלאי הדירות והדירות שנמכרו בפרט. שיפור מסד הנתונים צפוי לשפר את רמת הדיוק והאיכות של מדד מחירי הדירות, וכן יאפשר פיתוח מדדים נוספים המיועדים לשפוך אור על השינויים במחירי הדירות בפילוחים גיאוגרפיים שונים ולפי סוגי דירות.

היבט נוסף, הדורש טיפול בצד הנתונים קשור למהירות הדיווח של הנתונים אל הלמ"ס. מהירות הדיווח של הנתונים, ובפרט השלמת המידע על עסקאות שהתבצעו בעבר משפיע על גודל התיקון של האומדן הראשוני שמפורסם ביחס לתוצאה הסופית המתפרסמת לאחר שלושה תיקונים של האומדנים הזמניים. מאחר שהאומדן הראשוני מקבל את מירב תשומת הלב התקשורתית, ישנה חשיבות גדולה להקטנת התיקון ככל שניתן.

העבודה כוללת סקירה מתודולוגית של השיטות המדעיות המקובלת בעולם למדידת השינויים במחירי הדירות, סקירת התוצרים הסטטיסטיים המתפרסמים ברחבי העולם על-ידי לשכות סטטיסטיות, ריכוז ממצאים אמפיריים של בדיקות שנעשו בלמ"ס והצגת ההמלצות הראשוניות שגובשו על-ידי הוועדה המייעצת בנושא בינוי, דיור ונדל"ן במסגרת המועצה הציבורית לסטטיסטיקה.

שנתונים מתמחים - שנתון ירושלים ושנתון החברה החרדית

מאיה חושן, מכון ירושלים למחקרי מדיניות

אבסטרקט TBA

מושב III – סטטיסטיקה בתרופות (בשיתוף עם EMR-IBS)

יו"ר: ענת סאקוב, טבע

Immunogenicity of Biologic Drugs - With New Opportunities Come New Challenges

Oren Bar-Ilan, Non-Clinical Statistics, Global Specialty Development, Teva Pharmaceuticals Ltd,. Netanya, Israel.

Biologic drugs are defined as ones that are manufactured in living systems such as microorganisms, cells or tissues. These systems have been engineered to produce a specific molecule. Typically, biologic drugs are proteins that come to replace endogenous ones missing in treated patients or such that are targeted at

specific biological reactions to either inhibit or induce those reactions. Unlike chemical substances proteins are susceptible to create an immunological response by the immune system of patients. Such response may create allergic side effects or sometimes reduce the efficacy of a drug. Undesired immunogenicity response may have a severe effect on the safety and/or efficacy of a biologic product on patients. It is therefore important to identify and evaluate any such negative reactions at early stages of drug development. Regulatory bodies like the FDA and EMA require evaluation of the immunogenicity effects of a drug using statistical methods. The presentation will describe several types of immunogenicity studies and the statistical tools that are used to determine the immunogenicity effects of a drug.

Challenges and Opportunities of Data-Driven Approaches in Clinical Trials

Lena Granovsky, Shai Fine, Teva

The initiation of a new era of electronic health (eHealth) technology has erupted into industry, providing the ability to generate data on novel endpoints in clinical trials. However, no clear guidelines exist on how best to use the new technology, like wearable devices and mobile apps, in clinical trials, and to find ways that don't restrict the development of novel tech products. No clear pathway exists for validation and acceptance of trials which generate evidence through mobile technology and innovative methods versus a more controlled environment such as that routinely seen in pre-approval randomized trials.

One of the challenges is developing a robust methodological framework for design of clinical trials which utilise longitudinal data, such as activity data obtained from wearable devices. These studies can be used for developing and validating models for identifying patterns in the data, comparing patterns between populations, or investigating the relationship between patterns extracted from the longitudinal data and clinical outcomes. Crucial to the development of such a framework is the understanding the variability of the data within and between subjects, the occurrence of key features in the data and the relationship between these features and study endpoints. In this talk we suggest a practical methodology for calculating sample size and length of longitudinal data, which can be used in the clinical studies designed to prospectively validate algorithms that utilise wearable data.

ניסויים קליניים להערכת פוטנציאל שימוש לרעה בתרופות

סיון וייס, טבע

תרופות בעלות פוטנציאל שימוש לרעה ולהתמכרות הן תרופות המייצרות תחושת אופוריה, הזיות ושינויי מצב רוח אחרים. תרופות רבות לשיכוך כאבים הן בעלות פוטניצאל שימוש לרעה ומספר האנשים המתים משימוש יתר בתרופות אלו הולך ועולה. על מנת להילחם בתופעה זו, חברות התרופות מפתחות תרופות משככות כאבים (opioids) בעלות רכיב שמפחית פוטנציאל שימוש לרעה (abuse deterrent). על מנת להוכיח שתרופות אלו הן אכן בעלות פוטנציאל נמוך יותר לשימוש לרעה, יש לבצע ניסויים קליניים שיוכיחו זאת. ה-FDA דורש כיום לבצע ניסויים קליניים להוכחת פוטנציאל שימוש לרעה נמוך בתרופות כחלק מכל פיתוח תרופות בעלות רכיב הפעיל במערכת העצבים המרכזית.

בניסויים מסוג זה ישנם מספר אתגרים: מה תהיה אוכלוסיית הניסוי? מה תהיה קבוצת ההשוואה? מה יהיו המדדים שיבחנו את פוטנציאל השימוש לרעה? מה יהיו מבחני ההשערה שיבחנו את הפוטנציאל שימוש לרעה? מהו תכנון הניסוי המתאים לבחון השערות אלו? מה יהיה המודל הסטטיסטי המתאים?

בהרצאה זו אדון בשאלות הללו, אדבר על מאפייני הניסויים הקליניים של ניסויים אלו ואציג מספר דוגמאות.

בּוֹחַן כְּלָיוֹת וְלֵב

Johnathan Yefenof, Yuval Mathews, Daniel Odenheimer and Daniel Rothenstein, Quark Pharmaceuticals

The fundamental requirement for a “good statistical practice” is for the collected data to be reliable and supportive of its objectives. The data collection procedure has a crucial impact on the outcome. Hence, in the process of defining the study objectives the data collection methods has to be defined clearly as well. The importance of these issues is demonstrated in studies that look at Acute Kidney Injury (AKI) such as in an open heart surgery in which there is a risk that a patient undergoing surgery will develop AKI. The primary objective in such studies is to test whether the study drug decreases the rate of AKI in treated patients with respect to placebo treated patients. AKI is diagnosed on the basis of characteristic laboratory findings, such as urine volume over time. In the analysis presented it will be shown how the process of data collection may induce a technical bias which affects the outcome. Two approaches to overcome the bias issue will be introduced and discussed to enable to comply with the study objectives.

מושב IV – ביוסטטיסטיקה (בשיתוף עם EMR-IBS)

יו"ר: חבי מורד, מכון גרטנר

Standardization of mortality rates for hospital performance

Laurence Freedman[1], Ronen Fluss[1], Nethanel Goldschmidt[2] and Micha Mandel[3]

1. Gertner Institute for Epidemiology and Health Policy Research, Tel Hashomer, Israel
2. Ministry of Health, Jerusalem, Israel
3. Hebrew University, Jerusalem, Israel

When monitoring hospitals with regard to performance measures, it is now commonly accepted that some type of standardization be used to adjust for the distribution of profiles of patients treated at the hospital. In the epidemiologic literature two main types of standardization are described – direct and indirect. While direct standardized rates of different hospitals may be compared, the prevailing wisdom is that indirect rates are not strictly comparable. On the other hand, standard implementation of the direct method to the data of small hospitals is problematic, leading to large variances in the standardized rates. Perhaps because of this, in monitoring hospital performance indirect standardization is the more commonly used method. In this talk, we explore different methods of standardization and their advantages and disadvantages, illustrated by data on 30-day mortality after acute myocardial infarction in Israeli hospitals. The work is being carried out as part of the preparation for a new initiative of the Organisation for Economic Co-operation and Development (OECD) to monitor hospital performance in OECD countries.

שיטה לתחזית עכשווית של התפלגות הרוח המרחבית בגובה גגות הבתים בעיר מישורית

זיו קלויזנר, אייל פטל ,המכון למחקר ביולוגי, נס ציונה

תאור מאוחד ומייצג לרוח במרחב העירוני, ובכלל זה במקומות בהם לא קיימת מדידה, דורש התייחסות סטטיסטית לשדה הרוח המרחבי. הסיבה לכך היא שבאופן תדיר קיימות תקופות ביממה בהן שדה הרוח מאופיין באי אחידות מרחבית. תופעות כאלה יכולות לנבוע מתופעות זרימה שונות, כגון מעבר בין משטרי זרימה שונים או במהלך חדירת חזית מזג אוויר.

עבודה זו מתארת שיטה לתחזית עכשווית (nowcasting) סטטיסטית של הרוח במרחב. השיטה מבוססת על ניתוח אוסף הריאליזציות של הרוח בנקודות שונות מעל גגות הבתים המגדיר את ההתפלגות המרחבית. אנו מדגימים את

השיטה על גוש דן (איזור תל-אביב עד פתח-תקווה) ומראים שניתן באיזור זה לייצג בצורה טובה את ההתפלגות המרחבית באמצעות מדגם קטן של ארבע תחנות.

בהינתן מדידות זמן אמת של רוח מארבע התחנות המייצגות, השיטה מספקת הערכה עכשוית לאיזור טולראנס אליפטי סביב וקטור הרוח הממוצע. איזור טולראנס זה תוחם פרופורציה נתונה מתוך ריאליזציות הרוח בתא השטח העירוני.

Optimal flow rate sampling designs for studies with extended exhaled nitric oxide analysis

נועה מולשצקי, אוניברסיטת דרום קליפורניה

Introduction. The fractional concentration of exhaled nitric oxide (FeNO) is a biomarker of airway inflammation. Repeat FeNO maneuvers at multiple fixed exhalation flow rates (extended NO analysis) can be used to estimate parameters quantifying proximal and distal sources of NO in mathematical models of lower respiratory tract NO. A growing number of studies use extended NO analysis, but there is no official standard flow rate sampling protocol. In this paper, we provide information for study planning by deriving theoretically optimal flow rate sampling designs.

Methods. First, we reviewed previously published designs. Then, under a nonlinear regression framework for estimating NO parameters in the steady-state two compartment model of NO, we identified unbiased optimal four flow rate designs (within the range of 10–400 ml s−1) using theoretical derivations and simulation studies. Optimality criteria included NO parameter standard errors (SEs). A simulation study was used to estimate sample sizes required to detect associations with NO parameters estimated from studies with different designs.

Results. Most designs (77%) were unbiased. NO parameter SEs were smaller for designs with: more target flows, more replicate maneuvers per target flow, and a larger range of target flows. High flows were most important for estimating alveolar NO concentration, while low flows were most important for the proximal NO parameters. The Southern California Children's Health Study design (30, 50, 100 and 300 ml s−1) had.1.8 fold larger SEs and required 1.1–3.2 fold more subjects to detect the association of a determinant with each NO parameter as compared to an optimal design of 10, 50, 100 and 400 ml s−1.

Conclusions. There is a class of reasonable flow rate sampling designs with good theoretical performance. In practice, designs should be selected to balance the tradeoffs between optimality and feasibility of the flow range and total number of maneuvers.

Extent, Duration and Predictors of Exclusive Breastfeeding in a Longitudinal Study: Adjusting for missing data using an Accelerated Failure Time mode

סמאח חאיכ, המרכז הלאומי לבקרת מחלות

Background. The World Health Organization (WHO) recommends at least 6 months of exclusive breastfeeding (EBF). Longitudinal epidemiological studies facilitate estimation of the duration of EBF, but often suffer from loss to follow-up and missing information.

Objectives. While adjusting for missing data, (1) To estimate the proportion of Israeli women who practice EBF. (2) To estimate the distribution of duration of EBF. (3) To identify factors that predict the duration of EBF.

Methods. A longitudinal study was carried out including all women who gave birth between September 2009 and February 2010 in selected Israeli hospitals (N=2119). Participants reported information related to EBF, socio-demographic characteristics, and breastfeeding practices in the hospital and at two-monthly intervals thereafter. Information on EBF status and duration was missing for 35% of women. We imputed

EBF practice using logistic regression Multiple Imputation (MI) method with 20 repeats (procedure MI with option FCS, SAS 9.4) and using Rubin's rule estimated the probability of practicing EBF. Predicted probabilities of practicing EBF for women with missing information served as weights in the analyses of objectives 2-3. We imputed EBF duration based on an Accelerated Failure Time (AFT) model built on observed duration times, creating five complete data sets. We then estimated the distribution of duration in those practicing EBF using a weighted Kaplan-Meier curve (SAS 9.4) in each completed dataset and used Rubin's rule to estimate the time of EBF "survival" curve and its standard errors.

Results. 1: The observed proportion of women practicing EBF (complete case analysis) was 69% (95%CI; 66%-71%). After imputation, the estimated proportion changed to 65% (95%CI; 62%-68%).

2: After imputation, estimated percentiles the time of EBF among women practicing EBF were: 25%: 3.0m; 50%: 4.0m; 75%: 5.7m.

3: Predictors of EBF duration were: stated intention to BF - 50% increase (p=0.001); religious observance (secular vs. ultra-orthodox) - 22% decrease (p<0.001); giving formula milk in hospital - 11% decrease in EBF duration (p<0.001); using a pacifier in hospital - 10% decrease (p <0.001); ethnicity (Arab v Jew) – 9% decrease (p=0.06).

Conclusions. By imputing missing practice and duration of EBF we obtained estimated proportion and duration of EBF adjusted for the potential bias caused by missing information. Using an AFT model for EBF duration also allows direct interpretation of the impact of various factors on EBF duration.

Joint work with Havi Murad, Anneke Ifrah, Tamy Shohat, and Laurence Freedman.

מושב V – האתגרים הצפויים לתחום האקטואריה בעשור הקרוב

יו"ר: בני יקיר, האוניברסיטה העברית

משתתפים:

- סטיוארט קוטס, אקטואר יועץ (ביטוח כללי), ראש התכנית לאקטואריה במכון מגיד
- ענת אלעד, אקטוארית יועצת (בריאות)
- דוד אנגלמייר, אקטואר יועץ (חיים, פנסיה)

שינויים רבים בעולמנו התממשו או צפויים להתממש בשנים הקרובות. שינויים אלה ישפיעו על עולם הביטוח ויהפכו את פניו. הרשימה ארוכה. כדוגמאות בולטות ניתן לציין את התארכות משך החיים, וכפועל יוצא הנטל האקטוארי הנוצר על קרנות הפנסיה וכן הצורך הרפואי לטפל באוכלוסיות מזדקנות. הזינוק בהיקף האינפורמציה הנאספת על כל פרט ופרט ופיתוח הטכנולוגיות להפקת תובנות מנתונים אלה משנים את פני הרפואה וצפויות להן השלכות משמעותיות בתחום ביטוח הבריאות. במקביל, טכנולוגיות דומות לניתוח נתונים ממומשות כבר כיום בעזרים להגברת בטיחות הנהיגה ויאפשרו בקרוב נהיגה אוטונומית. שינויים אלה יחייבו ארגון מחודש של כל תחום הביטוח הרכב, תחום שהוא מרכזי למודלים העסקיים של חברות הביטוח.

מדע האקטואריה, העוסק בפיתוח מודלים סטטיסטיים ופיננסיים לפתרון בעיות בתחום הביטוח, יצטרך להיערך מחדש כדי להתמודד עם האתגרים שבפתח. בפרט, יהיה על העוסקים בתחום ללמוד כיצד לשלב את הכלים של הסטטיסטיקה המודרנית ומדע הנתונים בעולם הביטוח.

במושב זה נארח שלושה מומחים מתחום הביטוח שיציגו את האתגרים השונים מנקודת מבטם. בחלק הראשון של המושב יציג כל אחד מן המומחים היבטים רלוונטיים מתחומי מומחיותם. החלק השני יתנהל במבנה של שאלות ותשובות – שאלות שיציג המנחה ושאלות מן הקהל. המטרה של המושב היא לעורר את עניינם של אנשי הסטטיסטיקה ומדע הנתונים באתגרים שנובעים מתחום הביטוח ולעודד אותם להשתתף במאמץ לפתח פתרונות לצרכים החדשים.

Integrating R based engines in other systems

Adi Sarid, adisarid@gmail.com

One of R's limitations is the inability to compile it into binary files (i.e., executable programs). This makes it challenging to integrate statistical software written in R into other third party applications.

To illustrate, assume that a time series forecast engine has been developed in R, and is tailor made to a specific customer. The natural thing would be to integrate this engine back into the working environment of the customer (e.g., an ERP). However, migrating this engine might be an excruciating task, especially if the R engine has dependencies such as the `forecast` package (or other complex statistical packages).

In my talk, I will outline a number of methods to integrate an R based engine into an existing ecosystem (which is not written in R). Starting from the most basic methods such as scheduling automations, and up to more advanced techniques such as using cloud based (e.g., `Shiny`).

Short bio. Adi is a PhD student in the department of industrial engineering in Tel-Aviv university. His PhD is on the topic of electrical grid optimization and design for robustness. In parallel, Adi is also a partner and the head of operations research department in the Sarid Research Institute LTD.

MultiNav - Multivariate Exploratory Data Analysis

Efrat Vilenski, efratvil@gmail.com

Introducing MultiNav R package - a set of tools for semi-automatic multivariate EDA. MultiNav is an interactive visualization platform, incorporating ideas and algorithms from several fields to aid the analyst in the exploration. These include (naturally) multivariate statistics, but also robust statistics, process control, unsupervised machine learning, and network data analysis. The goal of MultiNav is to help to uncover interesting patterns and outliers as well as provide quick understanding of the important characteristics of a given multivariate dataset.

The package was developed based on use-cases from anomaly detection projects for large sensor networks, and machine failure. Early prototypes of MultiNav were generalized to receive standard multivariate inputs, so that it could be used on arbitrary data for multivariate EDA and process control problems.

Short bio. Efrat Vilenski is a PhD candidate of Industrial Engineering and Management at Ben Gurion University. Supervised by Jonathan Rosenblatt. Interest areas: visual analytics, multivariate methods, process monitoring.

R כאמצעי הוראה: הילכו סטטיסטיקה ועובדים סוציאליים יחדיו – אף אם לא נועדו?

יונתן אנסון, המחלקה לעבודה סוציאלית, אוניברסיטת בן-גוריון בנגב, jesa47@gmail.com

הסטטיסטיקה במחוזות מדעי החברה היא לרוב מעין טקס: מעלים עולה, בצורת נתונים, בפני אל המחשבים וממתינים למוצא פיו החורץ גורלות. המובהק הריח באפו, התפל הוא? הוא לא מדבר ישירות, כמובן, אלא דרך האורקל, דובר הנסתרים (SPSS), אך אם יודעים להצביע וללחוץ על הכפתורים הנכונים (point and click) סביר שהאל

יתרצה, הכוכביות יופיעו במקומות הנכונים ואפשר לרוץ ולפרסם! עתה נדרשים הכוהנים הוותיקים ללמד את טירוני הדת, החרדים בפני האל הנורא, את מהלכי הטקס . . .

מזה מספר שנים אני מנסה ללמד "שיטות כמותניות" (את המילה סטטיסטיקה הס מלהזכיר!) בדרך אחרת, לא כטקס אלא כתהליך אינטראקטיבי בו המטרה היא לתחקר את הנתונים. לרוב המספרים האימתניים מסתירים בחיקם סיפור והאתגר הוא לגלות מה הוא (משל הפסל המגלף את העץ כדי לגלות את הציפור היושבת בתוכו). לשם כך נדרשת שפת הידברות עם הנתונים, והשפה בה אני מדבר איתם היא R. נדרש, כמובן, מאמץ מצד הסטודנטים, להתגבר על החרדה הראשונית, ללמוד את השפה – וגם מצדי לתקן את כל הטעויות המשתקות הראשונות. אולם יש לזכור, החרדה לא פחותה בהוראת שיטת הטקס, והתוצאות הרבה פחות מספקות, גם לסטודנט (שלרוב לא מבין למה הוא עושה את מה שהוא עושה) וגם לא למרצה.

בהרצאה זו אציג את ההיגיון מאחורי הקורס, דרך התנהלותו, ההצלחות והכישלונות וקצת מתגובות הסטודנטים.

ביוגרפיה קצרה: דמוגרף / סוציולוג, דוקטורט 1985 מאוניבסיטת Brown, ארה"ב. לימדתי מ-1985 עד 2016 במח' לעבודה סוציאלית באוניברסיטת בן גוריון, לרוב קורסים בסוציולוגיה ובסטטיסטיקה. היום בפנסייה אך ממשיך ללמד סטטיסטיקה (שלוחת אילת) ולהנחות סטודנטים לתואר שני ושלישי

ישום יסודי של שיטות ביואינפרמטיות בניתוח נתוני החומרים הנדיפים במיצוי מאוכלוסיות הבר של אזוב מצוי בישראל

עמית פליס ונתיב דודאי, המחלקה לצמחי תבלין ורפואה, מרכז מחקר נוה יער, רמת ישי, ישראל, fliess@gmail.com

כדי לחקור את השונות בהרכב הכימי של החומרים הנדיפים באוכלוסיות הבר של אזוב מצוי בישראל, נעשה אישכול (clustering) היררכי דו-מימדי ע"י הפקודה heatmap על טבלה מנורמלת של נתוני מיצויי החומרים הנדיפים מצמחי אזוב שונים. למרות הנירמול, התוצאות עדיין אופיינו ע"י המרכיבים העיקריים. התקבלו 3 אשכולות: תימולי, קרבקרולי, ותימולי+קרבקרולי. אישכול ללא המרכיבים העיקריים תימול וקרבקרול נתן את אותה תוצאה. 3 הכמוטיפים אופיינו כמותית לפי החציון בפקודת median והסטיה ממנו בפקודת mad, כי נתונים ביולוגיים בדרך כלל לא מתפלגים נורמלית. החומרים, שנמצאו נבדלים משמעותית בין הכמוטיפים, קשורים לאותו מסלול ביוסינטטי. ב-2 הכמוטיפים העיקריים חושבה התפלגות צפיפות ההיסתברות בפקודה density של אחוזי חומרי המוצא, המרכיב העיקרי ותוצריו. בטיפוס התימולי היה אחוז גבוה יותר של חומרי מוצא ואחוז נמוך יותר של החומר העיקרי והתוצרים המשותפים מבטיפוס הקרבקרולי. למרות זאת יש חפיפה ניכרת בין הטיפוסים.

רקע ב-R: ד"ר בביואינפורמטיקה, כלומר עבודה בלמידה לא מונחת של נתונים בעלי התפלגות לא לינארית ולא נורמלית בעיקר באישכול (clustering) כולל אישכול היררכי דו-מימדי ודנדוגרמות לא מושרשות.

Exploring the exploration of pre-walking infants

Tzviel Frostig, tfrostig@gmail.com

We are analyzing the exploration of typically developed human pre walking infants in the presence of their relatively passive mothers in a new environment while using measures and structured established in previous work in origin related exploration of rodents. I will show how R's strong visualization capabilities can be used to shed light on some of those phenomenon, and how to use knitR to create a reproducible research.

Short bio: Tzviel Frostig is MSc student of statistics at Tel-Aviv University.

Predicting the probability of a Severe bodily injury in a Car accident in Tel Aviv

Jonathan Schwartz, yoni.schwartz1989@gmail.com

A major part of practical actuarial problems in general insurance is the rating of risk factors in motor insurance. The understanding of what affects the likelihood and severity of car accidents is important not just to the insurance sector but to anyone who uses a car as a means of transportation.

In Israel, every accident which involves a bodily injury must be reported to the police. Due to this a reliable database was put together by the central bureau of statistics. This can give way to an analysis of the main risk factors making bodily injury car accidents lethal in the city of Tel Aviv, i.e., what are the main risk factors that affect the severity of car accidents.

Short bio: Jonathan Schwartz is an Actuarial Science Masters student at the University of Haifa and an actuarial analyst at the Phoenix Insurance Company.

http://rpubs.com/jonathans/actuar
https://jonathans.shinyapps.io/Severity

RcppArmadillo – how to easily use C++ in R

Yair Goldberg, yair.goldy@gmail.com

The Rcpp package enables the use of C++ code in R packages. This integration typically leads to a code that can run much faster. However, not everyone knows C++ which makes the Rcpp package hard to access. Armadillo is a fast and easy to use C++ linear algebra library which uses similar syntax to that of R. In this talk, I will present the RcppAramdillo package which provides an interface between R and the C++ Armadillo library.

Short bio: Yair Goldberg is a researcher at the Department of Statistics at the University of Haifa and teaches R courses.

heatmaply: an R package for creating interactive cluster heatmaps for online publishing

Tal Galili, ta.galili@gmail.com

heatmaply is an R package for easily creating interactive cluster heatmaps that can be shared online as a stand-alone HTML file. Interactivity includes a tooltip display of values when hovering over cells, as well as the ability to zoom into specific sections of the figure from the data matrix, the side dendrograms, or annotated labels. Thanks to the synergistic relationship between heatmaply and other R packages, the user is empowered by a refined control over the statistical and visual aspects of the heatmap layout.

The heatmaply package is available under the GPL-2 Open Source license. It comes with a detailed vignette, and is freely available from: http://cran.r-project.org/package=heatmaply

Short bio: Tal Galili is PhD candidate of statistics at Tel-Aviv University. He blogs about R at r-statistics.com.

פאנל

סַקֶרֶת – מחלה או תרופה?

מנחה: צבי גילולה, האוניברסיטה העברית בירושלים

משתתפים:

- מנו גבע, מדגם – יעוץ ומחקר
- קמיל פוקס, אוניברסיטת תל-אביב
- רון קנת, KPA – תובנות באמצעות אנליטיקה

הפאנל ידון בשני נושאים:

1. סקרים תחת מתקפה – האם הרושם שמתקבל על אי-דיוק עקבי בסקרי יום בחירות בארץ (ובארה"ב בבחירות האחרונות) נכון? ואם כן, למה לא....?
2. סקרי פאנל אינטרנט – האם הם "עובדות אלטרנטיביות" בעלות איכות מתודולוגית ראויה?