קורס 52414 R - 2021-22 מבחן בית

הוראות:
יש להגיש את קובץ Rmd המלא וקובץ הhtml שנוצר ממנו. הוראות אלו כאן מסבירות בעברית את שאלות המבחן. בכל, מקרה הנוסח המחייב הוא בקובץ ה-Rmd של המבחן ועל קובץ הפתרון להשתמש בו ולהתייחס אליו.
אם נראה לכם שיש הבדל במשמעות ההוראות בין הניסוח כאן לניסוח ב-Rmd, שאלו בפורום/שעות קבלה.

1. בשאלה זו נדגום נקודות באופן אקראי מכדור הארץ. אנו ממדלים את כדור הארץ ככדור בעל רדיוס r=6371 ק"מ (מזניחים הבדלים קלים בגובה ובצורת כדור הארץ עצמו שאינה כדור מושלם).

ניתן לציין נקודה על כדור הארץ (כלומר על שפת הכדור) בקואורדינטות קרטזיות (x,y,z), כאשר הראשית (0,0,0) מציינת את מרכז הכדור, וכל נקודה על שפת הכדור מקיימת $x^2 + y^2 + z^2 = r^2$

באופן אלטרנטיבי, ניתן לייצג ע"י קואורדינטות ספריות (r, θ,Φ) כאשר Φ היא זוית בין 0 ל-2π המייצגת את קו האורך ברדיאנים, כאשר קו גריניץ' הוא 0, והזוית עולה כאשר נעים מזרחה: למשל היא שווה ל-π כאשר משלימים חצי סיבוב סביב כדור הארץ ומגיעים לקו התאריך הבינלאומי, ומתקרבת ל-2π כאשר משלימים סיבוב מלא.
θ היא זוית בין 0 ל-π המייצגת את קו הרוחב ברדיאנים, כאשר הערך הוא π עבור הקוטב הצפוני, π/2 עבור קו המשווה ו-0 עבור הקוטב הדרומי.

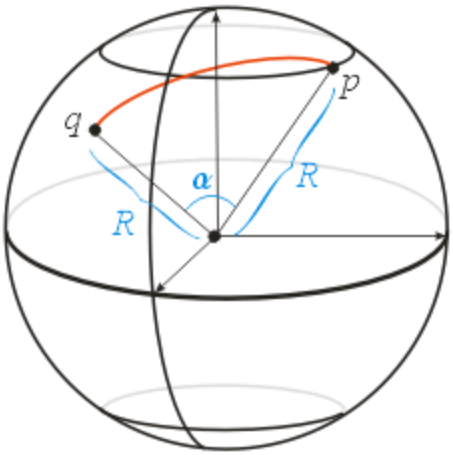
הערה: דרך מקובלת מעט אחרת לייצג נקודות על כדור הארץ היא כאשר קו האורך הוא בין 180- ל180 מעלות, כאשר קו גריניץ הוא 0, ומעלות חיוביות מציינות נקודות שהן מזרחית לו ומעלות שליליות מציינות נקודות שהן מערבית לו. קו הרוחב הוא בין 90- ל90 מעלות, כאשר הערך הוא 90 עבור הקוטב הצפוני, 0 עבור קו המשווה ו90- עבור הקוטב הדרומי. אנו נקרא לדרך זו בשם **קואורדינטות גאוגרפיות,** וניתן לעבור באופן פשוט בין טווחים אלו לטווחים של הקואורדינטות הספריות.

עבור שתי נקודות על שפת הכדור (הנקראית גם הספירה), ניתן להגדיר באופן הרגיל את המרחק האוקלידי, שהוא אורך הקו הישר המחבר בין הנקודות (עובר דרך הכדור). עבור הנקודות בקואורדינטות קרטזיות $(x_1, y_1, z_1), (x_2, y_2, z_2)$ המרחק הוא $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$

ניתן גם להגדיר את המרחק הגיאודזי (נקרא גם מרחק ספרי), שהוא אורך הקשת הקצרה ביותר המחברת את הנקודות על שפת הכדור, כפי שמסומן באדום בציור למטה (זהו המרחק בין הנקודות p,q בקו אווירי). עבור הנקודות בקואורדינטות ספריות $(r, \theta_1, \phi_1), (r, \theta_2, \phi_2)$ המרחק הוא

$r\, cos^{-1}(sin(\theta_1)sin(\theta_2) + cos(\theta_1)cos(\theta_2)cos(\phi_1 - \phi_2))$.

לדוגמא, אם $\phi_1 = \phi_2$ המרחק הופך ל $r|\theta_1 - \theta_2|$.



a. עליכם לאמוד את המרחק הגיאודזי הממוצע בין 2 נקודות המתפלגות באופן אחיד על שפת הכדור. עשו זאת בעזרת סימולציה של 1000 זוגות נקודות המתפלגות אחיד על שפת הכדור, וחישוב המרחק הגיאודזי עבור כל זוג.
ניתן לדגום נקודות באופן אחיד בעזרת הפונקציה runif_on_sphere מהחבילה uniformly.
כעת, חיזרו על הסימולציה, אבל הפעם חשבו את המרחק האוקלידי עבור כל זוג נקודות, כדי לאמוד את המרחק האוקלידי הממוצע בין נקודות המתפלגות באופן אחיד על שפת הכדור.
בונוס: פתחו נוסחאות מתמטיות עבור המרחק הגיאודזי והמרחק האוקלידי הממוצע, והשוו אותן לתוצאות אותן קיבלתם מהאומדים.

b. חזרו על הסימולציות בסעיף הקודם, אך הפעם שמרו את כל ההתפלגות האמפירית של המרחקים בין הזוגות והציגו את שתי ההתפלגויות.
חשבו והדפיסו את ה-skewness ואת ה-kurtosis של שתי ההתפלגויות.
איזו מהן נראית קרובה יותר להתפלגות הנורמלית?

c. כתבו פונקציה המקבלת כקלט מספר n ומחרוזת המייצגת ארץ מסויימת, ודוגמת n נקודות באופן אחיד מארץ זו בעזרת שיטת דגימת דחייה (rejection sampling). תוכלו לבדוק אם נקודה שייכת לארץ מסויימת בעזרת הפונקציה map.where מהחבילה maps (יש להשתמש בקואורדינטות גיאוגרפיות). הריצו את הפונקציה לדגימת 1000 זוגות נקודות אקראיות בארצות הברית והשתמשו בדגימה כדי לאמוד את ממוצע המרחק הגיאודזי בין 2 נקודות אקראיות בארצות הברית.
הסבירו במילים עבור אילו מדינות שיטת דגימת דחייה היא בעייתית ולמה, וכיצד הייתם משפרים אותה (אין צורך בקוד/ניתוח נתונים עבור חלק זה).

d. סטטיסטיקאי מציע לכם שתי שיטות לדגימת נקודות באופן אחיד על שפת הכדור:
בשיטה הראשונה דוגמים באופן בלתי תלוי מההתפלגויות האחידות, $\theta \sim U[0,\pi], \phi \sim U[0,2\pi]$. הוקטור המתקבל (r,θ,ϕ) מייצג את הנקודה בקואורדינטות ספריות. ניתן להשתמש בטרנספורמציה כדי לעבור לקואורדינטות קרטזיות.
בשיטה השניה, דוגמים באופן בלתי תלוי את x,y,z מתוך ההתפלגות הנורמלית הסטנדרטית N(0,1) . לאחר מכן מנרמלים את x,y,z ע"י הכפלת כל אחד מהם בערך $\frac{r}{\sqrt{x^2+y^2+z^2}}$. הערכים המתקבלים (x,y,z) לאחר הנרמול מייצגים את הנקודה הנדגמת בקואורדינטות קרטזיות. ניתן להשתמש בטרנספורמציה כדי לעבור לקואורדינטות ספריות.
עבור כל אחת מהשיטות, עליכם לקבוע האם היא אכן מובילה לדגימת נקודות מתוך ההתפלגות האחידה על הספירה. מותר להשתמש בטיעונים מתמטיים ו/או פלטי סימולציה/גרפים לבחירתכם. זכרו שידוע שהפונקציה runif_on_sphere אכן דוגמת באופן נכון מתוך ההתפלגות האחידה על הספירה, וניתן להשוות את תוצאות שתי השיטות לדגימות המתקבלות בעזרת פונקציה זו.

2. בשאלה זו נקרא וננתח נתונים גיאוגרפיים על מדינות העולם.

a. קראו את דף הויקיפדיה המכיל נתונים על מדינות העולם על פי שטח. קראו את ה-html, הוציאו ממנו את הטבלה ושמרו את הנתונים כ-data-frame. ניתן לעשות זאת בעזרת פונקציות מהחבילה rvest.
הסירו שורות בהן שם המדינה מכיל סוגריים. הפכו את העמודה המייצגת את השטח לעמודה נומרית עם שטח בקילומטרים רבועים. במקרים בהם יש יותר משטח אחד עבור מדינה, השתמשו במספר הראשון. הציגו את שתי השורות הראשונות ושתי השורות האחרונות של ה-data-frame המתקבל.

b. חזרו על הסעיף הקודם, אך הפעם עבור דף הויקיפדיה המכיל נתונים על מדינות העולם על פי אוכלוסיה. כאן במקום השטח, יש להפוך לנומרית את העמודה המכילה את האוכלוסייה של כל מדינה.

c. אחדו את שני ה-data-frames והוסיפו עמודה נוספת הנקראית pop.density המכילה את צפיפות האוכלוסייה (מספר אנשים לק"מ מרובע) בכל מדינה. הציגו מפה של העולם (ניתן להשתמש בחבילה rworldmap) בה כל מדינה צבועה על פי צפיפות האוכלוסין שלה.
בנוסף, הציגו את 3 המדינות הצפופות ביותר ו3 המדינות הדלילות ביותר.

d. נניח שבוחרים אנשים באופן אקראי ואחיד מכלל אוכלוסיית העולם, כאשר מניחים שהמיקום של כל אדם מתפלג באופן אחיד בשטח המדינה שלו.
אנו רוצים לאמוד את המרחק הגיאודזי הממוצע בין זוג אנשים תחת הנחות אלו. דיגמו 1000 זוגות אנשים באופן זה: תחילה את המדינה עבור כל אדם, ואח"כ את המיקום בהינתן המדינה. חשבו את המרחקים הגיאודזים עבור זוגות אנשים שדגמתם בדרך זו והשתמשו בהם לאמידת המרחק הגיאודזי הממוצע בין אנשים תחת הנחות אלו. השוו את המרחק המתקבל למרחק הממוצע בין זוג נקודות אקראיות על שפת הכדור.

e. נניח שלא ידענו את שטחי המדינות, והיינו רוצים לאמוד אותם בעזרת דגימה. כתבו פונקציה המקבלת וקטור עם שמות מדינות, וכן מספר דגימות מבוקש n, ומחזירה אומדים לשטח של כל מדינה. הפונקציה תעשה זו בעזרת דגימת n נקודות באופן אקראי על שפת הכדור, וחישוב השכיחות היחסית של הנקודות המתקבלות בתחומי כל מדינה.
הריצו את הפונקציה עבור n=10000 ועבור רשימת המדינות המופיעות ב-data-frame המתקבל מויקיפדיה. הציגו בטבלאות את 10 המדינות עבורן האומדן לשטח הוא הגדול ביותר, ואת 10 המדינות עבורן השטח האמיתי הוא הגדול ביותר. עד כמה יש הסכמה בין שתי הרשימות?
בנוסף, מזגו את תוצאות האומדנים לשטח עם ה-data-frame מויקיפדיה, והציגו גרף המשווה את השטח האמיתי (בציר x) מול השטח הנאמד (בציר y). מהו ה-R^2 בין השטחים הנאמדים לאמיתיים?

3. בשאלה זו קוראים ומנתחים נתוני רעידות אדמה

a. קראו את נתוני רעידות האדמה מכאן. יש לבחור רעידות מכל העולם עם עוצמה של מעל 2.5 (בסולם ריכטר), במהלך כל שנת 2022. יש לשמור את הנתונים בקובץ csv ולאחר מכן לטעון אותו ל-R כ-data-frame.
הציגו את 5 רעידות האדמה המאוחרות ביותר, וכן את 5 רעידות האדמה העוצמתיות ביותר על פי magnitude.

b. הציגו גרף עם רעידות האדמה כנקודות על פני העולם, כאשר ציר x מייצג את קו האורך (longitude), ציר y מייצג את קו הרוחב (latitude) והגודל מייצג את עוצמת הרעידה (mag). הציגו זאת בעזרת ggplot על פני מפת העולם. את מפת העולם ניתן להציג בעזרת הפונקציה geom_map מתוך החבילה ggmap. את ה-data-frame המייצג את המפה עצמה ניתן לייצר מתוך הפלט של הפקודה map_data("world")

c. חזרו על הגרף מהסעיף הקודם, אבל הפעם הפרידו את רעידות האדמה לקבוצות על פי העוצמה (בין 2 ל-3, בין 3 ל-4,..., בין 7 ל-8), כאשר לכל קבוצה גרף נפרד (על פני המפה). מה ניתן להסיק לגבי מיקומים של רעידות אדמה חזקות וחלשות?

d. קראו את הנתונים בעמודה המייצגת את הזמן, לצורך יצירת שתי עמודות חדשות מספריות: עמודה חדשה בשם day-of-year המייצגת את היום מתחילת השנה, ועמודה חדשה נוספת בשם time-of-day המייצגת את השעה מחצות (כשבר).
לדוגמא, עבור "2022-03-10T17:40:28.123Z" היום המתקבל הוא 10+28+31=69 (מספרי הימים בחודשים הראשונים בשנה הם ינואר:31, פברואר:28, מרץ:31, אפריל:30, מאי:31), והשעה המתקבלת היא 28/3600+40/60+17=17.6744 שעות מאז חצות (מתעלמים מחלקי שניה).

e. לאחר מכן, הציגו גרפים (scatter) וקורלציות עבור כל זוג משתנים מבין: mag, depth,day-of-year,time-of-day.
ניתן לעשות זאת בעזרת הפונקציה ggpairs מתוך החבילה GGally.
עבור אילו זוגות הקורלציה היא סיגניפיקנטית?

f. רוצים לבחון האם יש זמנים במהלך היממה בהם יש יותר/פחות רעידות אדמה.
באופן פורמלי, אם t_i הזמן מחצות בשעות של רעידת האדמה ה-i, אז בודקים את השערת האפס $H_0: t_i \sim U[0,24]$ מול האלטרנטיבה המורכבת H_1 עבורה ההתפלגות של t_i אינה ההתפלגות האחידה $U[0,24]$. לשם כך מחשבים את הסטטיסטי: $S = \sum_{i=1}^{24} \frac{(o_i - e_i)^2}{e_i}$ כאשר o_i מספר רעידות האדמה הנצפה בשעה i, למשל בין 00:00:00 לבין 00:59:59 עבור i=0, וכאשר $e_i = \frac{n}{24}$ הוא המספר הצפוי על פי ההתפלגות האחידה, כש-n הוא המספר הכולל של רעידות האדמה.
חשבו את הסטטיסטי על הנתונים, וכן חשבו p-value בעזרת קירוב להתפלגות תחת השערת האפס ע"י התפלגות חי-בריבוע עם 1-24 דרגות חופש.
האם תדחו את השערת האפס עבור רמת מובהקות $\alpha = 0.01$?