**פתיח על קובץ הנתונים: student performance**

מקור הקובץ שלנו הוא באתר Kaggle (<https://www.kaggle.com/datasets>). קובץ הנתונים שלנו מכיל נתונים על מדגם של 1000 ילדים בגילאים זהים מתוך אוכלוסייה בעלת 5 קבוצות אתניות שונות. הילדים עברו מבחן הכולל 3 חלקים (מתמטיקה, קריאה, כתיבה). ציוני הילדים המוצגים הם על כל אחד מן החלקים במבחן הכולל.  
**המשתנים:**

* ID- מספור של כל תצפית/ נבדק. משתנה קטגוריאלי (בסולם שמי- אין חשיבות לסדר המספור היות שזה מקרי לחלוטין). מספור מ-1 עד 1000.
* gender- מגדר הילד. משתנה קטגוריאלי (בסולם שמי). מוגדר ע"י הערכים: "male"   
  ו-"female" (בן/ בת).
* race/ethnicity- גזע/ מוצא אתני. משתנה קטגוריאלי (בסולם שמי). מוגדר ע"י הערכים: "group A", "group B", "group C", "group D", ו-"group E", היות שבאוכלוסייה ממנה נלקח המדגם קיימות 5 קבוצות אתניות שונות.
* parental level of education- רמת השכלה של הורה. משתנה קטגוריאלי (בסולם סדר). מוגדר ע"י הערכים (מסודר מהגבוה לנמוך): "master's degree"- תואר שני, "bachelor's degree"- תואר ראשון, "associate's degree"- לימודי תעודה,   
  "some college", "high school", "some high school".
* lunch- ארוחת צהרים. משתנה קטגוריאלי (בסולם סדר). מוגדר ע"י הערכים: ארוחת "standard"- ארוחה במחיר מלא, ו"free/reduced"- ארוחת צהרים מוזלת או חינמית.
* test preparation course- קורס הכנה למבחן. משתנה קטגוריאלי (בסולם שמי). מוגדר ע"י הערכים: "none" ו-"completed", כלומר- ישנם ילדים שהשלימו קורס הכנה לכלל חלקי המבחן ויש כאלה שכלל לא עשו קורס הכנה.
* math score- ציון הילד בחלק המתמטיקה במבחן. משתנה רציף (בסולם רווח).
* reading score- ציון הילד בחלק הקריאה במבחן. משתנה רציף (בסולם רווח).
* writing score- ציון הילד בחלק הכתיבה במבחן. משתנה רציף (בסולם רווח).

**הערות שרלוונטיות למהלך העבודה:**

* הוספנו בR משתנה רציף נוסף (בסולם רווח), באמצעות פונקציית apply. המשתנה הוא TestAVG- ממוצע הציונים של 3 החלקים במבחן של כל ילד. כלומר- ציון ממוצע של כל ילד. את המשתנה הנ"ל הוספנו מטעמי נוחות (הקשורים בשאלה השנייה שלנו).
* גם במקרים בהם קיבלנו שהנחות של מבחן מסוים אינן מתקיימות (באם נורמליות או שוויון שונויות), המשכנו בניתוח באמצעות המבחן, תוך ציון בתת-סעיף 'הנחות' שהנחה מסוימת אינה מתקיימת, בהתאם להנחיות העבודה.
* שימוש בקובץ הנתונים, מלווה בהנחה שהנתונים המוצגים בו הם מדגם המהווה דגימה מקרית ומייצגת מתוך אוכלוסיית הילדים בני אותו הגיל. כלומר, בכל אחד מן הסעיפים, גם אם לא צוין בקוד ב-R, קיימת הנחה של דגימה מקרית.

**שאלה מספר 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| שם הקבוצה | שם המשתנה | מספר תצפיות (גודל המדגם) | ממוצע | סטיית תקן |
| male | BoysMathGrades | 482 | 68.7282 | 14.35628 |
| female | GirlsMathGrades | 518 | 63.6332 | 15.49145 |

**נתונים- סטטיסטיקה תיאורית:**

(שני המשתנים המופיעים הם וקטורי הציונים שיצרנו ב-R עבור כל אחת מהקבוצות). **שאלת המחקר:**

האם קיים הבדל בין תוחלות הציונים של בנים ובנות במתמטיקה?  
שיערנו שברמת מובהקות של 5%, קיים הבדל בין תוחלת הציונים במתמטיקה של הבנים ותוחלת הציונים במתמטיקה של הבנות באוכלוסייה.   
**הסבר לגבי הניתוח:**

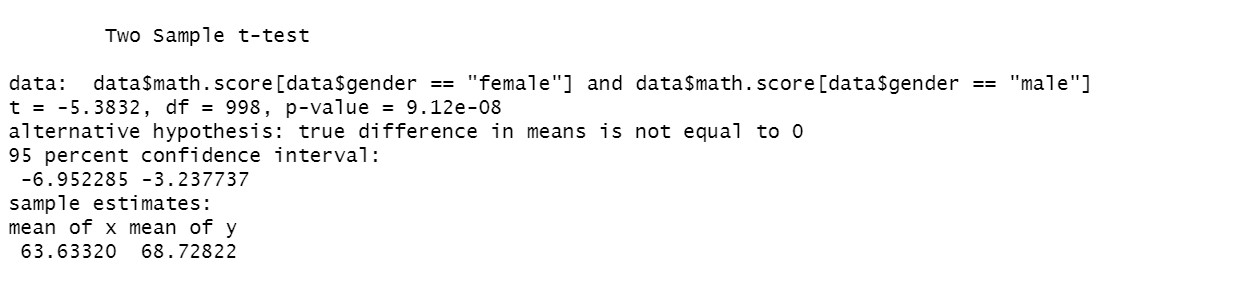
הפרמטר הוא ההפרש בתוחלות הציונים, ההשערה דו-זנבית, ושונות האוכלוסייה לא ידועה לנו. לכן, המבחן הסטטיסטי המתאים הוא מבחן t להפרש תוחלות, דו-זנבי למדגמים בלתי-תלויים.  **הנחות המבחן:**

1. דגימה מקרית
2. המשתנה מתפלג נורמלית באוכלוסיות- השתמשנו ב3 פרוצדורות לבדיקת הנחת נורמליות: הסתכלנו על הנתונים הגולמיים בהיסטוגרמות, שימוש ב-QQ plot (שיטה גראפית הבוחנת האם מדגם מתפלג נורמלית), הראה כי הנקודות אינן מסתדרות על קו ישר, ובנוסף קיבלנו תוצאה מובהקת במבחן הסטטיסטי לנורמליות- Shapiro-Wilk. הממצאים מראים על סטיה מנורמליות. לכן, קבענו כי הנחת הנורמליות לא מתקיימת.
3. שיווין שונויות בשתי האוכלוסיות- השתמשנו ב- Welch t testלהשוואת ממוצעים ב"ת שאינו דורש שוויון שונויות. בכל זאת, וכפי שהוצג בקורס, ערכנו מבחן F להשוואת שונויות ע"י שימוש ב-Levene's test שנתן תוצאה לא מובהקת, כלומר- השונויות שוות.

**תוצאות:**

t (df= 998) = -5.3832 , p = p value = 9.12\*10^-8   
התקבלה תוצאה מובהקת (p=9.12\*10^-8), לכן נדחה את השערת האפס.   
(הוספנו בקוד שהשונויות שוות למרות שהמבחן לא מחייב, וזו הסיבה לכך שdf=998).   
**מסקנה:**

נסיק כי תוחלות הציונים במתמטיקה של בנים ובנות באוכלוסיה שונות באופן מובהק.

**הפלט ב-R:**

**מענה על סעיף 6:**

לטובת מענה על השאלה ביצענו מבחן T בייסאני בתוכנת Jasp. ניתן לראות כי BF10 הינו 93698 – המשמעות הינה שהנתונים משמעותית יותר סבירים תחת השערת H1, הטוענת כי ציוני המתמטיקה של בנים ובנות מגיעים מוכלוסיות שונות.

**שאלה מספר 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| שם הקבוצה | שם המשתנה | מספר תצפיות (גודל המדגם) | ממוצע | סטיית תקן |
| group A | grades- group A | 89 | 62.9925 | 14.4446 |
| group B | grades- group B | 190 | 65.4684 | 14.7321 |
| group C | grades- group C | 319 | 67.1317 | 13.8722 |
| group D | grades- group D | 262 | 69.1794 | 13.2528 |
| group E | grades- group E | 140 | 72.7524 | 14.565 |

**נתונים- סטטיסטיקה תיאורית:**

בקוד לא יצרנו שם משתנה לווקטור הציונים הממוצעים של כל קבוצה, אלא כתבנו כל פעם: data$TestAVG [data$race.ethnicity == "group x"]. בטבלה כתבנו שמות משתנים (הציונים הממוצעים) של כל קבוצה בקיצור, כאשר עבור קבוצה x שם המשתנה יהיה:"grades- group x".  
**שאלת המחקר:**

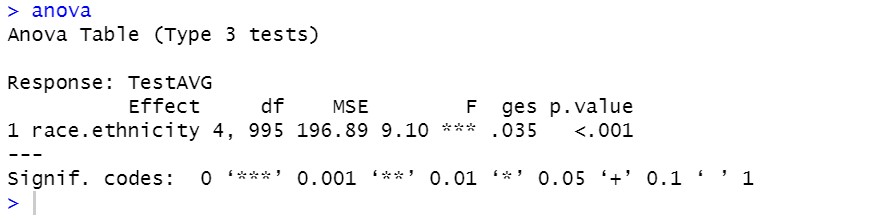
האם השתייכות לקבוצה אתנית מסוימת משפיעה על הציון הממוצע של ילד?   
שיערנו שברמת מובהקות של 5%, לפחות תוחלת ציוני ממוצעים של קבוצה אתנית אחת שונה מהתוחלות של יתר הקבוצות האתניות באוכלוסיה.   
**הסבר לגבי הניתוח:**

על מנת לבדוק האם תוחלת אחת לפחות שונה מהיתר, נשתמש בניתוח שונות חד- גורמי (כך נתגבר על בעיית תיסוף האלפא שנגרמת עקב עריכת סדרת מבחני t להפרש תוחלות במדגמים ב"ת, שבכל אחד מהם H0 נכונה). הניתוח נעשה ע"י מבחן F להשוואת תוחלות ב"ת (ניתוח שונות פשוט/חד-גורמי-ANOVA)הבודק בבת אחת אם התוחלות של האוכלוסיות מהן נלקחו המדגמים שוות או שונות. למבחן5 רמות: קבוצות A-E. השתמשנו בפקודה aov\_ez כדי לערוך את הניתוח. **הנחות המבחן:**

1. דגימה מקרית
2. המשתנה מתפלג נורמלית באוכלוסיות- השתמשנו ב3 פרוצדורות לבדיקת הנחת נורמליות: הסתכלות על היסטוגרמות, שימוש ב-QQ plot הראה כי הנקודות של חלק מהקבוצות מסתדרות על מעין קו ישר ושל חלק לא ממש, ובנוסף קיבלנו תוצאה מובהקת במבחן Shapiro-Wilk עבור קבוצות B ו-C. הממצאים מראים על סטיה מנורמליות של קבוצות B ו-C, לכן עבורן קבענו כי הנחת הנורמליות לא מתקיימת. בעוד שעבור יתר הקבוצות (A, D, E), הממצאים מראים שהנחת הנורמליות כן מתקיימת.
3. שיווין שונויות בשתי האוכלוסיות- ערכנו מבחן F להשוואת שונויות ע"י שימוש ב-Levene's test שנתן תוצאה לא מובהקת, כלומר- השונויות שוות.

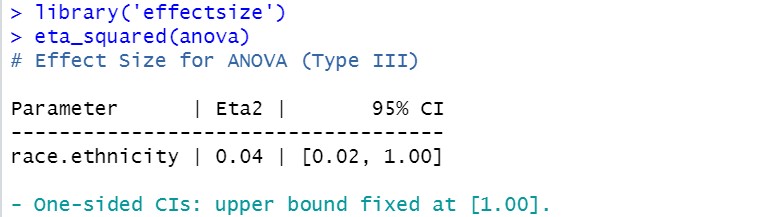
**תוצאות:**

F (4, 995) = 9.10 , p = p value < 0.001 , df= 4 , 995  
התקבלה תוצאה מובהקת (p < 0.001), לכן נדחה את השערת האפס.   
**מסקנה:**

נסיק שתוחלת של לפחות קבוצה אתנית אחת שונה מיתר התוחלות באופן מובהק.  
כלומר, יש אפקט- יש השפעה מובהקת של השתייכות לקבוצה אתנית מסוימת על הציון הממוצע (יש הבדל בין תוחלות התנאים).  
**הפלט ב-R:**

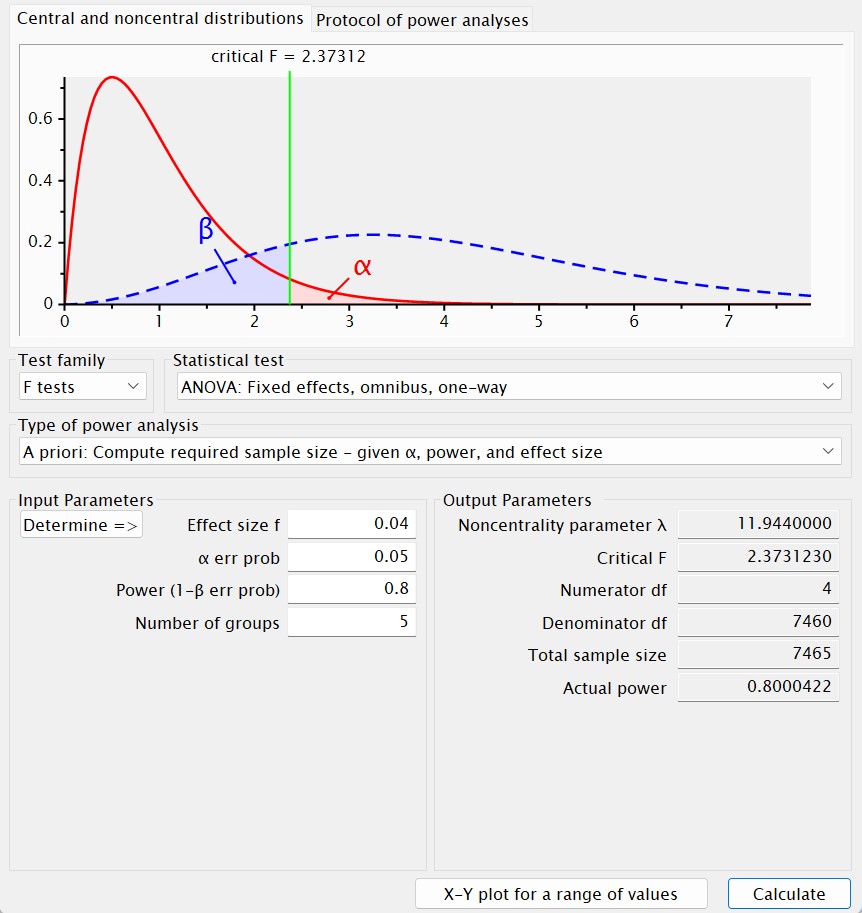
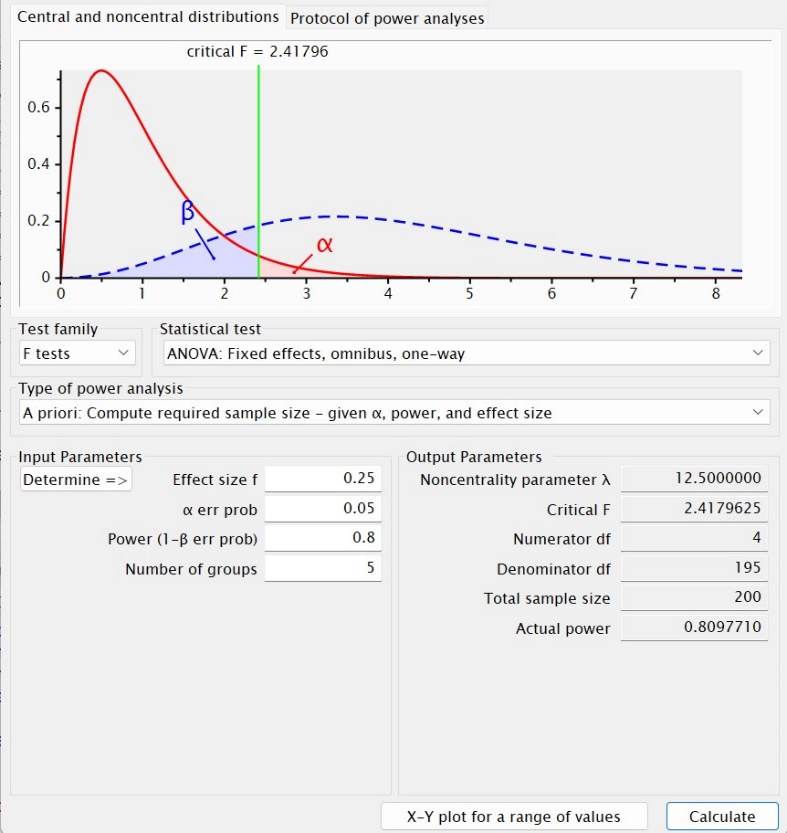
**מענה על סעיף 7:**

מדד גודל אפקט η2 – אמנם בפלט שקיבלנו ב-R קיים מדד גודל האפקט (ges= 0.035), אך בכל זאת בחרנו לערוך חישוב נפרד הכולל גם את רווח הסמך למדד גודל האפקט. ניתן לראות כי יש חלק של השונות הכללית המוסבר על ידי ההבדלים בין הקבוצות, לכן- יש אפקט.



**מענה על סעיף 5:**

בצילום המסך השמאלי, ניתן לראות כי גודל המדגם הדרוש בעבור עוצמה סטטיסטית של 80% במבחן דו-זנבי, עבור אפקט בגודל בינוני (שבגלל שערכנו מבחן F הוא שווה ל0.25 ולא ל0.5), ואלפא של 5%, הוא: 200 תצפיות. ובצילום המסך הימני, ניתן לראות שגודל המדגם הדרוש בעבור עוצמה סטטיסטית של 80% במבחן דו-זנבי, עם אלפא של 5%, ועבור גודל האפקט הקיים (כפי שחישבנו בסעיף 7), הוא: 7465 תצפיות.



**שאלה מספר 3- ניתוחי המשך**

**נתונים- סטטיסטיקה תיאורית:**

זהים לנתונים של שאלה מספר 2 (היות שאלו הם ניתוחי המשך).  
**שאלת המחקר:**

בהמשך לניתוח שערכנו בשאלה 2, רצינו לבדוק איזו קבוצה אתנית הביאה לקבלת התוצאה המובהקת, כלומר- התוחלת של איזו קבוצה אתנית (אחת/ יותר) שונה מיתר התוחלות?   
רצינו לעבוד בצורה שיטתית ולכן שיערנו שברמת מובהקות של 5%, תוחלת ציונים ממוצעים של קבוצה אתנית מסוימת תהיה שונה מזו של הקבוצות האתניות שבעלות ממוצע ציונים ממוצעים גבוה ממנה. ניסחנו 4 השערות (רמת ההצלחה = תוחלת הציונים הממוצעים):  
1. רמת ההצלחה של קבוצה A במבחן תהיה שונה מרמת ההצלחה של כל שאר הקבוצות בממוצע.  
2. רמת ההצלחה של קבוצה B במבחן תהיה שונה מרמת ההצלחה של קבוצות C, D, E בממוצע.  
3. רמת ההצלחה של קבוצה C במבחן תהיה שונה מרמת ההצלחה של קבוצות D, E בממוצע.  
4. רמת ההצלחה של קבוצה D במבחן תהיה שונה מרמת ההצלחה של קבוצה E במבחן.  
**הסבר לגבי הניתוח:**

על מנת לבדוק את ההשערות, נשתמש בניתוחי המשך, המתבצעים באמצעות השוואה (=קונטרסט) בין תנאים שונים. כל השוואה מיוצגת על ידי משקולות, שמתארות מהן הקבוצות שביניהן משווים (קבוצות שביניהן משווים מקבלות סימנים הפוכים, קבוצה שלא נכללת בהשוואה מקבלת את הסימן 0, סכום המשקולות של כל הקבוצות צריך להיות אפס). כלל ההשוואות שערכנו ב"ת זו בזו, כאשר התוצאה של אחת לא תלויה בתוצאה של אחרת (=השוואות אורתוגונליות, פירוט בסעיף ההנחות), היות שסדרה של J-1 השוואות אורתוגונליות תפרק את השונות Between באופן היעיל ביותר, שיגרום לניפוח מינימלי של אלפא. יצרנו מסגרת נתונים שהכנסנו לפקודה contrast על מנת לערוך את הניתוח.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| group E | group D | group C | group B | group A |  |
| -1 | -1 | -1 | -1 | 4 | comp1 |
| -1 | -1 | -1 | 3 | 0 | comp2 |
| -1 | -1 | 2 | 0 | 0 | comp3 |
| -1 | 1 | 0 | 0 | 0 | comp4 |

A\_vs\_all = comp1 , B\_vs\_CDE = comp2 , C\_vs\_DE = comp 3 , D\_vs\_E = comp4))

**הנחות המבחן:**

1. דגימה מקרית
2. המשתנה מתפלג נורמלית באוכלוסיות- בדיקה זהה לזו בשאלה 2.
3. שיווין שונויות בשתי האוכלוסיות- בדיקה זהה לזו בשאלה 2.
4. מדגמים בלתי תלויים- על ההשוואות להיות ב"ת (ממצות ומוציאות), לכן:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | group E | group D | group C | group B | group A |  |
| 0= | -1 | -1 | -1 | -1 | 4 | comp1 |
| 0= | -1 | -1 | -1 | 3 | 0 | comp2 |
| 0= | -1 | -1 | 2 | 0 | 0 | comp3 |
| 0= | -1 | 1 | 0 | 0 | 0 | comp4 |

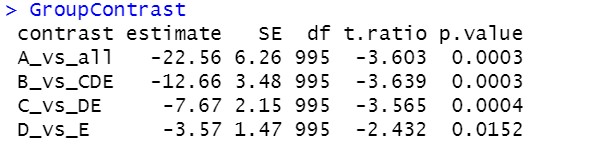
נבדוק שההשוואות הן גם ב"ת זו לזו (השוואות הן בלתי-תלויות כאשר המתאם בין המשקולות שלהן הוא אפס), לצורך הבדיקה, עלינו לסכום את מכפלת המשקולות:

∑comp1\*comp2= 4\*0 + (-1)\*3+(-1)\*(-1)+(-1)\*(-1)+(-1)\*(-1)= 0-3+1+1+1= 0  
∑comp1\*comp3= 4\*0 + (-1)\*0 + (-1)\*2 + (-1)\*(-1)+(-1)\*(-1)= 0+0-2+1+1= 0  
∑comp1\*comp4= 4\*0 + (-1)\*0 + (-1)\*0+ (-1)\*1 + (-1)\*(-1)= 0+0+0-1+1= 0  
∑comp2\*comp3= 0\*0 + 3\*0 + (-1)\*2+ (-1)\*(-1) + (-1)\*(-1)= 0+0-2+1+1= 0  
∑comp2\*comp4= 0\*0 + 3\*0 + (-1)\*0 + (-1)\*1 + (-1)\*(-1)= 0+0+0-1+1= 0  
∑comp3\*comp4= 0\*0 + 0\*0 + 2\*0 + (-1)\*1 + (-1)\*(-1)= 0+0+0-1+1= 0

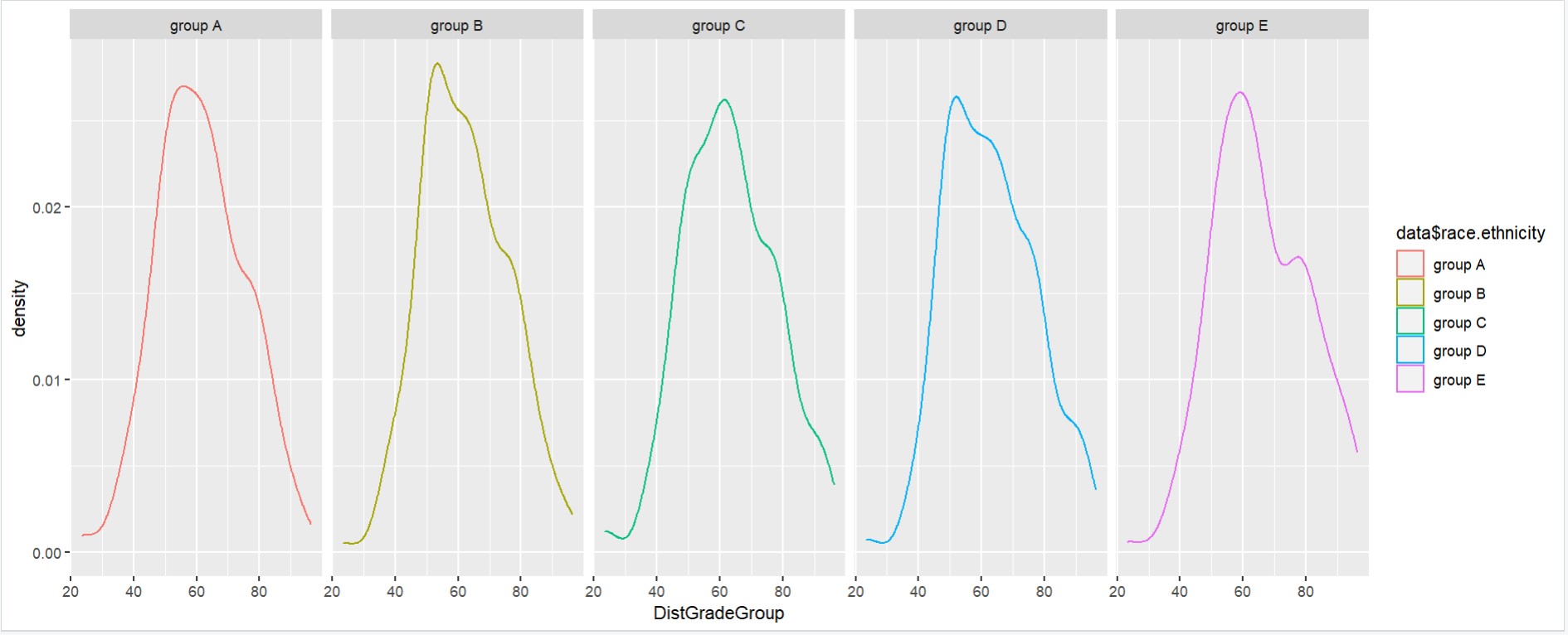
סכום כל אחת מהמכפלות בין משקולות הוא 0, לכן- הנחת ההשוואות האורתוגונליות מתקיימת.  
**תוצאות:**

Contrast A\_vs\_all: p = p value = 0.0003 , df= 995  
התקבלה תוצאה מובהקת (p = 0.0003), לכן נדחה את השערת האפס.   
Contrast B\_vs\_CDE: p = p value = 0.0003 , df= 995  
התקבלה תוצאה מובהקת (p = 0.0003), לכן נדחה את השערת האפס.  
Contrast C\_vs\_DE: p = p value = 0.0004 , df= 995  
התקבלה תוצאה מובהקת (p = 0.0004), לכן נדחה את השערת האפס.  
Contrast D\_vs\_E: p = p value = 0.0152 , df= 995  
התקבלה תוצאה מובהקת (p = 0.0152), לכן נדחה את השערת האפס.

**מסקנה:**

נסיק שרמת ההצלחה של קבוצה אתנית מסוימת שונה מרמת ההצלחה של כל הקבוצות האתניות שממוצע הציונים הממוצעים שלהן גדול משלה, באופן מובהק. כלומר, התוחלת שלה נלקחת מהתפלגות שונה, שאינה שייכת להתפלגות של יתר הקבוצות הנ"ל. מבחינת ניתוחי המשך, אין רק תוחלת של קבוצה אתנית אחת שהביאה לתוצאות, היות שכל התוצאות יצאו מובהקות.  
**הפלט ב-R:**

**מענה על סעיף 4:**

**** התרשים מתאר את ההתפלגות של הציונים הממוצעים של כל אחת מן הקבוצות זו ליד זו, כך שנוכל לראות את ההבדלים בהתפלגויות כפי שקיבלנו בתוצאות השאלה.