**עבודה ב 206836009 – 318862398**

השתמשנו בסט הנתונים עבור Social Media Influencers, משם התמקדנו בטבלה של ידוענים באינסטגרם. סט הנתונים הכיל את העמודות הבאות:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Influencer insta name** | category\_1 | category\_2 | Followers | Audience country(mostly) | Authentic engagement | Engagement avg |

ניקינו את מאגר הנתונים בעזרת סקריפטים נפרדים שכתבנו.

* החלפנו את הערכים המספריים לפורמט שניתן לעבוד איתו (לדוגמה השרשור 17K הומר ל 17000)
* הסרנו שורות בהן היו ערכים לא תקינים (שורות ללא קטגוריה, ללא כמות עוקבים)
* הוספנו עמודת Engagement Ratio המציגה עבור כל עמוד אינסטגרם את כמות הengagement האותנטי שלו חלקי סך כל הengagement שלו

הסקריפטים נמצאים בגיטהב של הפרויקט.

**שאלה ראשונה**

***שאלת המחקר***

שיערנו שהודים אוהבים עמודי אינסטגרם מהקטגוריה Cinema & Actors ביחס גדול יותר משאר האוכלוסייה. כלומר, שברמת מובהקות של 5%, עמודי אינסטגרם עם דמוגרפיה הודית יותר סבירים להיות עמודים מהקטגוריה Cinema & Actors.

***הסבר לגבי הניתוח***

מכיוון שאנו בוחנים פרופורציה בין מדגם של ערוצים עם דמוגרפיה הודית לכלל הערוצים, נשתמש במבחן רווח סמך לפרופורציה.

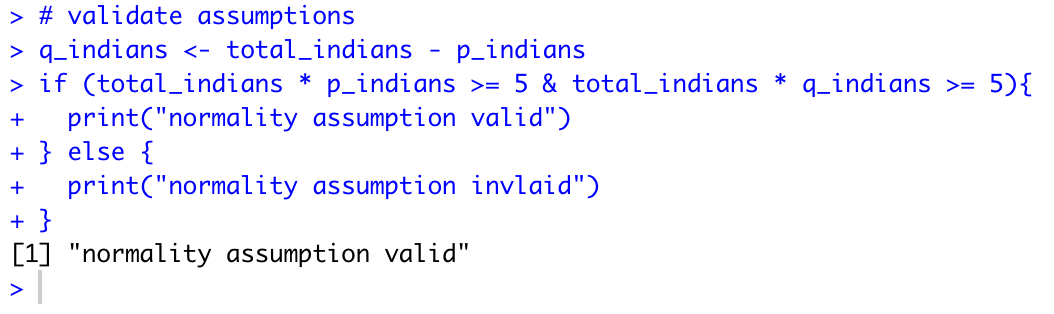
נסמן את פרופורציית הערוצים מקטגוריית Cinema & Actors מתוך כלל הערוצים עם דמוגרפיה הודית  כP.

נסמן את פרופורציית הערוצים מקטגוריית Cinema & Actors מתוך כלל הערוצים כ Ṕ.

השערות:

H0: P = Ṕ

H1: P > Ṕ

הסטטיסטי הוא  Ṕ.

***בדיקת ההנחות***

* דגימה מקרית
* התפלגות הדגימה נורמלית בקירוב

ניתן להניח נורמליות מכיוון ש n\*p > 5 וגם n\*q > 5

***תוצאות***

תיאור התוצאות ברמה הסטטיסטית צריך לכלול לכל הפחות את סטטיסטי המבחן, ד"ח (במידה ויש), p-value. לדוגמה, במבחן t לרוב נראה את הדיווח הסטטיסטי הבא:

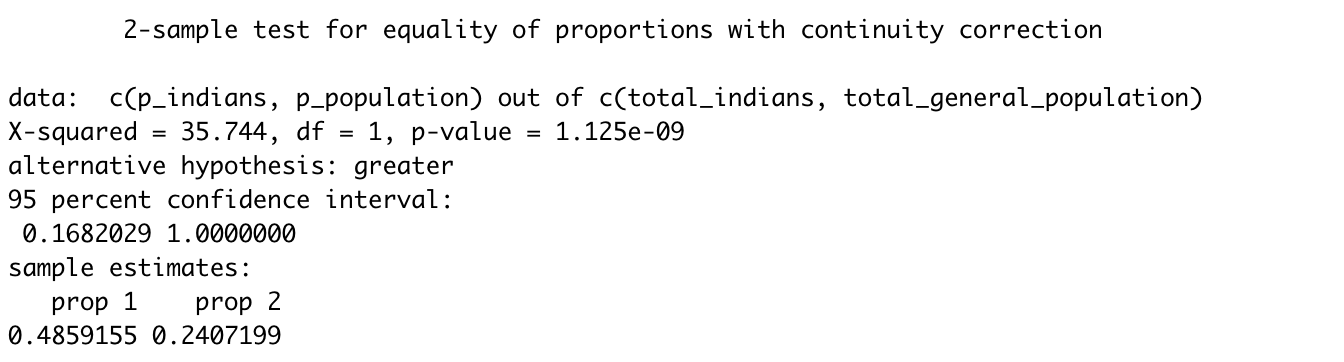
t(df) = t value, p = p value - המסקנה הסופית לגבי השאלה ששאלתם.

***תוצאות***

df = 1    p-value = 1.125e-09    α = 0.05    N = 142

ערך הסטטיסטי שהתקבל קטן מערך האלפא, לכן התוצאה מובהקת ונדחה את השערת האפס.

***מסקנה***נסיק שברמת בטחון של 95%, עמודים עם דמוגרפיה הודית נוטים להיות בנושא Cinema & Actors באופן יחסי יותר מאשר מעמודים עם דמוגרפיה כללית (ללא התייחסות לדמוגרפיה).

***פלט ב-R* **

**שאלה 2**

***שאלת המחקר***

שיערנו שיש הבדל בין פעילות העוקבים הממוצעת בחשבונות שהדמוגרפיה העיקרית שלהם היא מהמדינות הגדולות (ארה״ב, אינדונזיה, הודו, ברזיל), ברמת מובהקות של 5%.

***הסבר לגבי הניתוח***

מכיוון שאנו בוחנים הבדל תוחלות בין מספר קבוצות ואיננו רוצים לנפח את ערך האלפא (או להקטין אותו), מבחן שונות חד גורמי למספר משתנים הוא הניתוח המתאים ביותר

***בדיקת ההנחות***

* דגימה מקרית
* התפלגות נורמלית (ההנחה לא התקבלה - קיבלנו p=2.2e-16 < 0.05 במבחן שפירו-ווילקס)
* שוויון שונויות (ההנחה לא התקבלה - קיבלנו p=2.2e-16 < 0.05 במבחן ברטלט)
* משתנים ב״ת (התקבלה ההנחה - קיבלנו p=0.47 > 0.05 במבחן חי בריבוע לאי-תלות)

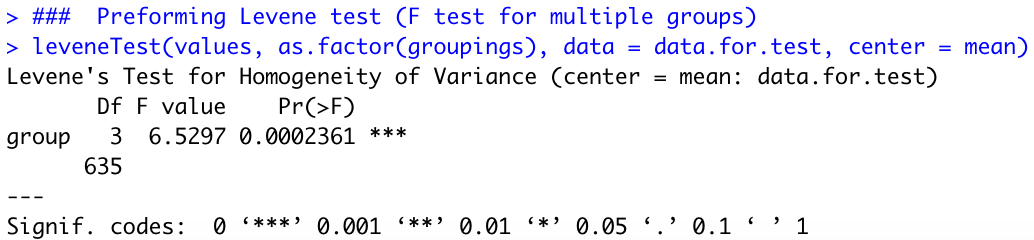
***תוצאות***

p-value = 0.00024    df = (3,635)    α = 0.05    N = 639

ערך הסטטיסטי שהתקבל קטן מערך האלפא, לכן התוצאה מובהקת ונדחה את השערת האפס.

***מסקנה***נסיק כי לפחות אחד מממוצעי הפעילות בחשבונות להם רוב עוקבים מברזיל, ארה״ב, אינדונזיה והודו שונה מהשאר.

***פלט ב-R***

****

**שאלה 3**

***שאלת המחקר***

אנו משערים כי לכמות העוקבים אחר חשבונות מוזיקליים יש השפעה על היחס בין הפעילות הלא אותנטית בחשבון לפעילות האותנטית. ברמת מובהקות של 5%.

***הסבר לגבי הניתוח***

מהנתונים נחלץ סטיות תקן וממוצעי מדגמים של חשבונות גדולים וקטנים מהממוצע, הניתוח המתאים ביותר יהיה ביצוע רווח סמך להפרש תוחלות בעזרת מבחן t למדגמים בלתי תלויים.

***בדיקת ההנחות***

* דגימה מקרית
* מדגמים ב״ת (ההנחה התקבלה - קיבלנו p=0.35 > 0.05 במבחן חי בריבוע לאי תלות)
* שוויון שונויות (ההנחה לא התקבלה - קיבלנו p=0.00056 < 0.05 במבחן ברטלט)
* התפלגות נורמלית (ההנחה לא התקבלה - קיבלנו p=0.006 < .05 במבחן שפירו-ווילקס)

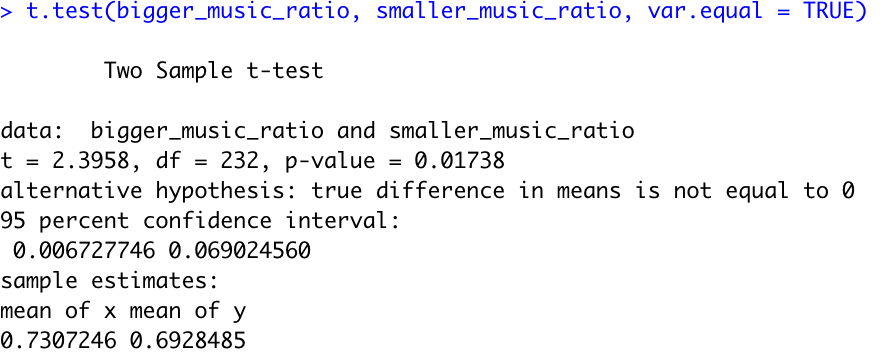
***תוצאות***   
tc = ∓1.65    𝝰=0.05    df=232        t=2.396    p=0.0174    Ŝ(x̄1-x̄2)=0.048

התקבלה תוצאה מובהקת (p<𝝰) ולכן נדחה את השערת האפס

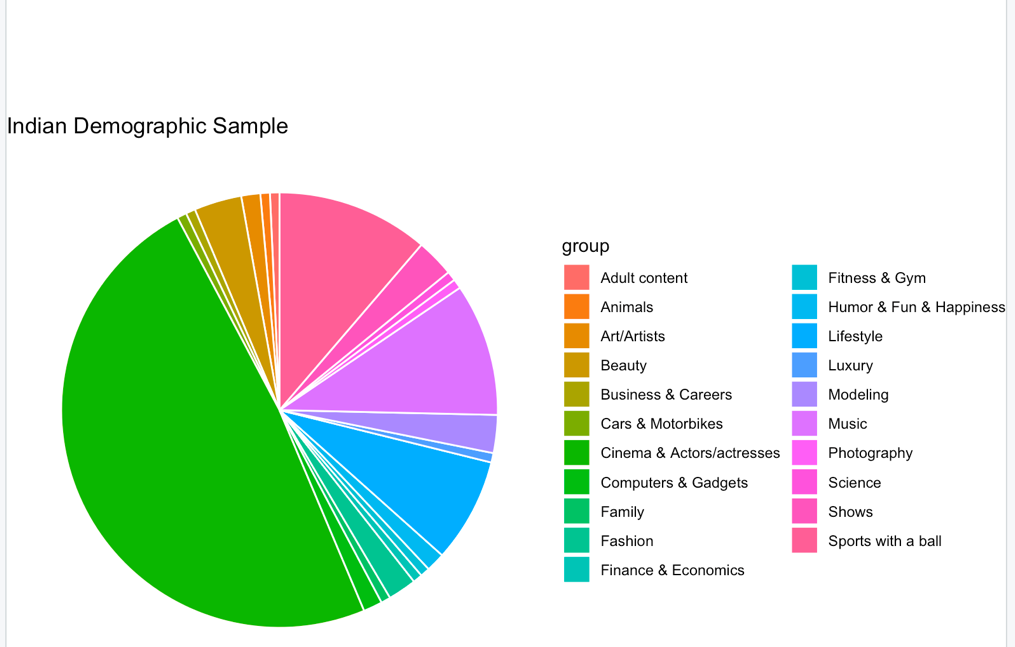
***מסקנה***

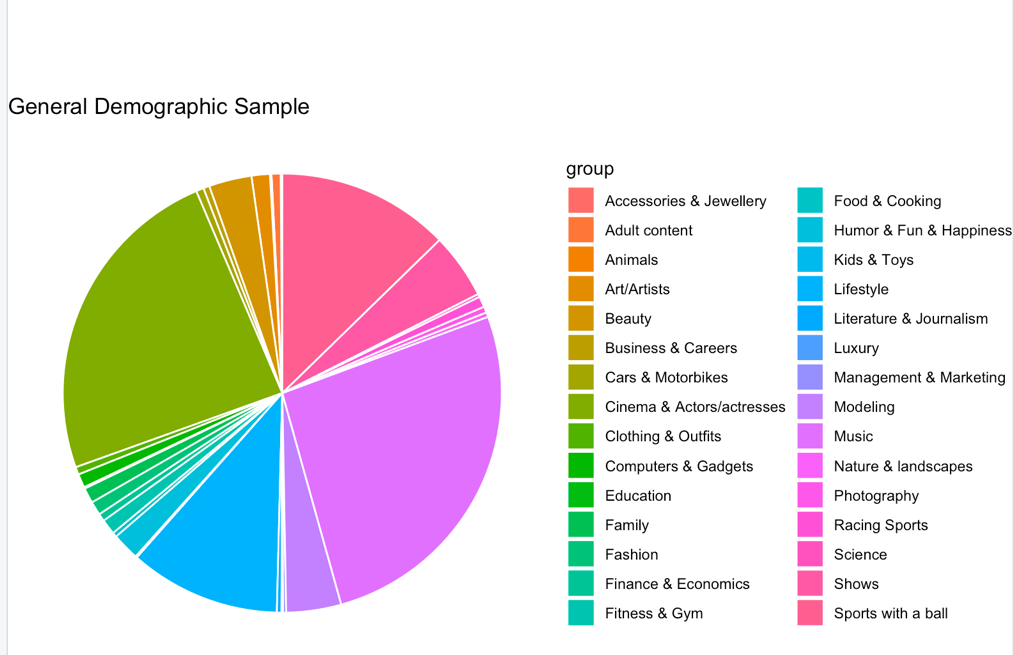
נסיק כי ברמת ביטחון של 95%, יש הבדל ביחס בין הפעילות הלא אותנטית לאותנטית בעמודים מוסיקליים גדולים (בעלי הרבה עוקבים) למול יחס הפעילות בעמודים קטנים יותר.

***פלט ב-R***

******

***סעיף 4 לשאלה 1***

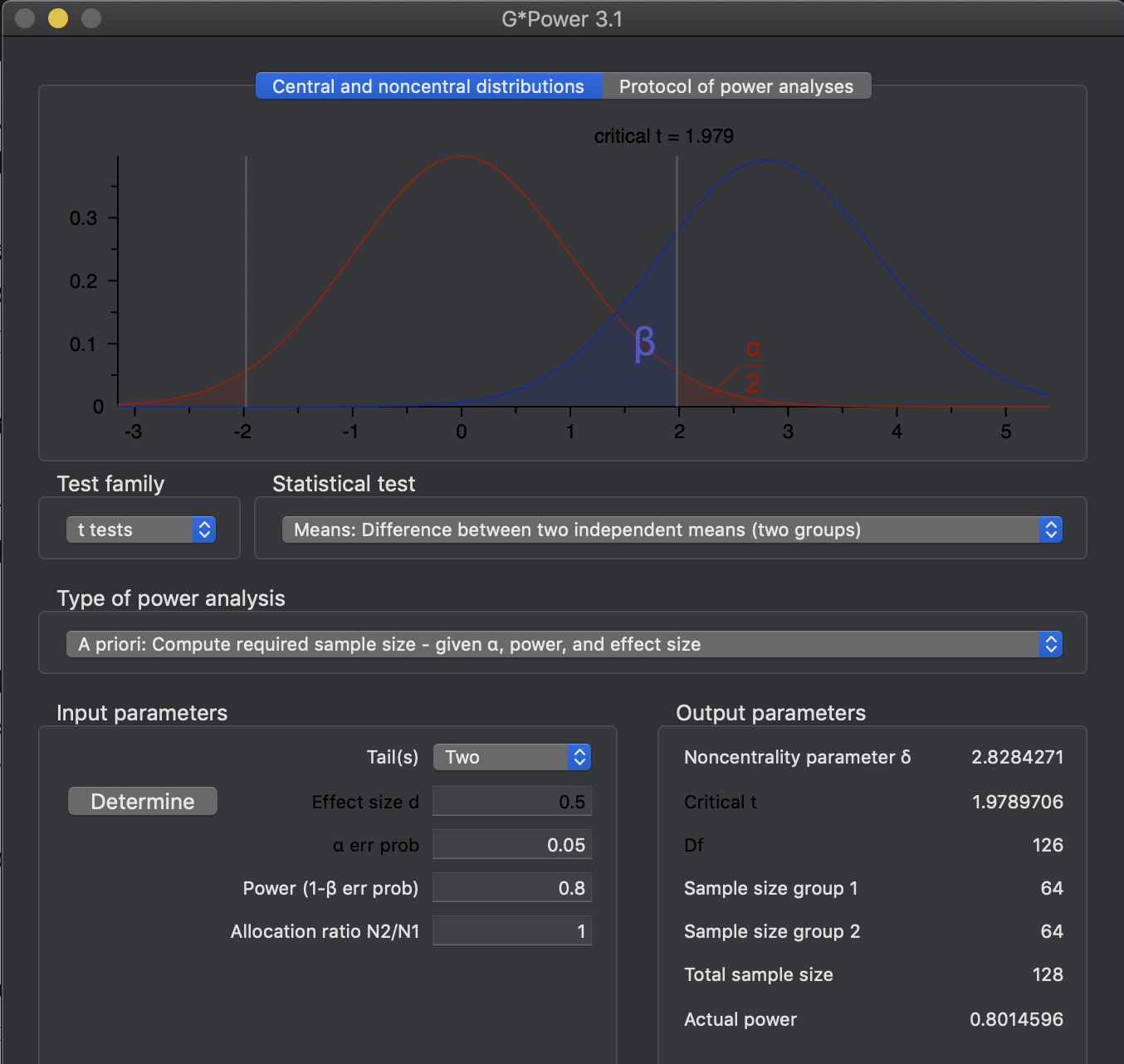
******

******

הצגנו את הנתונים משאלה 1 בגרף פיי.

כפי שניתן לראות ובהתאם למסקנה שקיבלנו בשאלה 1, ניתן לראות בקלות שבקרב העמודים הפופולרים בהודו, יותר עמודים עוסקים בקלונוע ושחקנים באופן יחסי מאשר בקרב האוכלוסיה הכללית.

***סעיף 5 - G\*Power לשאלה 3***

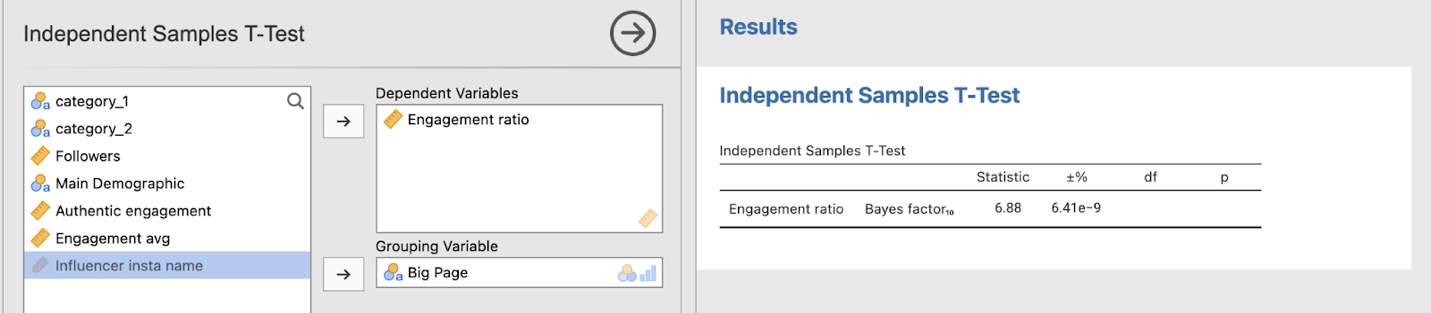
חישבנו עוצמה סטטיסטית עבור שאלה שלוש בעזרת G\*Power, ומצאנו כי גודל המדגם הדרוש - בהינתן הנתונים בהוראות השאלה - הינו 128 (כמו שניתן לראות בתמונה)  


***סעיף 6  - bayes factor לשאלה 3***

יצרנו קובץ נתונים חדש מבוסס על האחד המקורי בו הצגנו רק עמודים בנושא מוזיקה.

בנוסף, הוספנו עמודה נוספת בשם "Big Page" שערכיה עשויים להיות "big"/"small". בדקנו Bayes Factor עם המשתנה התלוי Engagement Ratio לפי עמוד גודל או קטן וקיבלנו 6.41e-9. זהו מספר די קטן שמעיד שבסבירות די גבוהה עלינו לדחות את השערת ה0. (כלומר, יש הבדל משמעותי בין תוחלת Engagement Ratio של עמודים גדולים לעומת קטנים)

מצורף צילום מסך של Jamovi:

******

***סעיף 7 - גודל אפקט לשאלה 3***

על מנת לחשב גודל אפקט לשאלה זו נשתמש במדד d של כהן (Cohen's D)

אמדנו ב-R את טעות התקן, והשתמשנו בממוצעי המדגמים כאומדים קונסיסטנטים וחסרי הטיה ל- H0 ו-H1  
חישבנו ב-R את גודל האפקט וקיבלנו כי d = 0.797 (אפקט גדול)

