סטטיסטיקה הסקית

ראשית, נזכיר כי המשתנה המסביר שלנו הוא מספר הפלייליסטים שבהם מופיע שיר בספוטיפיי (נמדד במאות), ואילו המשתנה המוסבר הוא סך כל ההשמעות של אותם שירים (נמדד במיליארדים). במדגם שלנו יש 4467 תצפיות. מכאן והאלה לא נזכיר כל פעם יחידות מדדה אלא נקרא משתנים: כמות פלייליסטים, כמות שמעות.

שאלה 1

א. לפי הנחה שכמות שמעות מתפלגות נורמלית אמדנו תוחלת בשיטת המומנטים, היא שווה לממוצע המדגם. קיבלנו כי אומד לתוחלת שווה ל 443.839
נזכיר שהאומד לשונות הוא ממוצע ריבועי התצפיות מינוס הממוצע בריבוע. אחרי שימוש בנוסחה התקבל ערך

281500.2 = mean((Streams)^2) – (mean(Streams))^2

ב. אחרי חישוב של ערך המינימלי של משתנה כמות הפלייליסטים, חיסרנו אותו מכלל התצפיות והנחנו שמשתנה חדש W שייך למשפחת התפלגות גמא. בחרנו בשיטת המומנטים לאמידת פרמטרים

$$\hat{\alpha}_{mom} = \frac{(\overline{x}_n)^2}{\hat{x}_n^2 - (\overline{x}_n)^2} = 0.7137 \qquad \hat{\beta}_{mom} = \frac{\overline{x}_n}{\hat{x}_n^2 - (\overline{x}_n)^2} = 0.0012$$

ג. חישבנו אחוזונים הנדרשים על פי התפלגויות לפי פרמטרים שחושבו, להלן טבלה עם תוצאות.

משתנה	האחוזון ה-10	האחוזון ה-50	האחוזון ה-75	האחוזון ה-90
Y (כמות השמעות)	-236.10	443.83	801.7	1123.78
W	30.02	354.49	828.07	1504.64

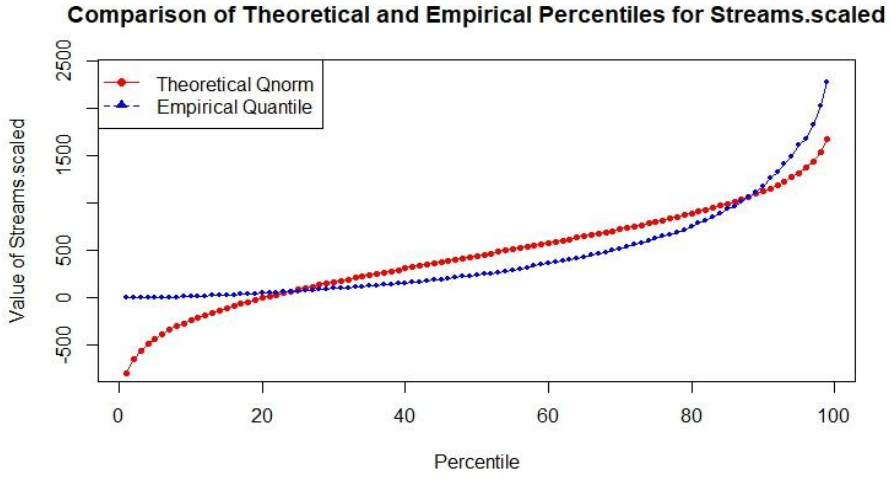
ד. נציג טבלה של אחוזונים אימפרים לעומת אחוזונים שהתקבלו מהתפלגויות שחישבנו

משתנה	האחוזון ה-10	האחוזון ה-50	האחוזון ה-75	האחוזון ה-90
Y (כמות השמעות)	14.2	239.62	623.9	1178.1
W	9.46	331.50	870.94	1639.7

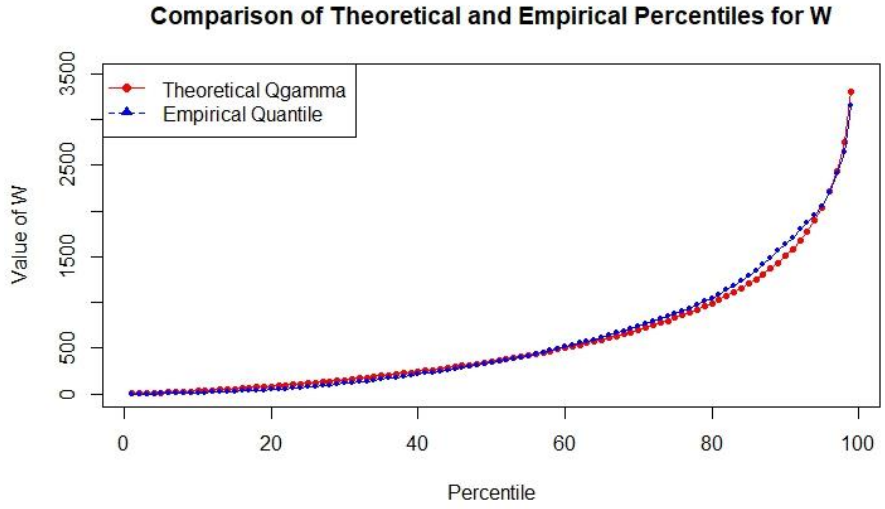
לאחר השוואה בין הטבלאות של האחוזונים שהתקבלו מהמדגם לבין אלו שנחזו על ידי המודל עבור כמות ההשמעות (משתנה Y) ,ניתן להבחין בפערים בין שתי השיטות. עם זאת, באחוזונים 75 ו-90, הערכים נעשו קרובים יותר, מה שעשוי להעיד על התאמה טובה יותר של המודל הנורמלי בטווחים אלו.

משתנה W, המייצג את מספר הפלייליסטים בהם מופיע שיר לאחר החסרת הערך המינימלי, נמצא כי האחוזונים שהתקבלו מהמדגם ומהמודל דומים יותר זה לזה, דבר המעיד על התאמה טובה יותר של המודל גמא לנתונים אלו.

ה. אחרי שחישבנו אחזונים אימפירים ואחוזונים בעזרת מודלים, התקבלו גרפים הבאים



הגרף מראה שהמודל המבוסס על התפלגות נורמלית לא מייצג באופן מדויק את האחוזונים האמיתיים של מספר ההשמעות, במיוחד בקצוות ההתפלגות. עם זאת, סביב האחוזון ה-90, הערכים קרובים יחסית זה לזה, כפי שציינו בסעיף הקודם.



במשתנה W המייצג את מספר הפלייליסטים בהם מופיע שיר, לאחר החסרת הערך המינימלי,
ניתן לראות מהגרף כי באמצעות המודל המבוסס על התפלגות גמא התקבלו ערכי אחוזונים הקרובים מאוד לערכים שנמצאו בדרך אמפירית. הדבר מעיד על כך שהמודל שנבחר מתאים לנתונים שלנו.

<u>שאלה 2</u>

רווח הסמך לתוחלת של משתנה Y המתאר את כמות השמעות, ברמת סמך של 97% התקבל בטווח (בהנחה ש Y מתפלג נורמלי): [426.61 ,461.07].

CI_low = mean(Streams)-qnorm(0.985)*sqrt(var/4467)=426.61

CI_up = mean(Streams)+qnorm(0.985)*sqrt(var/4467)=461.07

במקום התפלגות t, השתמשנו בהתפלגות נורמלית, כיוון שבמדגם שלנו יש 4467 תצפיות, מה שמתאים להתפלגות t עם 4466 דרגות חופש – וכאשר מספר דרגות החופש גדול, התפלגות t שואפת להתפלגות נורמלית.

רווח הסמך לשונות ברמת סמך של 92% נמצא בטווח: [271363.6 ,292363.3]. לחישוב רווח הסמך השתמשנו בהתפלגות חי-בריבוע, על פי החישוב הבא.

CI_var_Low = 4466*var/qchisq(0.96,4466) = 271363.6

CI_var_Up = 4466*var/qchisq(0.04, 4466) = 292236.3

כאשר לחישוב var - אומד לשונות של משתנה Y (כמות השמעות) נעשה שימוש באומד חסר הטיה לשונות

<u>שאלה 3</u>

א. נרצה לבחון האם קיים הבדל בתוחלת מספר ההשמעות בספוטיפיי בין שירים הנמצאים בכמות פלייליסטים קטנה מהחציון לבין שירים הנמצאים בכמות פלייליסטים גדולה מהחציון. חציון של כמות הפלייליסטים שווה ל – 331.51
<u>השערת האפס</u> (H_0): תוחלת מספר ההשמעות בספוטיפיי זהה עבור שתי הקבוצות.
<u>השערה אלטרנטיבית</u> (H_1): תוחלת מספר ההשמעות בספוטיפיי גבוה יותר עבור השירים הנמצאים בכמות פלייליסטים גדולה מהחציון.

ב.

$$H_0: E(Y_i|X_i < 331.51) = E(Y_i|X_i \geq 331.51)$$
$$H_1: E(Y_i|X_i < 331.51) > E(Y_i|X_i \geq 331.51)$$

נזכיר כי סטטיסטי המבחן להפרש בין תוחלות כאשר השונויות אינן ידועות ואינן שוות מוגדר באופן הבא:

$$T = \frac{\bar{y}_m - \bar{y}_n - (\mu_{y_m} - \mu_{y_n})}{\sqrt{\frac{S_{y_m}{}^2}{m} + \frac{S_{y_n}{}^2}{n}}} = \frac{\bar{y}_m - \bar{y}_n}{\sqrt{\frac{S_{y_m}{}^2}{m} + \frac{S_{y_n}{}^2}{n}}} \sim N_{H_0}(0,1)$$

ג. כאשר כאן Y_m מייצג את מספר ההשמעות בספוטיפיי עבור שירים הנמצאים בכמות פלייליסטים גדולה מהחציון, ו-Y_n מייצג את מספר ההשמעות בספוטיפיי עבור שירים הנמצאים בכמות פלייליסטים קטנה מהחציון. נזכיר כי הסטטיסטי מתפלג בקירוב נורמלי סטנדרטי תחת השערת האפס, בהתאם למשפט הגבול המרכזי (בכל קבוצה יש מספר רב של תצפיות: 2233 ו-2234). מכיוון שהשערה האלטרנטיבית מניחה כי תוחלת מספר ההשמעות גבוה יותר עבור שירים הנמצאים בפלייליסטים רבים יותר, נבצע את המבחן על ידי השוואת p-value לערך הסף שנקבע מראש לרמת המובהקות. אם p-value קטן מרמת המובהקות, נדחה את השערת האפס, אחרת לא נדחה אותה.

בחישוב סטטיסטי T בנוסחה המצוינת למעלה התקבל ערך: T = 43.66.
אנו מחשבים p-value בצורה הבאה: p-value = 1 − Φ(T) ≅ 0
המספר שהתקבל זעיר עד כדי כך שהמחשב מציג אותו כ-0, וזה הגיוני משום שערך T שחישבנו יצא גדול מאוד.

ד. מכיוון שערך p-value שהתקבל קטן מ-0.03, נדחה את H_0 השערת האפס ברמת מובהקות של 3% ונכריע את H_1. כלומר, אנו מסיקים כי ישנו הבדל מובהק בין הקבוצות, מספר ההשמעות גבוה יותר בממוצע עבור שירים שנמצאים בכמות פלייליסטים גדולה מהחציון, בהשוואה לשירים הנמצאים בכמות פלייליסטים קטנה מהחציון. תוצאה זו מעידה על קשר חיובי בין מספר הפלייליסטים בהם מופיע שיר לבין מספר ההשמעות שלו: שירים שמופיעים ביותר פלייליסטים זוכים ליותר השמעות.