# Лабораторна робота №2

## “Статистичне виведення”

Підготували: Руслан Дюбакін, Петриченко Нікіта, Пріхно Ілля, Резниченко Єлизавета.

Група: КМ-01

# Вступ

У рамках цієї роботи були спроби відповісти на такі питання:

* Чи можна справді вважати обрані в першій роботі 5 жанрів музики найпопулярнішими за середнім та медіанним значенням?
* Чи можна справді вважати обрані в першій роботі 5 артистів найпопулярнішими за середнім та медіанним значенням?
* Які треки є популярнішими: з ненормативною лексикою чи без неї?
* Які треки мають більший рівень мовності: з ненормативною лексикою чи без неї?
* Чи є статистично значущою виявлена достатньо висока кореляція між змінною energy та loudness?

Для дослідження вище наведених питань були побудовані довірчі інтервали для статистик, що становлять для нас інтерес (вибіркове середнє), були побудовані бутстреп довірчі інтервали для статистик, дисперсія яких не має асимптотично нормального розподілу (медіана, кореляція Пірсона). Використано тести Волда для тестування гіпотез.

# Звіт

[**Лабораторна робота №2 1**](#_zdokrnwpxg11)

[“Статистичне виведення” 1](#_wb3592r5eqjq)

[**Вступ 1**](#_nigdwc97siog)

[**Звіт 2**](#_fe3ltv5do0ky)

[**Результати досліджень 3**](#_vpla2jfgu3fb)

[5 жанрів музики 3](#_42yjzgbrkchx)

[5 артистів 11](#_pcppia5pn3ja)

[Ненормативна лексика і популярність 19](#_dtdvqb79m4f1)

[Ненормативна лексика і мовність 26](#_j8y6d8pay6hs)

[Кореляція energy та loudness 33](#_i4vv5sdgcty7)

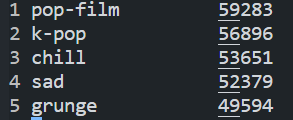
[**Висновки 36**](#_2dvcxhxnxrk7)

[**Список літератури 38**](#_ahcfu55s60cx)

# Результати досліджень

## 5 жанрів музики

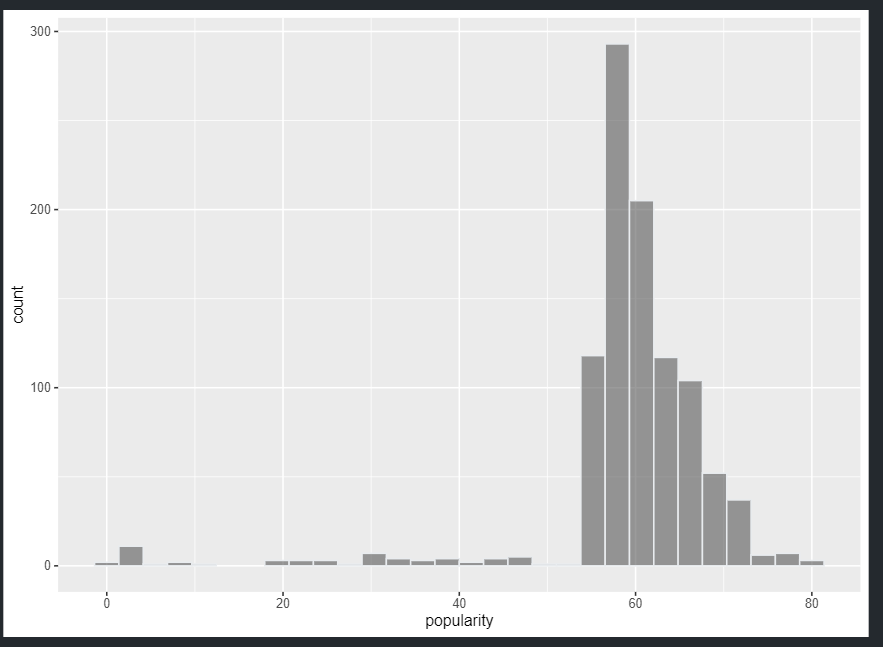
У першій лабораторній роботі було визначено 5 жанрів музики, що мають в сумі найвищу популярність.

Було отримано такі жанри:

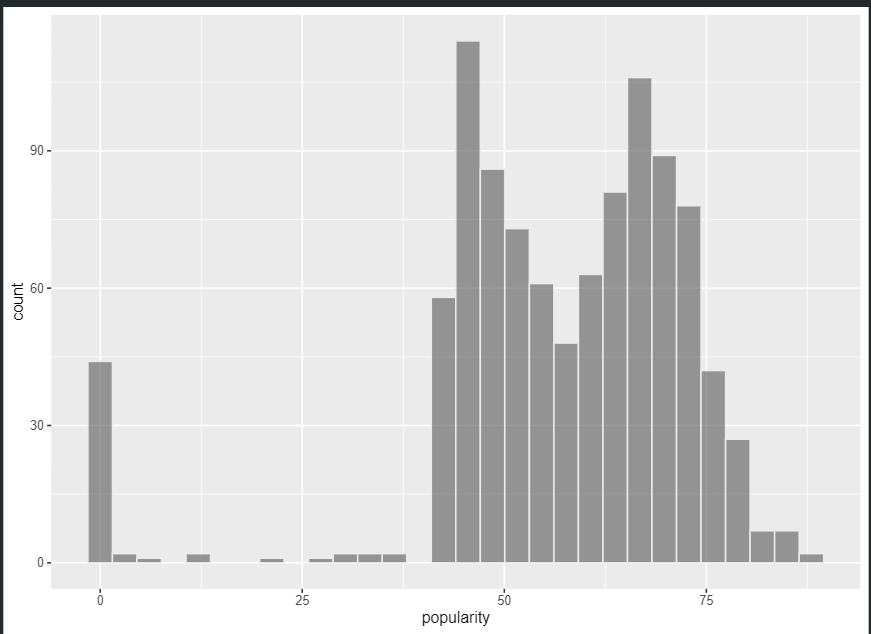
У нас виникло питання, чи справді можна вважати ці жанри найпопулярнішими.

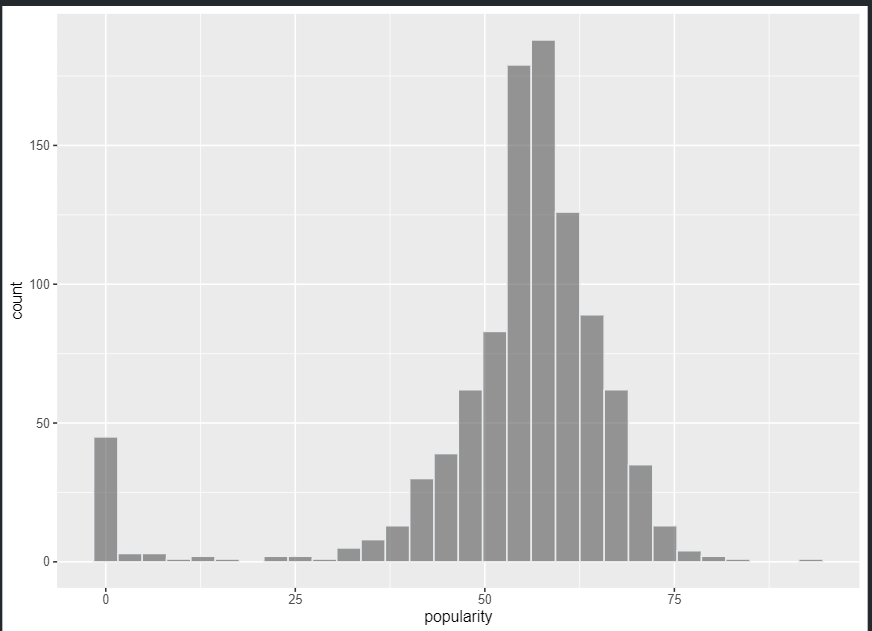
Подивимося на розподіли популярності кожного жанру:

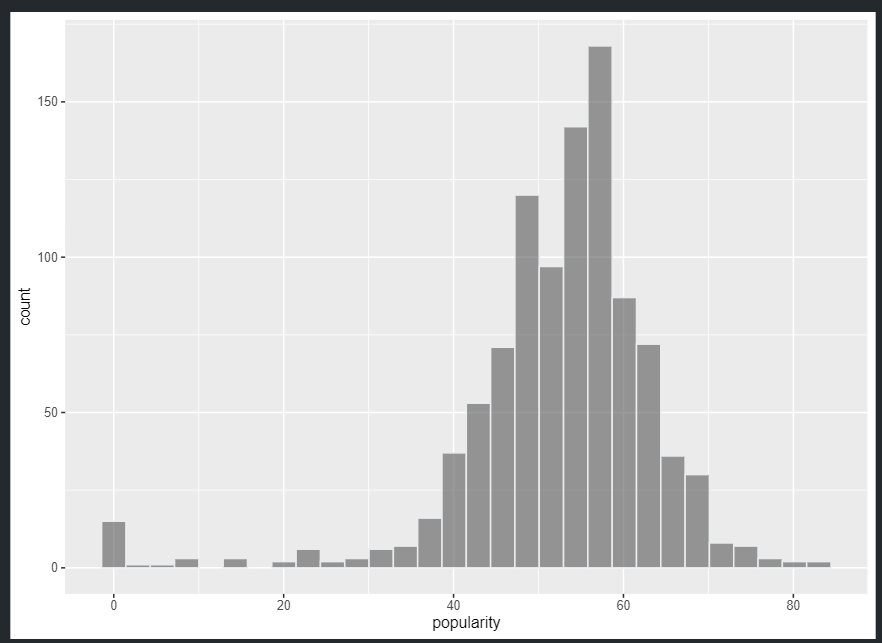
1. Розподіл популярності pop-film

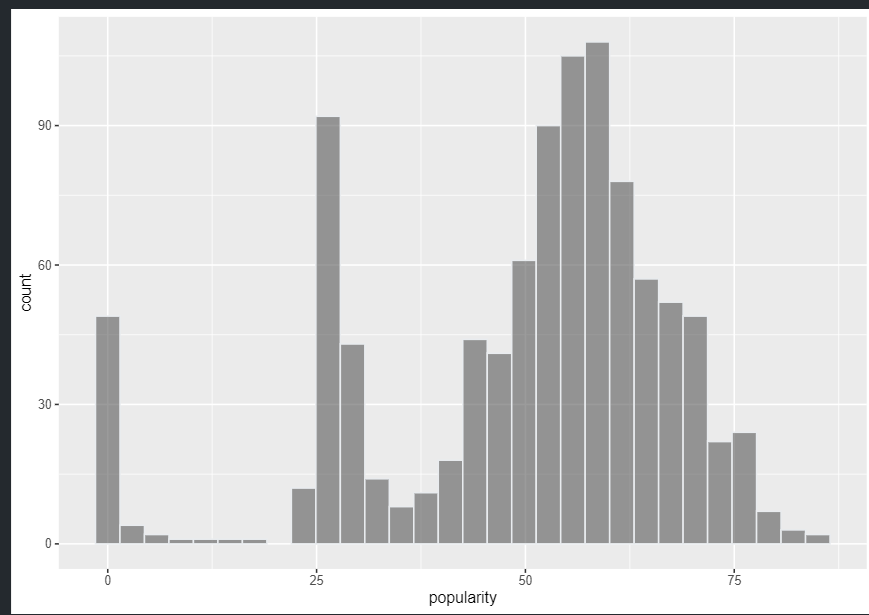


1. Розподіл популярності k\_pop\_genre:



1. Розподіл популярності для chill\_genre:
2. Розподіл популярності для sad\_genre:

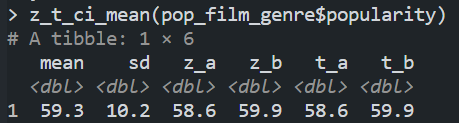


1. Розподіл популярності для grunge\_genre:

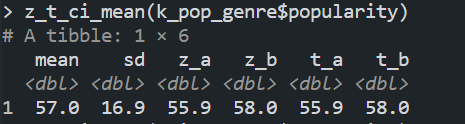
Як бачимо, кожен жанр має скошений асимптотично нормальний розподіл. Отже, для оцінки популярності жанрів доречніше орієнтуватися на медіанне значення, адже медіана є стійкішою за вибіркове середнє значення.

Оцінимо середнє значення популярності кожного жанру та побудуємо довірчі інтервали на основі квантилів стандартного нормального розподілу (z-розподілу) та t-розподілу.

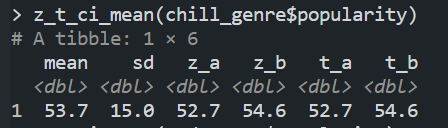
1. Довірчі інтервали для pop-film



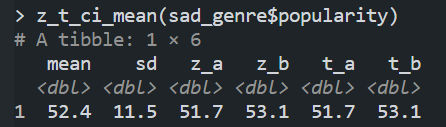
1. Довірчі інтервали для k-pop



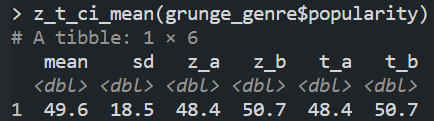
1. Довірчі інтервали для chill genre



1. Довірчі інтервали для sad genre

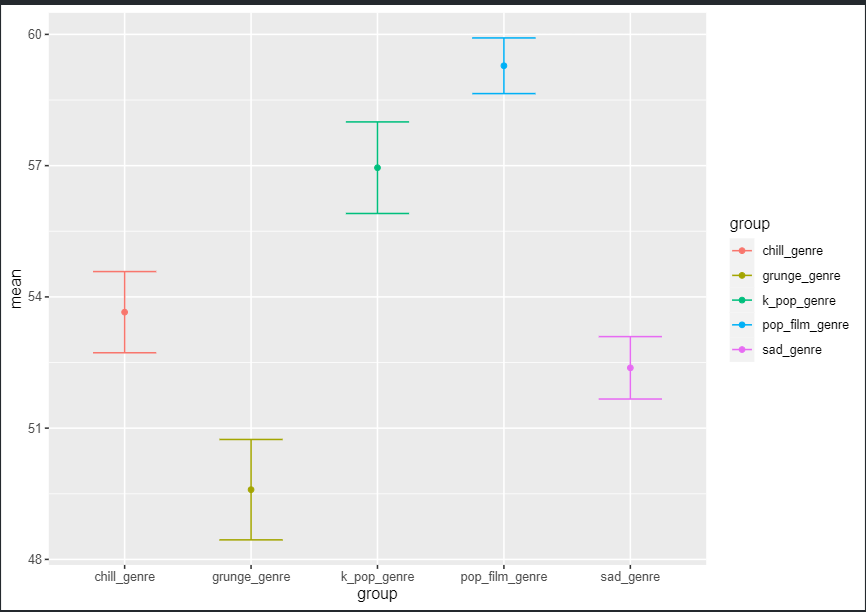


1. Довірчі інтервали для grunge genre



Бачимо, що довірчі інтервали з квантилями z-розподілу та t-розподілу майже співпадають.

Зобразимо на спільному графіку обраховані довірчі інтервали для популярності кожного жанру:

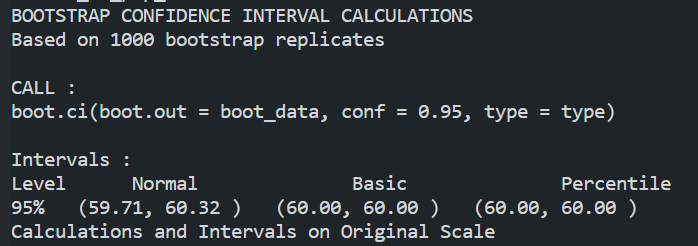


На графіку бачимо, що довірчі інтервали є доволі вузькими для кожного жанру. За середніми значеннями жанр pop-film справді можна вважати найпопулярнішим із обраних жанрів. Довірчі інтервали жанрів chill та sad перетинаються, отже, не можна вважати точно, що chill є популярнішим за sad.

Оцінимо медіанне значення популярності кожного жанру та побудуємо довірчі інтервали на основі бутстрепу, так як важко оцінити асимптотичну дисперсію медіани. Для бутстрепу оберемо кількість реплікацій - 1000. Для кожного жанру обраховані довірчі інтервали типу Normal, Basic, Percentile.

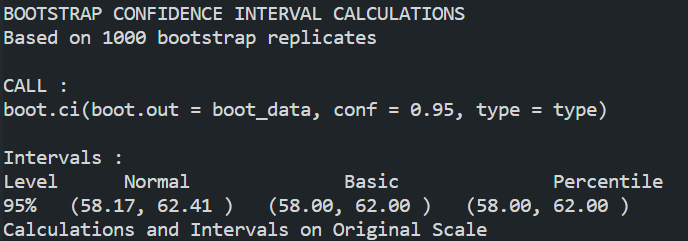
1. Довірчі інтервали для жанру pop-film





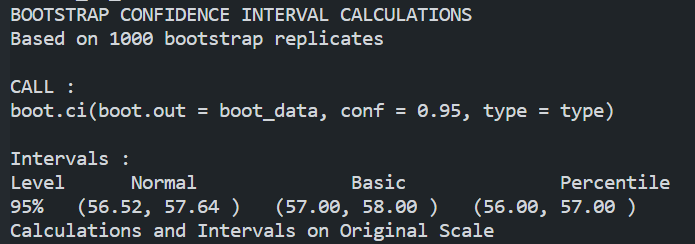
1. Довірчі інтервали для жанру k-pop:





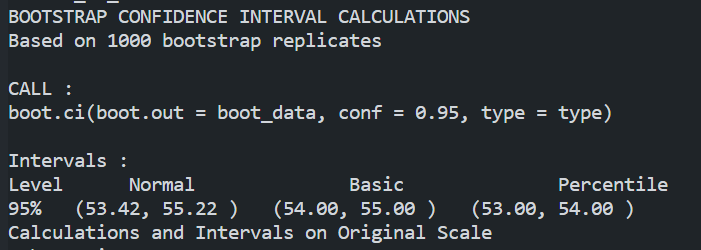
1. Довірчі інтервали для жанру chill:





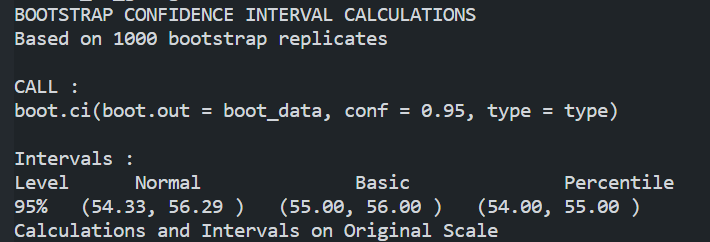
1. Довірчі інтервали для жанру sad:





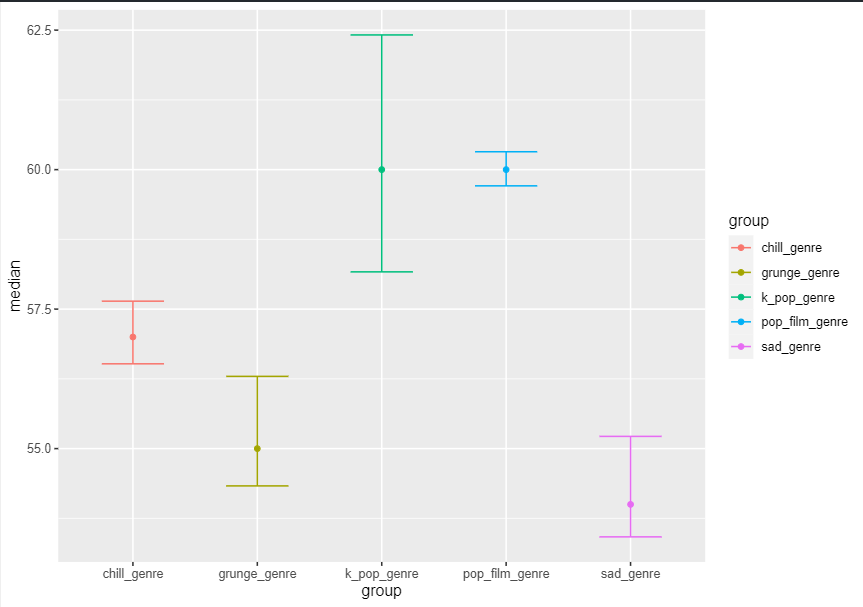
1. Довірчі інтервали для жанру grunge:





Видно, що довірчі інтервалів різних типів є скошеними для медіанних значень. Це пояснюється тим, що розподіли є скошеними в ліву або праву сторону.

Побудуємо спільний графік для кожного жанру та їхні довірчі інтервали типу Normal.



Як бачимо, за медіанними значеннями популярності та їх довірчими інтервалами pop-film не можна вважати найпопулярнішим, адже його інтервали перетинаються з інтервалами k-pop. Також треба зазначити, що рейтинг популярності змінився. Так як медіанне значення є стійкішим до зміщенності розподілу, неможливо вважати, що обрані жанри музики є найпопулярнішими.

## 5 артистів

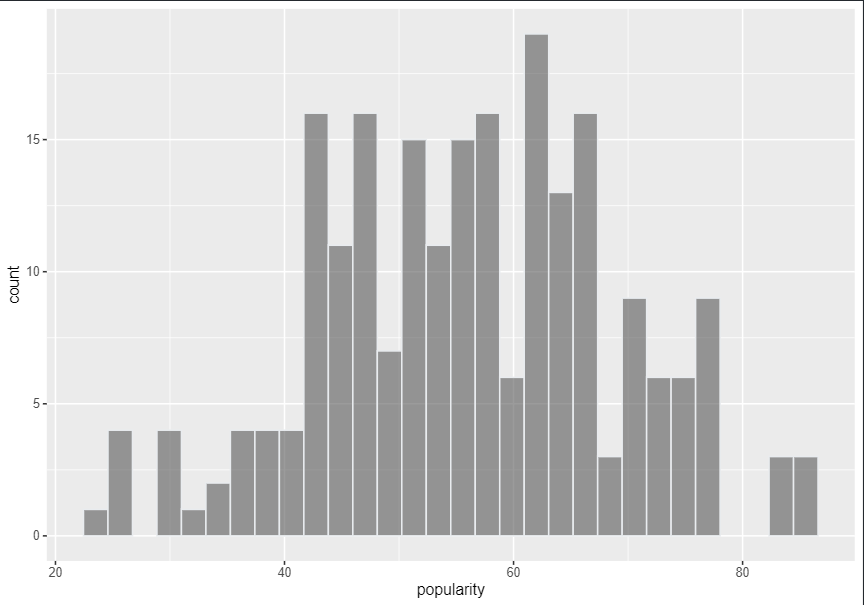
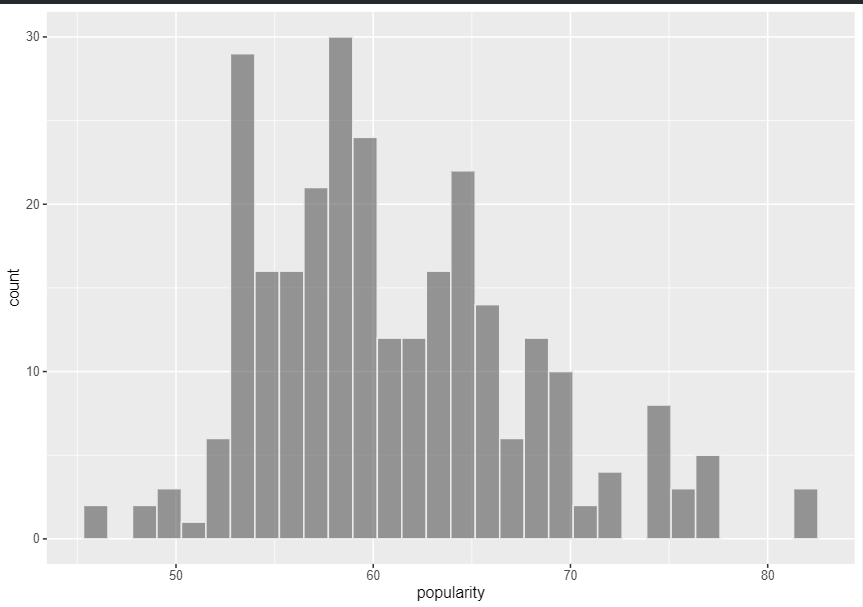
У першій лабораторній роботі було визначено 5 артистів, що мають в сумі найвищу популярність.

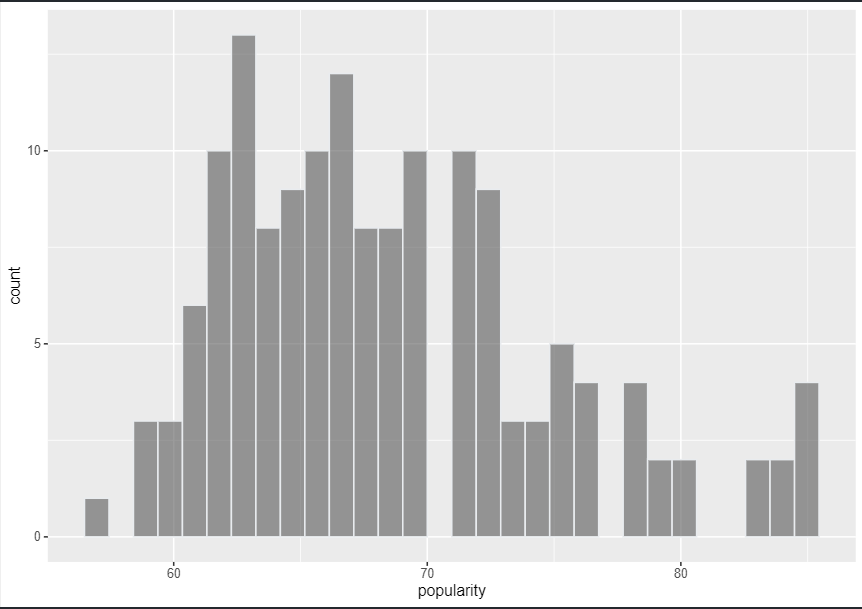
Було отримано такі імена/назви:



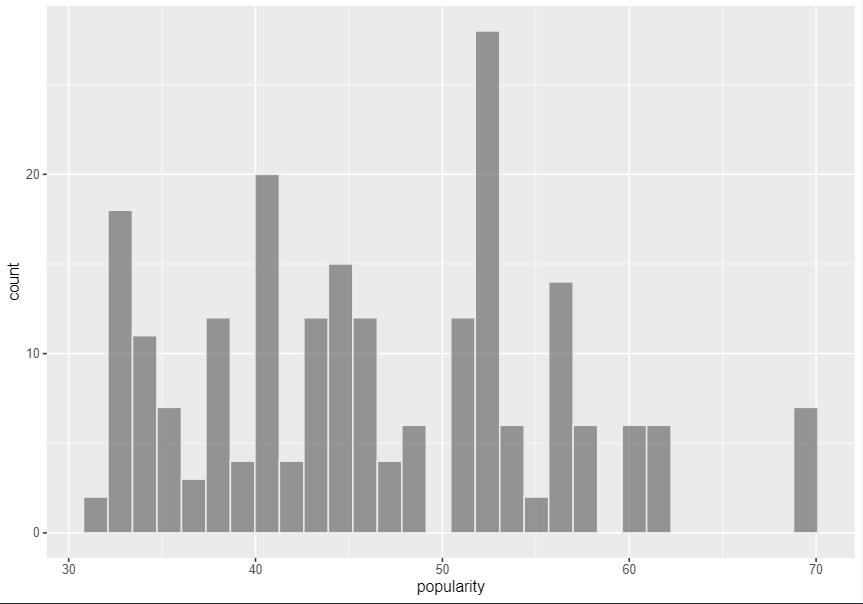
У нас виникло питання, чи справді можна вважати ці жанри найпопулярнішими.

Подивимося на розподіли популярності кожного артисту:

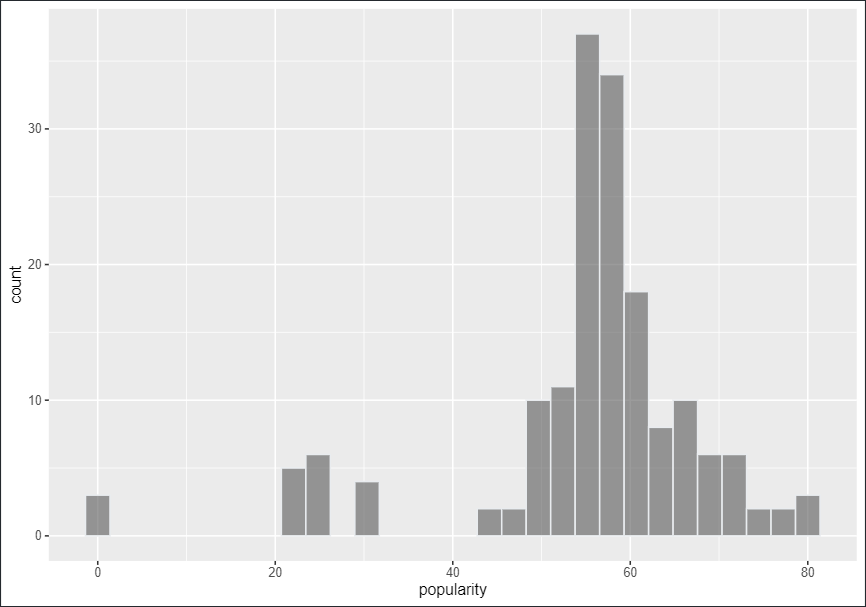
1. Розподіл популярності The beatles
2. Розподіл популярності Linkin Park
3. Розподіл популярності BTS



1. Розподіл популярності Prateek Kuhad



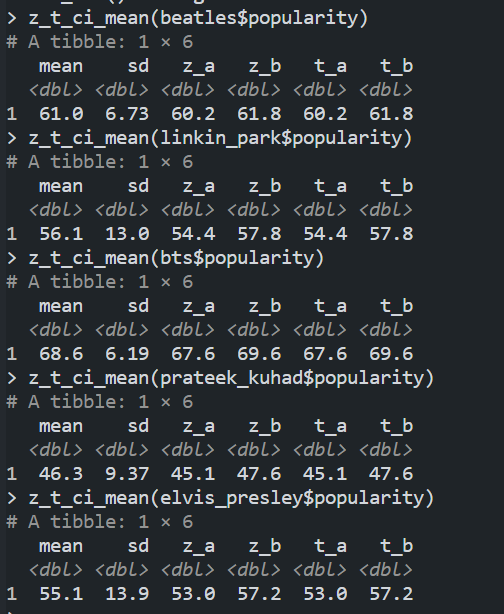
1. Розподіл популярності Elvis Presley



Як бачимо, кожен артист має скошений асимптотично нормальний розподіл. Отже, для оцінки популярності артистів доречніше орієнтуватися на медіанне значення, адже медіана є стійкішою за вибіркове середнє значення.

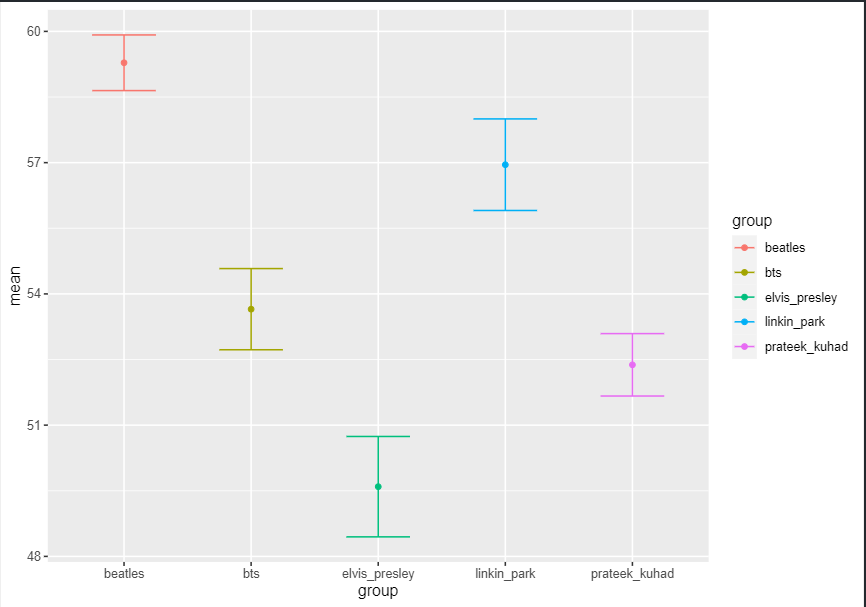
Оцінимо середнє значення популярності кожного артисту та побудуємо довірчі інтервали на основі квантилів стандартного нормального розподілу (z-розподілу) та t-розподілу.

Довірчі інтервали для The beatles, Linkin Park, BTS, Prateek Kuhad, Elvis Presley:



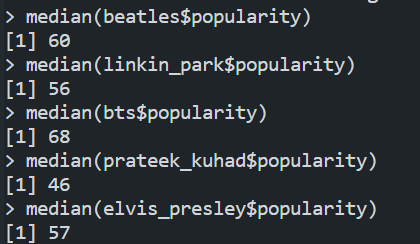
Бачимо, що довірчі інтервали з квантилями z-розподілу та t-розподілу майже співпадають.

Зобразимо на спільному графіку обраховані довірчі інтервали для популярності кожного артиста:

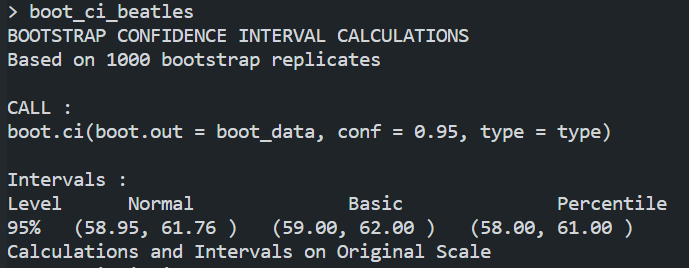
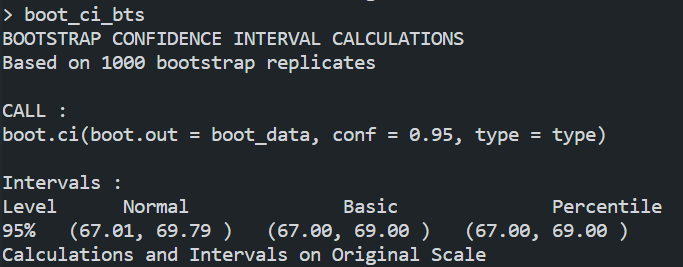


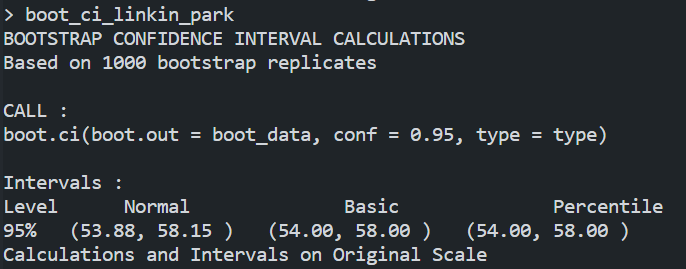
На графіку бачимо, що довірчі інтервали є доволі вузькими для кожного артиста. За середніми значеннями beatles справді можна вважати найпопулярнішим із обраних артистів. Довірчі інтервали для bts та prateek kuhad перетинаються, отже, не можна вважати точно, що bts є популярнішим за prateek kuhad.

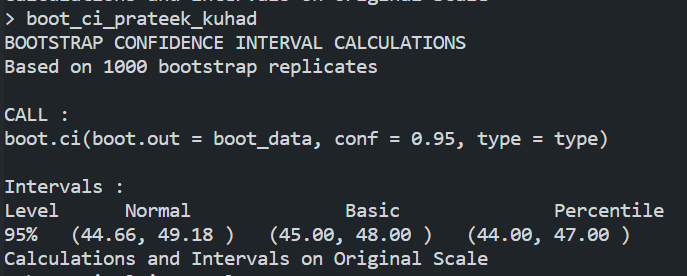
Оцінимо медіанне значення популярності кожного артиста та побудуємо довірчі інтервали на основі бутстрепу, так як важко оцінити асимптотичну дисперсію медіани. Для бутстрепу оберемо кількість реплікацій - 1000. Для кожного артиста обраховані довірчі інтервали типу Normal, Basic, Percentile.

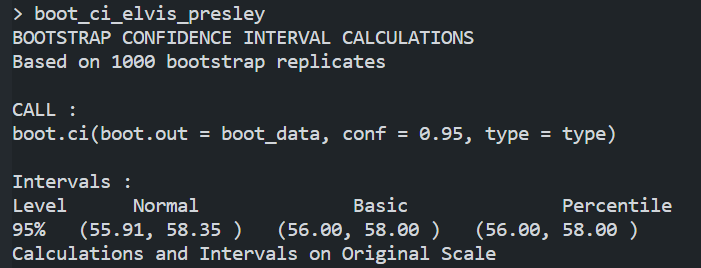
Медіани для всіх артистів: 

Довірчі інтервали для всіх артистів:



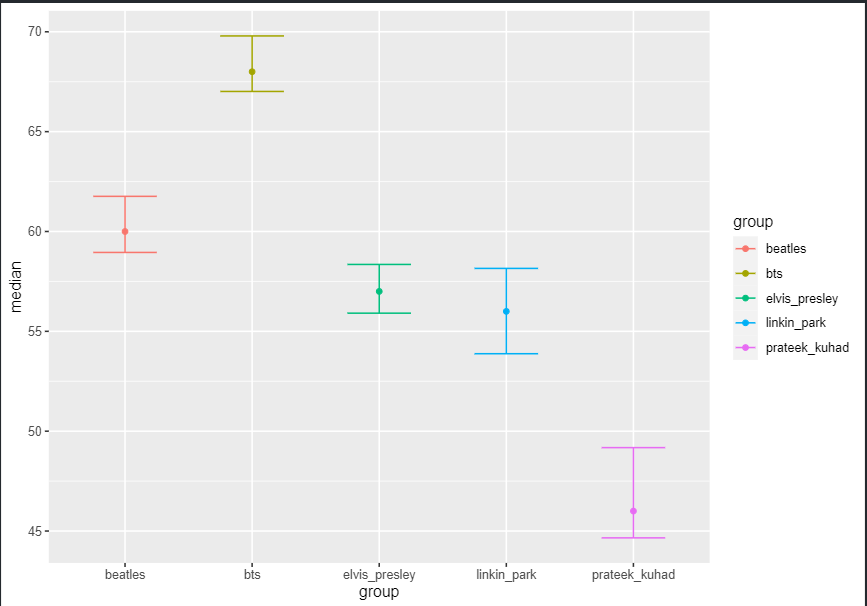






Видно, що довірчі інтервалів різних типів є скошеними для медіанних значень. Це пояснюється тим, що розподіли є скошеними в ліву або праву сторону.

Побудуємо спільний графік для кожного артиста та їхні довірчі інтервали типу Normal.



Бачимо, що рейтинг популярності змінився. Так як медіанне значення є стійкішим до зміщенності розподілу, неможливо вважати, що обрані артисти є найпопулярнішими.

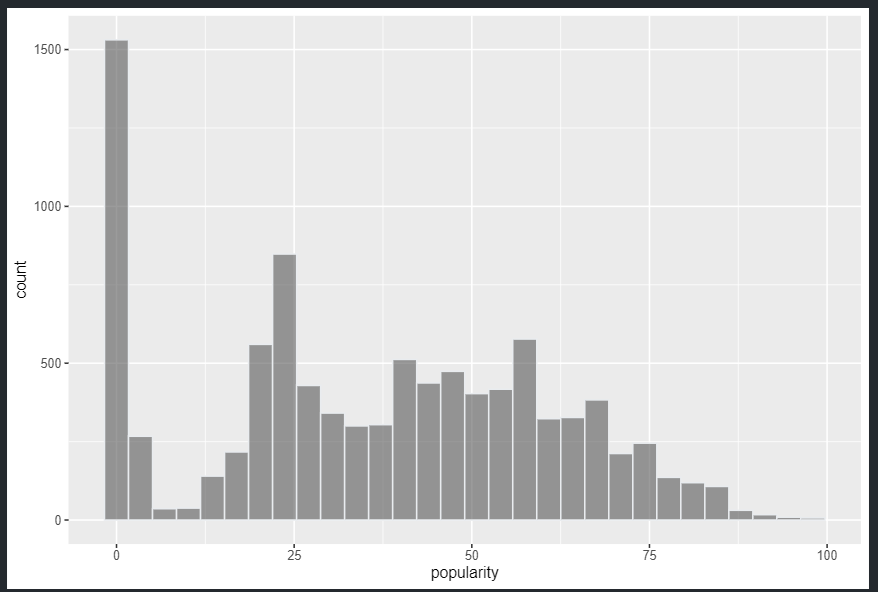
## Ненормативна лексика і популярність

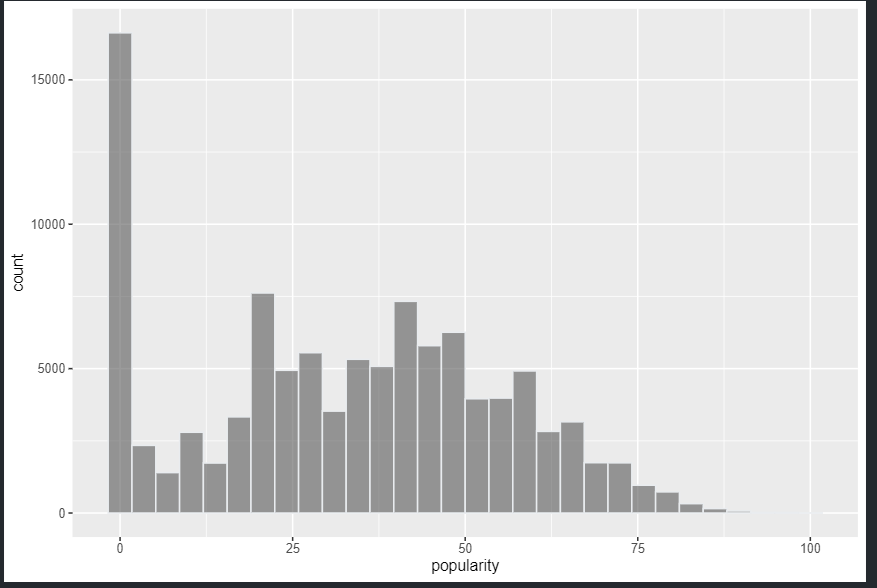
У минулій лабораторній роботі було розглянуто, які треки мають більшу популярність та мовність, з ненормативною лексикою чи без неї. Графічно було показано, що треки з ненормативною лексикою є популярнішими та мають вищий рівень мовності аніж треки без неї.

Спробуємо це підтвердити за допомогою довірчих інтервалів для середнього та медіанного значення популярності та мовності.

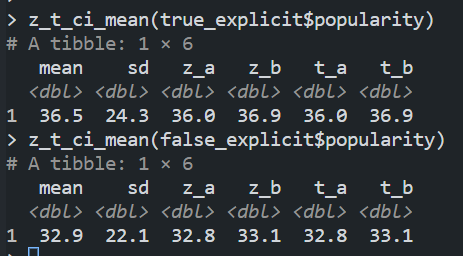
Розглянемо розподіли популярності для треків з explicit = True, та explicit = False.

**histogram\_plot\_builder(true\_explicit, aes(x=popularity))**

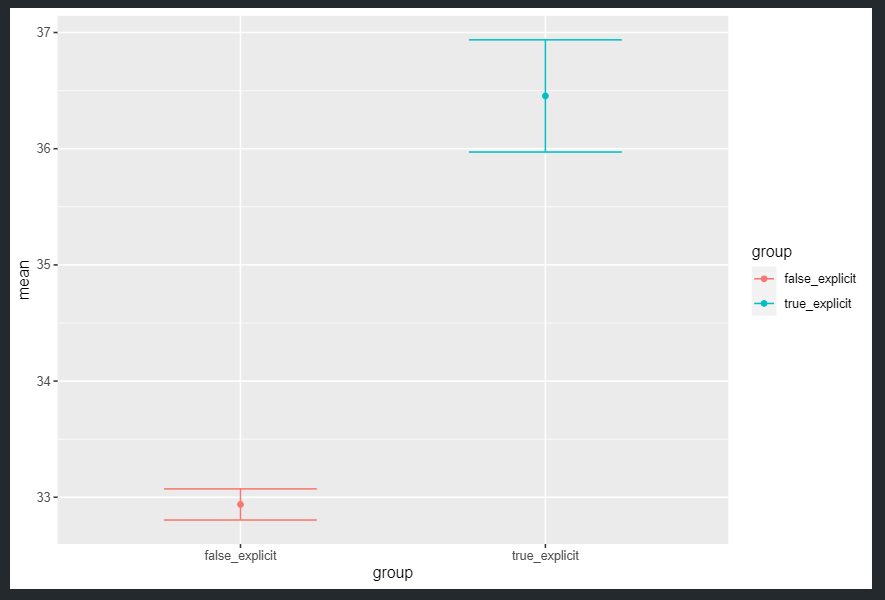


**histogram\_plot\_builder(false\_explicit, aes(x=popularity))**

Як бачимо розподіли є зміщеними, а отже, медіанні значення будуть точнішими для порівняння, адже медіана є стійкішою за середнє значення до зміщенності розподілів та викидів.

Оцінимо середнє значення для популярності двох категорій треків та побудуємо довірчі інтервали:

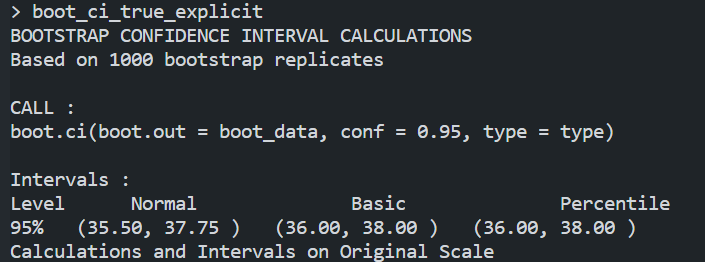
Побудуємо спільний графік довірчих інтервалів на основі квантилів t-розподілу.



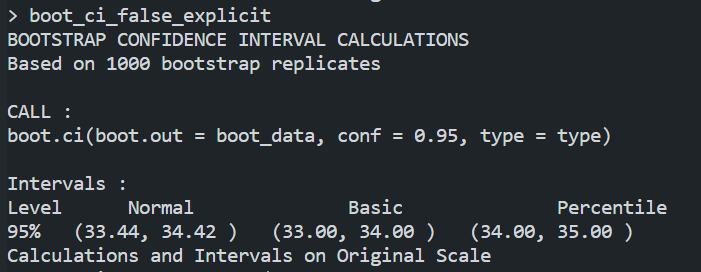
На графіку бачимо, що довірчі інтервали є доволі вузькими та знаходяться достатньо далеко один від одного. Отже, і справді треки з ненормативною лексикою є популярнішими за треки без ненормативної лексики за середнім значенням.

Оцінимо медіанне значення для популярності кожної групи та побудуємо бустреп довірчі інтервали:

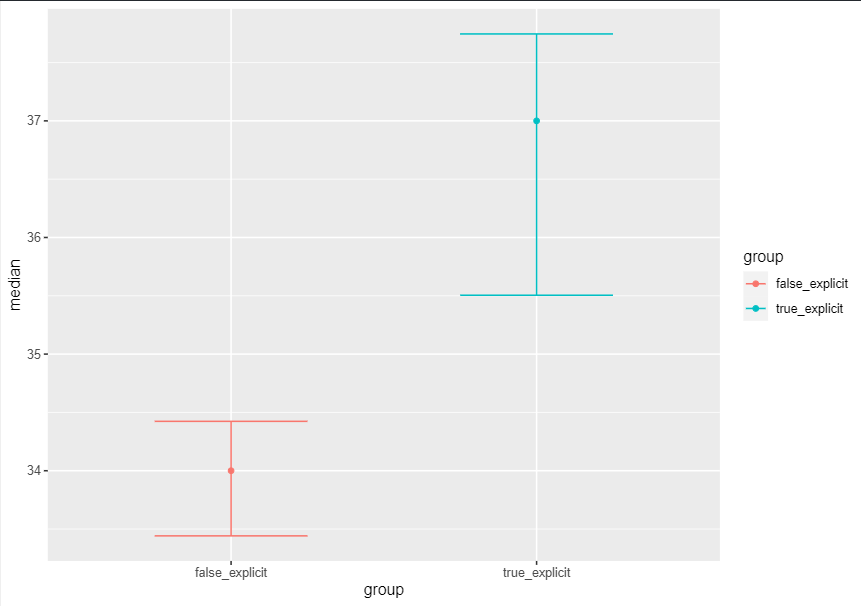








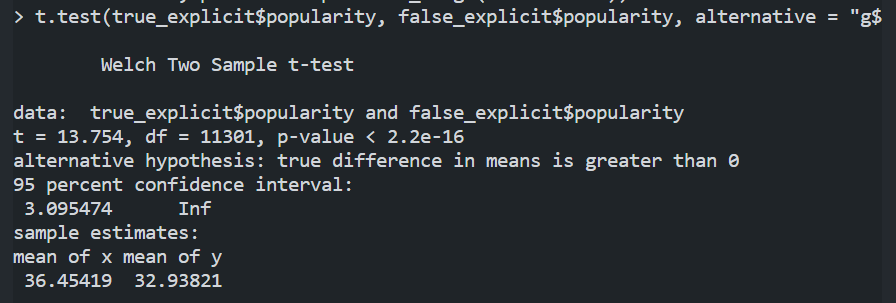
Побудуємо графічно бутстреп довірчі інтервали типу Normal:



Бачимо, що довірчі інтервали є трохи скошеними через скошеність розподілів, але також можна вважати, що за медіанним значенням трек з ненормативною лексикою є популярнішими за треки без ненормативної лексики.

Протестуємо правосторонню гіпотезу про рівність середнього значення популярності двох груп.

H0 (нульова гіпотеза): Середня популярність в true\_explicit$popularity не більша за середню популярність в false\_explicit$popularity.

H1 (альтернативна гіпотеза): Середня популярність в true\_explicit$popularity більша за середню популярність в false\_explicit$popularity.

Значення p-значення < 2.2e-16, що набагато менше за типовий рівень значущості 0.05. Це означає, що є статистично значущі докази в підтримку альтернативної гіпотези. Тобто, ми маємо достатні докази твердження, що середня популярність в true\_explicit$popularity є більшою за середню популярність в false\_explicit$popularity.

Додатково, інтервал довіри 95% для різниці середніх між двома групами (true\_explicit$popularity та false\_explicit$popularity) не має верхньої межі, яка означає, що різниця є статистично значущою та необмежено зверху.

Отже, на підставі цього тестування можна зробити висновок, що середня популярність в групі true\_explicit$popularity статистично значущо вища за середню популярність в групі false\_explicit$popularity.

Проведемо тест Волда для перевірки рівності медіанного значення популярності двох груп.

Код виконує тест на відмінність медіан між двома групами (true\_explicit$popularity та false\_explicit$popularity) за допомогою методу bootstrap та оцінює статистичну значущість цієї різниці. Нижче наведено пояснення для кожного кроку коду:

1. boot\_data\_true\_explicit і boot\_data\_false\_explicit: Виконується bootstrap-процедура, де медіана розраховується за допомогою функції boot\_median на вибірках true\_explicit$popularity та false\_explicit$popularity. Виконується 1000 ітерацій.
2. var\_median\_te і var\_median\_fe: Розраховуються дисперсії медіан з вибірок true\_explicit$popularity та false\_explicit$popularity на основі bootstrap-результатів.
3. median\_te і median\_fe: Розраховуються медіани оригінальних вибірок true\_explicit$popularity та false\_explicit$popularity.
4. se\_median: Обчислюється стандартна помилка різниці медіан на основі розрахованих дисперсій.
5. T: Обчислюється статистика T, яка представляє собою відношення різниці медіан до стандартної помилки.
6. p\_value: Розраховується p-значення на основі статистики T та нормального розподілу.
7. conf.int: Розраховується довірчий інтервал на рівні значущості 0.05 (95% довірчий інтервал) на основі розрахованої стандартної помилки.
8. median\_te, median\_fe, p\_value, conf.int: Виводяться результати тесту, включаючи значення медіан для обох груп, p-значення та довірчий інтервал.

Отже, цей тест використовує метод bootstrap для оцінки розподілу медіан в двох групах та розраховує статистику T для визначення статистичної значущості різниці між медіанами.

Результати цього тесту на відмінність медіан між двома групами (true\_explicit$popularity та false\_explicit$popularity) за допомогою методу bootstrap можна описати наступним чином:

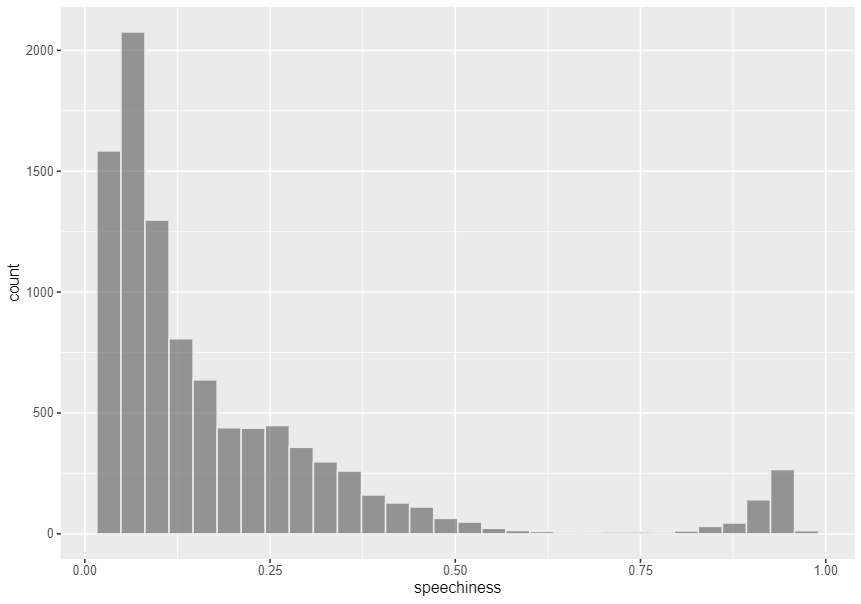
1. median\_te: Медіана в групі true\_explicit$popularity становить 37.
2. median\_fe: Медіана в групі false\_explicit$popularity становить 34.
3. p\_value: Значення p-значення дорівнює 3.271208e-07. Це значення набагато менше за типовий рівень значущості 0.05, що свідчить про наявність статистично значущої різниці між медіанами двох груп.
4. conf.int: Довірчий інтервал на рівні значущості 0.05 (95% довірчий інтервал) для різниці між медіанами становить від 2.008024 до нескінченності. Це означає, що різниця між медіанами є статистично значущою та необмеженою знизу.

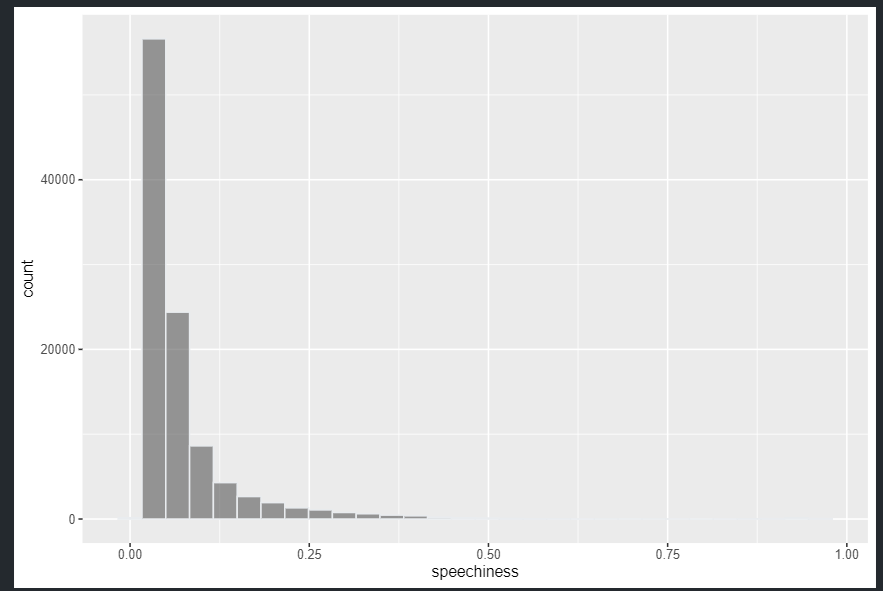
Отже, на підставі цих результатів можна зробити висновок, що медіана в групі true\_explicit$popularity статистично значущо вища за медіану в групі false\_explicit$popularity. Також, довірчий інтервал підтверджує, що різниця між медіанами є статистично значущою та необмеженою знизу.

## Ненормативна лексика і мовність

Розглянемо розподіли популярності для треків з explicit = True, та explicit = False.

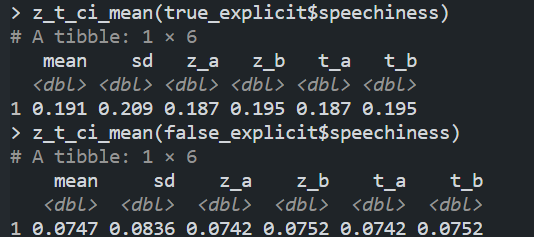
**histogram\_plot\_builder(true\_explicit, aes(x=speechiness))**



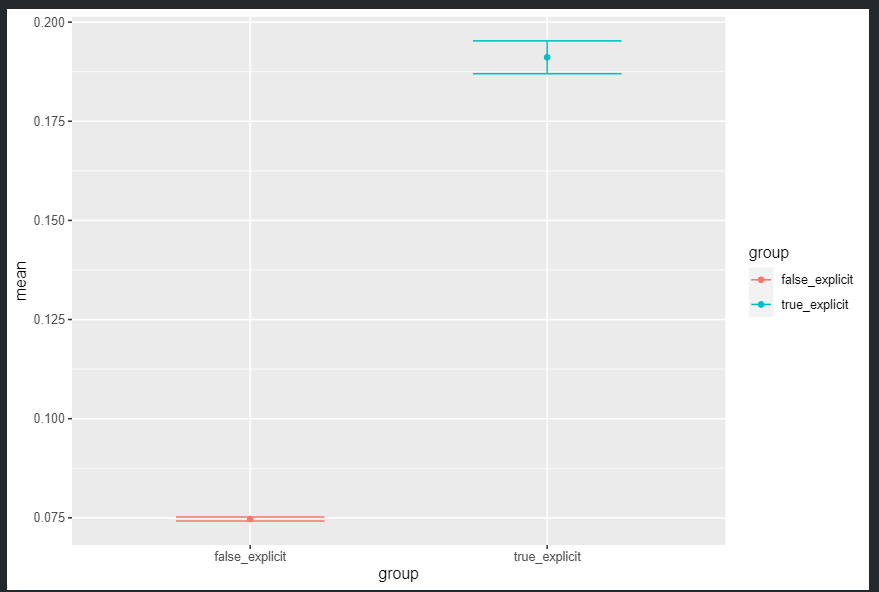
**histogram\_plot\_builder(false\_explicit, aes(x=speechiness))**

Як бачимо розподіли є зміщеними, а отже, медіанні значення будуть точнішими для порівняння, адже медіана є стійкішою за середнє значення до зміщенності розподілів та викидів.

Оцінимо середнє значення для мовності двох категорій треків та побудуємо довірчі інтервали:

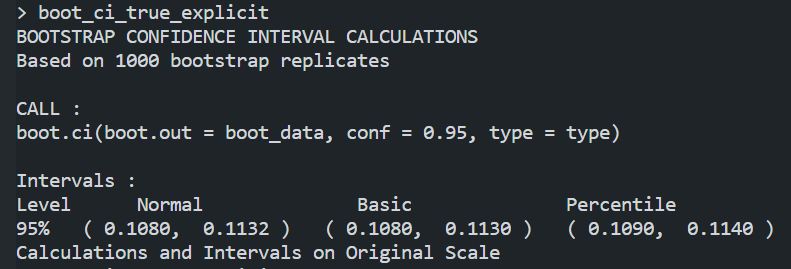


Побудуємо спільний графік довірчих інтервалів на основі квантилів t-розподілу.

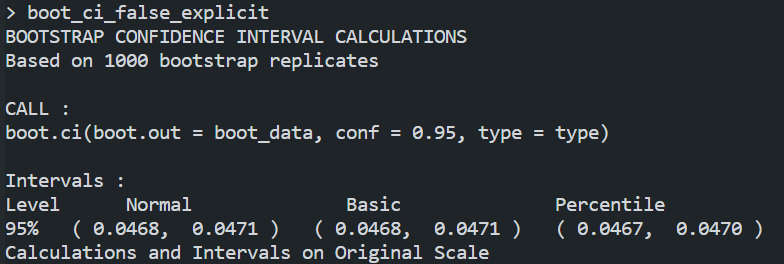


На графіку бачимо, що довірчі інтервали є доволі вузькими та знаходяться достатньо далеко один від одного. Отже, і справді треки з ненормативною лексикою мають вищий рівень мовності аніж треки без ненормативної лексики за середнім значенням.

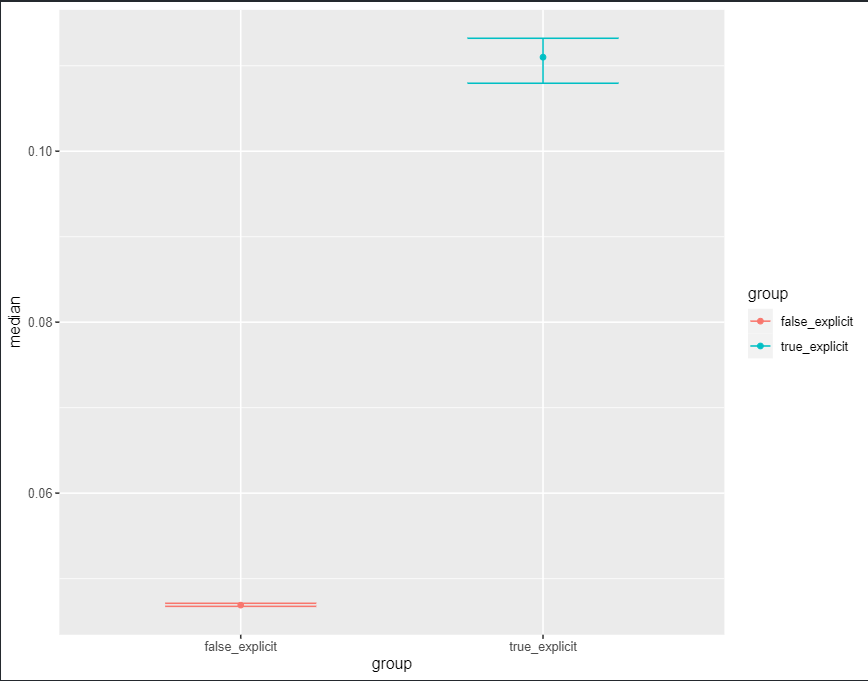
Оцінимо медіанне значення для мовності кожної групи та побудуємо бустреп довірчі інтервали:







Побудуємо графічно бутстреп довірчі інтервали типу Normal:

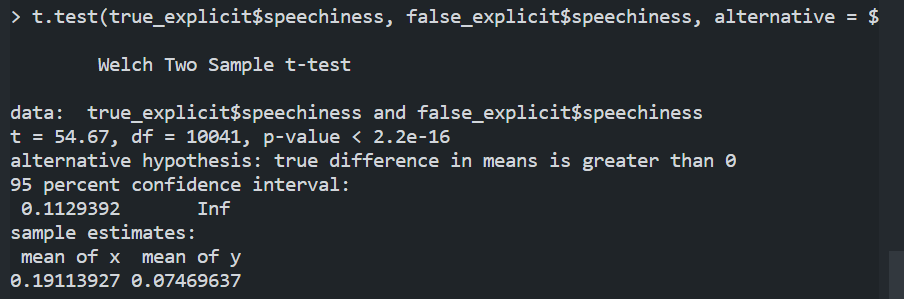


Бачимо, що довірчі інтервали є трохи скошеними через скошеність розподілів, але також можна вважати, що за медіанним значенням трек з ненормативною лексикою має більший рівень мовності аніж треки без ненормативної лексики.

Протестуємо правосторонню гіпотезу про рівність середнього значення мовності двох груп.

H0 (нульова гіпотеза): Середня рівень "speechiness" в true\_explicit$speechiness не більша за середню рівень "speechiness" в false\_explicit$speechiness.

H1 (альтернативна гіпотеза): Середня рівень "speechiness" в true\_explicit$speechiness більша за середню рівень "speechiness" в false\_explicit$speechiness.



Значення p-значення < 2.2e-16, що набагато менше за типовий рівень значущості 0.05. Це означає, що є статистично значущі докази в підтримку альтернативної гіпотези. Тобто, ми маємо достатні докази твердження, що середня рівень "speechiness" в true\_explicit$speechiness є більшою за середню рівень "speechiness" в false\_explicit$speechiness.

Додатково, інтервал довіри 95% для різниці середніх між двома групами (true\_explicit$speechiness та false\_explicit$speechiness) не має верхньої межі, що означає, що різниця є статистично значущою та необмежено зверху.

Отже, на підставі цього тестування можна зробити висновок, що середня рівень "speechiness" в групі true\_explicit$speechiness статистично значущо вища за середню рівень "speechiness" в групі false\_explicit$speechiness.

Проведемо тест Волда для перевірки рівності медіанного значення мовності двох груп.

Код виконує тест на відмінність медіан між двома групами (true\_explicit$speechiness та false\_explicit$speechiness) за допомогою методу bootstrap та оцінює статистичну значущість цієї різниці. Нижче наведено пояснення для кожного кроку коду:

1. boot\_data\_true\_explicit і boot\_data\_false\_explicit: Виконується bootstrap-процедура, де медіана розраховується за допомогою функції boot\_median на вибірках true\_explicit$speechiness та false\_explicit$speechiness. Виконується 1000 ітерацій.
2. var\_median\_te і var\_median\_fe: Розраховуються дисперсії медіан з вибірок true\_explicit$speechiness та false\_explicit$speechiness на основі bootstrap-результатів.
3. median\_te і median\_fe: Розраховуються медіани оригінальних вибірок true\_explicit$speechiness та false\_explicit$speechiness.
4. se\_median: Обчислюється стандартна помилка різниці медіан на основі розрахованих дисперсій.
5. T: Обчислюється статистика T, яка представляє собою відношення різниці медіан до стандартної помилки.
6. p\_value: Розраховується p-значення на основі статистики T та нормального розподілу.
7. conf.int: Розраховується довірчий інтервал на рівні значущості 0.05 (95% довірчий інтервал) на основі розрахованої стандартної помилки.
8. median\_te, median\_fe, p\_value, conf.int: Виводяться результати тесту, включаючи значення медіан для обох груп, p-значення та довірчий інтервал.

Отже, цей тест використовує метод bootstrap для оцінки розподілу медіан в двох групах та розраховує статистику T для визначення статистичної значущості різниці між медіанами.

Результати цього тесту на відмінність медіан між двома групами (`true\_explicit$speechiness` та `false\_explicit$speechiness`) за допомогою методу bootstrap можна описати наступним чином:

1. `median\_te`: Медіана в групі `true\_explicit$speechiness` становить 0.111.

2. `median\_fe`: Медіана в групі `false\_explicit$speechiness` становить 0.0469.

3. `p\_value`: Значення p-значення дорівнює 0. Це означає, що спостережувана різниця між медіанами двох груп є статистично значущою на будь-якому рівні значущості (p-value < 0.05).

4. `conf.int`: Довірчий інтервал на рівні значущості 0.05 (95% довірчий інтервал) для різниці між медіанами становить від 0.06194288 до нескінченності. Це означає, що різниця між медіанами є статистично значущою та необмеженою зверху.

Отже, на підставі цих результатів можна зробити висновок, що медіана в групі `true\_explicit$speechiness` статистично значущо вища за медіану в групі `false\_explicit$speechiness`. Довірчий інтервал підтверджує, що різниця між медіанами є статистично значущою та необмеженою зверху.

## Кореляція energy та loudness

У минулій лабораторній було побудовано графік кореляцій між змінними:

Нас зацікавила достатньо висока додатна кореляція між змінними energy та loudness. Ми вирішили перевірити її статистичну значущість, побудувавши бутстреп довірчі інтервали.

У коді використовується метод bootstrap для оцінки довірчих інтервалів для коефіцієнта кореляції між змінними `energy` та `loudness` у вихідному наборі даних `data`. Основні кроки та пояснення для кожної функції:

1. `boot\_cor`: Ця функція обчислює коефіцієнт кореляції Пірсона між змінними `energy` та `loudness` на вибірці, що вибирається з даних `data` за допомогою методу bootstrap. Вона використовує функцію `cor` з параметром `method = "pearson"` для обчислення кореляції.

2. `boot\_ci\_cor`: Ця функція використовує метод bootstrap для розрахунку довірчих інтервалів для коефіцієнта кореляції. Вона викликає функцію `boot` з параметром `boot\_cor` для виконання bootstrap-процедури та обчислення розподілу коефіцієнта кореляції. Потім вона викликає функцію `boot.ci` з параметрами `type` і `conf` для розрахунку довірчих інтервалів. Параметр `type` визначає методи розрахунку довірчих інтервалів, а `conf` встановлює рівень довіри (у цьому випадку 0.95).

3. `boot\_ci`: Ця змінна містить результати функції `boot\_ci\_cor` - довірчі інтервали для коефіцієнта кореляції. Вона містить інформацію про нижню та верхню межі довірчих інтервалів для кожного вибраного методу.

Довірчі інтервали (`boot\_ci`) мають три типи:

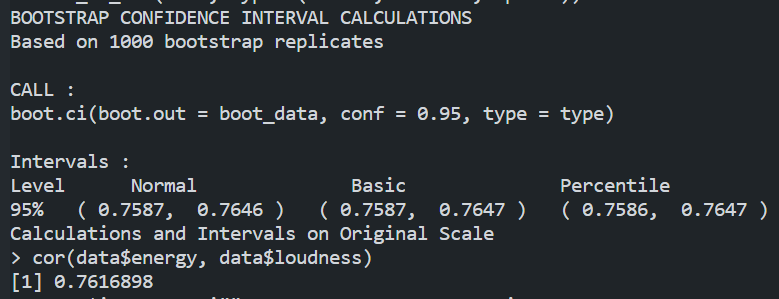
1. "norm" (нормальний): Цей тип довірчих інтервалів базується на нормальному розподілі. Він використовує стандартні властивості нормального розподілу для розрахунку довірчих інтервалів.

2. "basic" (базовий або

простий): Цей тип довірчих інтервалів базується на простих математичних формулах. Він використовує базові статистичні методи для розрахунку довірчих інтервалів.

3. "perc" (перцентильний): Цей тип довірчих інтервалів базується на розподілі перцентилів. Він використовує перцентильний метод для розрахунку довірчих інтервалів.

Кожен тип довірчих інтервалів містить нижню та верхню межу довірчого інтервалу для коефіцієнта кореляції. Вони представлені у вигляді числових значень.



Як бачимо, довірчі інтервали трьох типів є доволі вузькими, а отже, ми можемо довіряти, що між змінними energy та loudness існує тісний зв’язок.

# Висновки

У результаті виконання лабораторної роботи було знайдено відповіді на наші початкові питання такі як:

* Чи можна справді вважати обрані в першій роботі 5 жанрів музики найпопулярнішими за середнім та медіанним значенням?
* Чи можна справді вважати обрані в першій роботі 5 артистів найпопулярнішими за середнім та медіанним значенням?
* Які треки є популярнішими: з ненормативною лексикою чи без неї?
* Які треки мають більший рівень мовності: з ненормативною лексикою чи без неї?
* Чи є статистично значущою виявлена достатньо висока кореляція між змінною energy та loudness?

Ми виявили, що обрані нами 5 жанрів музики за медіанними значеннями популярності не можна вважати найпопулярнішими, адже їх довірчі інтервали перетинаються між собою.

Обрані нами 5 артистів також не можна вважати найпопулярнішими за медіаною.

Ми підтвердили за допомогою довірчих інтервалів та тестів Волда, що треки з ненормативною лексикою є популярнішими за треки без ненормативної лексики.

Також таким же чином підтверджено, що треки з ненормативною лексикою мають вищий рівень мовності аніж без.

З допомогою бустреп довірчих інтервалів виявлено, що кореляція між energy та loudness є статистично значущою.

# Список літератури

* Конспекти лекцій з предмету “Аналіз даних” [Лекції – Google Диск](https://drive.google.com/drive/folders/1LsWpozxahLgv51DOZKrIhKdIvhYMfEbX)
* Записи лекцій з предмету “Аналіз даних” [Записи лекцій – Google Диск](https://drive.google.com/drive/folders/1A9hTKNKyxRqC62QM7EN6FBaGJT7aV8tI)