Дослідимо структуру нашого датасету:

```
> head(birth)
  births wday month day_of_year day_of_month day_of_week
1   8068  Thu     1           1            1           5
2  10850  Fri     1           2            2           6
3   8328  Sat     1           3            3           7
4   7065  Sun     1           4            4           1
5  11892  Mon     1           5            5           2
6  12425  Tue     1           6            6           3
```

```
> str(birth)
'data.frame':   365 obs. of  6 variables:
 $ births      : int  8068 10850 8328 7065 11892 12425 12141 12094 11868
 8014 ...
 $ wday        : Factor w/ 7 levels "Fri","Mon","Sat",..: 5 1 3 4 2 6 7 5
 1 3 ...
 $ month       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ day_of_year : int  1 2 3 4 5 6 7 8 9 10 ...
 $ day_of_month: int  1 2 3 4 5 6 7 8 9 10 ...
 $ day_of_week : int  5 6 7 1 2 3 4 5 6 7 ...
```

Знайдемо топ 10 днів в році, коли народилося найбільше дітей:

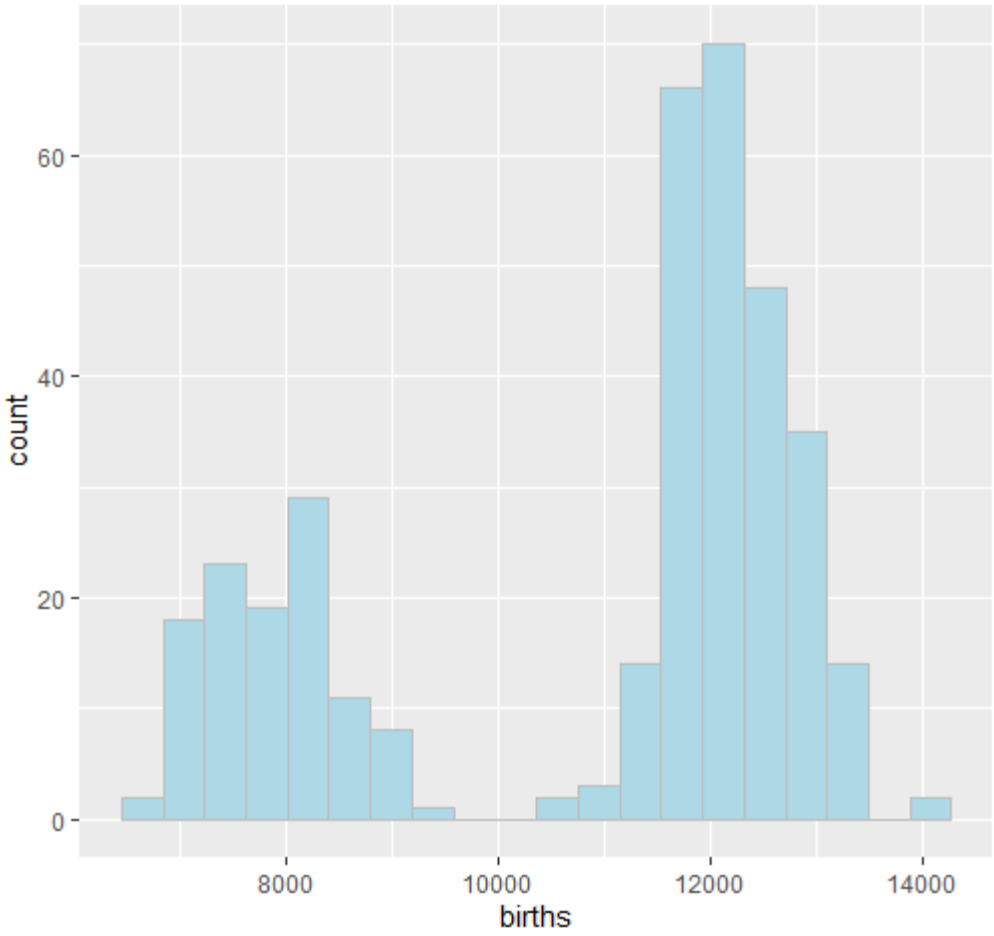
```
> birth %>%
+   top_n(n = 10, births) %>%
+   arrange(desc(births))
   births wday month day_of_year day_of_month day_of_week
1   13949  Wed     9         252            9           4
2   13925  Thu     9         253           10           5
3   13458  Tue    12         363           29           3
4   13439  Tue     9         258           15           3
5   13426  Tue     7         188            7           3
6   13319  Thu     9         260           17           5
7   13264  Tue     7         202           21           3
8   13240  Tue     7         209           28           3
9   13231  Tue     9         244            1           3
10  13229  Fri     9         261           18           6
```

Також знайдемо топ днів тижня по кількості народжуваних дітей:

```
> temp = aggregate(birth[, 1], list(birth$wday), mean)
> arrange(temp, desc(x))
  Group.1         x
1     Tue 12585.808
2     Wed 12279.096
3     Thu 12083.434
4     Fri 11834.558
5     Mon 11739.385
6     Sat  8357.096
7     Sun  7397.808
```

Побудуємо гістограму для кількості народжуваних:

```
ggplot(birth, aes(x=births)) +
  geom_histogram(bins=20, color="grey", fill="lightblue") +
  ylab("Days")
```



Знайдемо середнє значення:

```
> births_mean <- mean(birth$births)
> print(births_mean)
[1] 10899.99
```

Знайдемо середньоквадратичне відхилення:

```
> births_sd <- sd(birth$births)
> print(births_sd)
[1] 2076.684
```

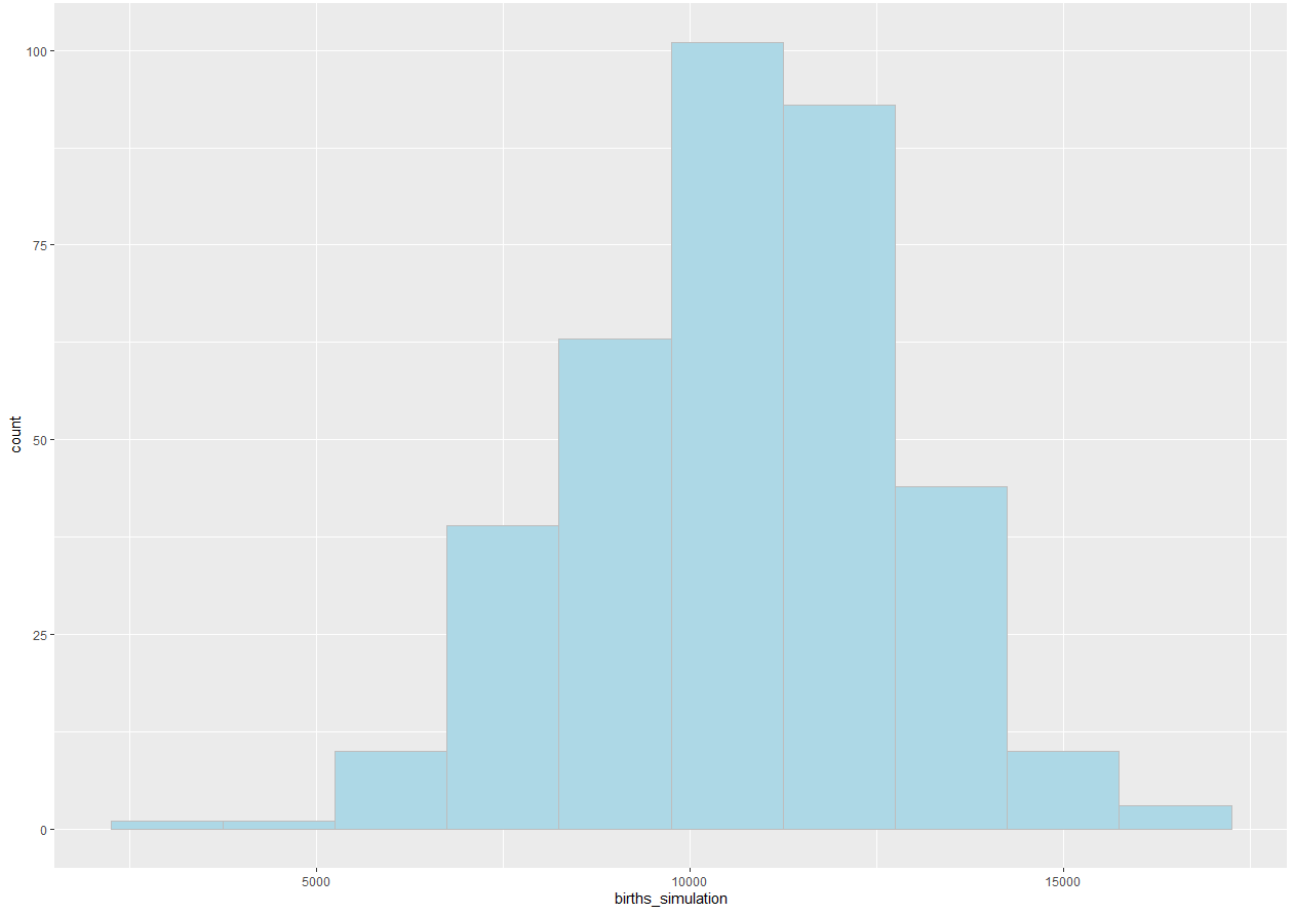
Згенеруємо нормальний розподіл, який має середнє значення births_mean та середньоквадратичне відхилення births_sd. Для цього використаємо функцію rnorm. Для того, щоб послідовність, яка генерується була сталою, при кожному виконанні нашого коду, встановимо параметр set.seed

```
set.seed(900)
births_simulation <- rnorm(n=nrow(birth), mean = births_mean, sd
= births_sd)
birth$births_simulation <- births_simulation
str(birth)
```

```
> set.seed(900)
> births_simulation <- rnorm(n=nrow(birth), mean = births_mean, sd =
hs_sd)
> birth$births_simulation <- births_simulation
> str(birth)
'data.frame':   365 obs. of  7 variables:
 $ births           : int  8068 10850 8328 7065 11892 12425 12141 12
1868 8014 ...
 $ wday             : Factor w/ 7 levels "Fri","Mon","Sat",..: 5 1 3
 6 7 5 1 3 ...
 $ month            : int  1 1 1 1 1 1 1 1 1 1 ...
 $ day_of_year      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ day_of_month     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ day_of_week      : int  5 6 7 1 2 3 4 5 6 7 ...
 $ births_simulation: num  11154 8619 10347 8474 11495 ...
```

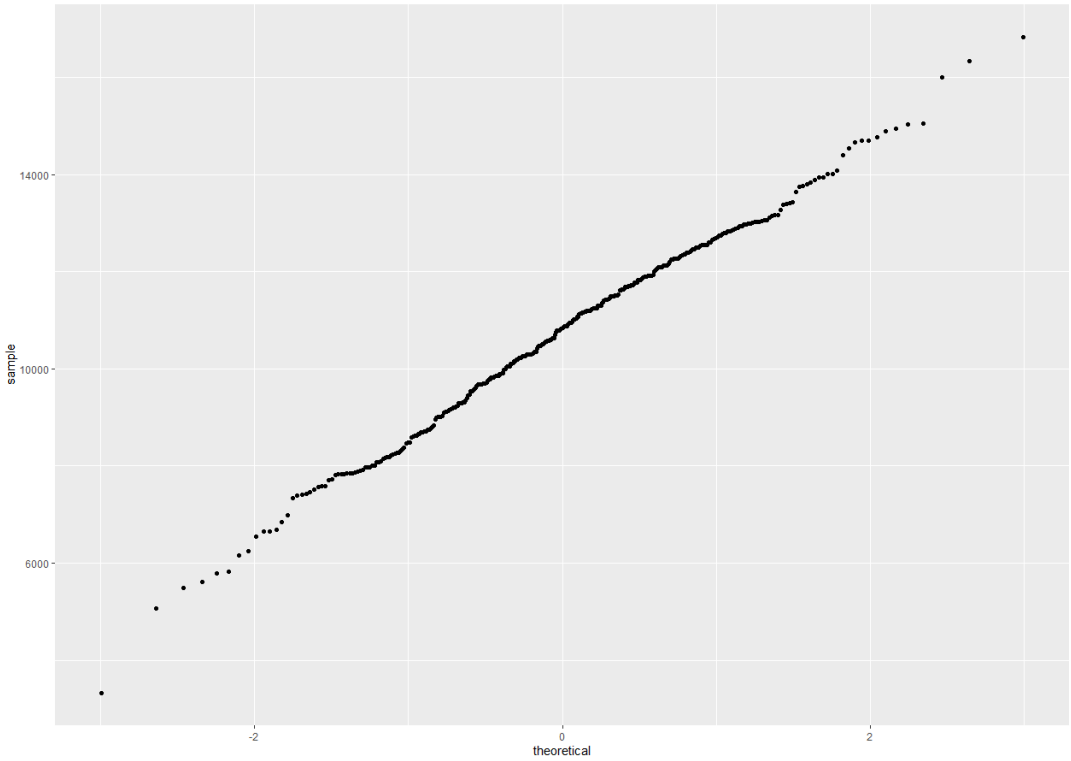
Побудуємо гістограму для цієї симуляції:

```
ggplot(birth, aes(x=births_simulation)) +
geom_histogram(bins=10, color="grey", fill="lightblue")
```

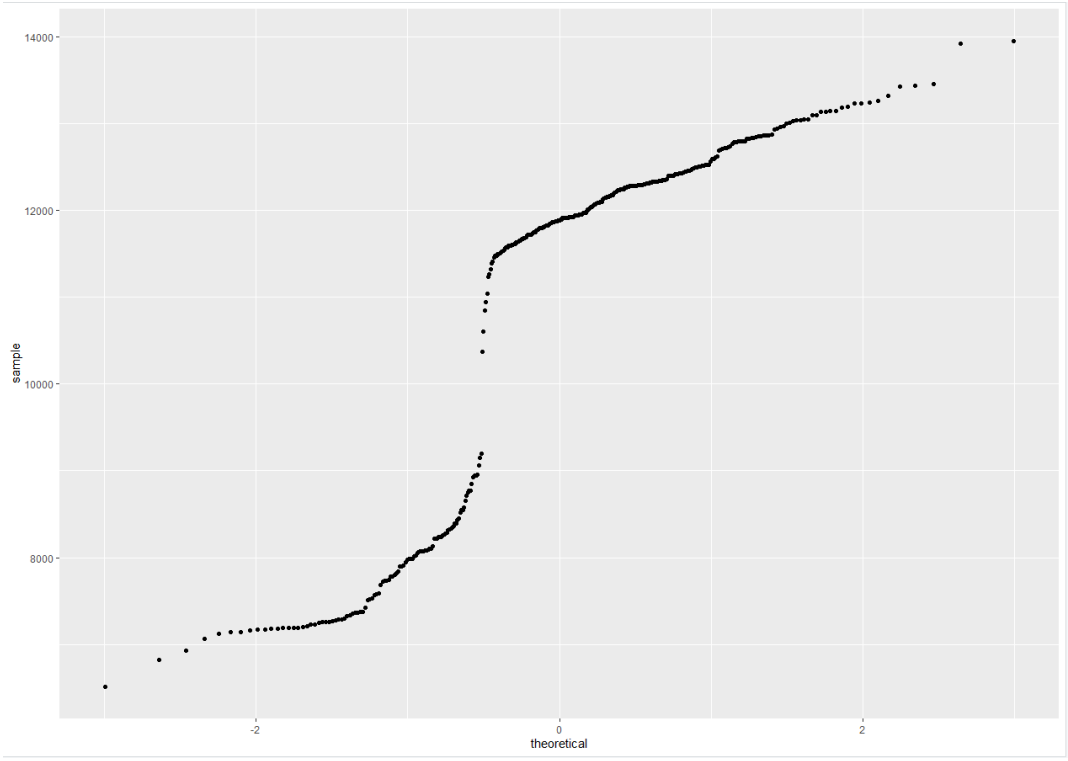


Для перевірки, чи є розподіл нормальним, використовується функція qqplot. Скористаємося нею для перевірки чи є нормально розподілені дані народжуваності. Спочатку побудуємо qqplot для нашої симуляції births_simulation:

ggplot(birth, aes(sample = births_simulation)) + stat_qq()



А тепер для справжніх даних: **ggplot**(movie_body_counts, **aes**(sample = IMDB_Rating)) + **stat_qq**()



Висновок

Знайшов середнє значення та середньоквадратичне відхилення для рейтингу даних фільмів, також побудував гістограму народжуваності та її симуляцію у вигляді нормального розподілу, в кінці перевірив чи нормально розподілена симуляція та оригінальні дані.