

Lab1

Завантаження пакетів

```
if (!require("pacman")) install.packages("pacman")

## Loading required package: pacman

pacman::p_load(pacman, ggplot2,
               plotly, rio, rmarkdown, moments)

library(pacman)
library(moments)
```

Завантаження даних

```
# CSV
data_csv <- import("/Users/victoria/Documents/weatherAUS.csv")
```

Опис даних

Датасет містить 10-річні щоденні спостереження погоди з багатьох метеорологічних станцій Австралії. Він містить 23 колонки та 145460 рядків. Джерело: <https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>

Розглянемо перші рядки датасету:

```
head(data_csv)

##           Date Location MinTemp MaxTemp Rainfall Evaporation Sunshine WindGustDir
## 1 2008-12-01  Albury      13.4    22.9      0.6          NA          NA          W
## 2 2008-12-02  Albury      7.4    25.1      0.0          NA          NA          WNW
## 3 2008-12-03  Albury     12.9    25.7      0.0          NA          NA          WSW
## 4 2008-12-04  Albury      9.2    28.0      0.0          NA          NA          NE
## 5 2008-12-05  Albury     17.5    32.3      1.0          NA          NA          W
## 6 2008-12-06  Albury     14.6    29.7      0.2          NA          NA          WNW
##   WindGustSpeed WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm Humidity9am
## 1             44          W          WNW           20           24          71
## 2             44         NNW          WSW           4            22          44
## 3             46          W          WSW          19           26          38
## 4             24          SE           E          11            9          45
## 5             41         ENE          NW           7           20          82
## 6             56          W           W          19           24          55
##   Humidity3pm Pressure9am Pressure3pm Cloud9am Cloud3pm Temp9am Temp3pm
## 1           22      1007.7      1007.1         8         NA      16.9      21.8
## 2           25      1010.6      1007.8        NA         NA      17.2      24.3
## 3           30      1007.6      1008.7        NA         2      21.0      23.2
## 4           16      1017.6      1012.8        NA         NA      18.1      26.5
## 5           33      1010.8      1006.0         7         8      17.8      29.7
## 6           23      1009.2      1005.4        NA         NA      20.6      28.9
##   RainToday RainTomorrow
## 1          No          No
## 2          No          No
## 3          No          No
## 4          No          No
## 5          No          No
## 6          No          No
```

Подальші обчислення будемо проводити над змінною MaxTemp - максимальна температура за Цельсієм за добу. Позначимо цю змінну X.

```
X <- data_csv$MaxTemp
head(X)
```

```
## [1] 22.9 25.1 25.7 28.0 32.3 29.7
```

Знайдемо середнє значення та медіану:

```
mean_max_temp<-mean(X,na.rm = TRUE) # mean
mean_max_temp
```

```
## [1] 23.22135
```

```
median_max_temp<- median(X, na.rm = TRUE)
median_max_temp
```

```
## [1] 22.6
```

Підрахуємо показники варіації: дисперсія, стандартне відхилення, коефіцієнт варіації та інтерквартильний розмах.

```
v<-var(X, na.rm = TRUE)
v
```

```
## [1] 50.68086
```

```
sd<-sd(X, na.rm = TRUE)
sd
```

```
## [1] 7.119049
```

```
sd/mean_max_temp # Coefficient of variation
```

```
## [1] 0.3065734
```

```
max(X, na.rm = TRUE)-min(X, na.rm = TRUE) # Range
```

```
## [1] 52.9
```

```
IQR(X, na.rm = TRUE) # Interquartile range
```

```
## [1] 10.3
```

Побудуємо коробковий графік:

```
boxplot(X, xlab = "Діаграма температури повітря в Австралії", col="orange")
```



Діаграма температури повітря в Австралії

Знайдемо квартилі:

```
quantile(X, na.rm = TRUE)
```

```
##      0%   25%   50%   75%  100%
## -4.8 17.9 22.6 28.2 48.1
```

Знайдемо перший та дев'ятий децилі:

```
quantile(X, na.rm = TRUE, prob = seq(0, 1, length = 11), type = 5)#Deciles
```

```
##      0%   10%   20%   30%   40%   50%   60%   70%   80%   90%  100%
## -4.8 14.5 16.9 18.9 20.7 22.6 24.7 27.0 29.6 32.9 48.1
```

Знайдемо коефіцієнт асиметрії та коефіцієнт ексцесу:

```
skewness(X, na.rm = TRUE) # Skewness
```

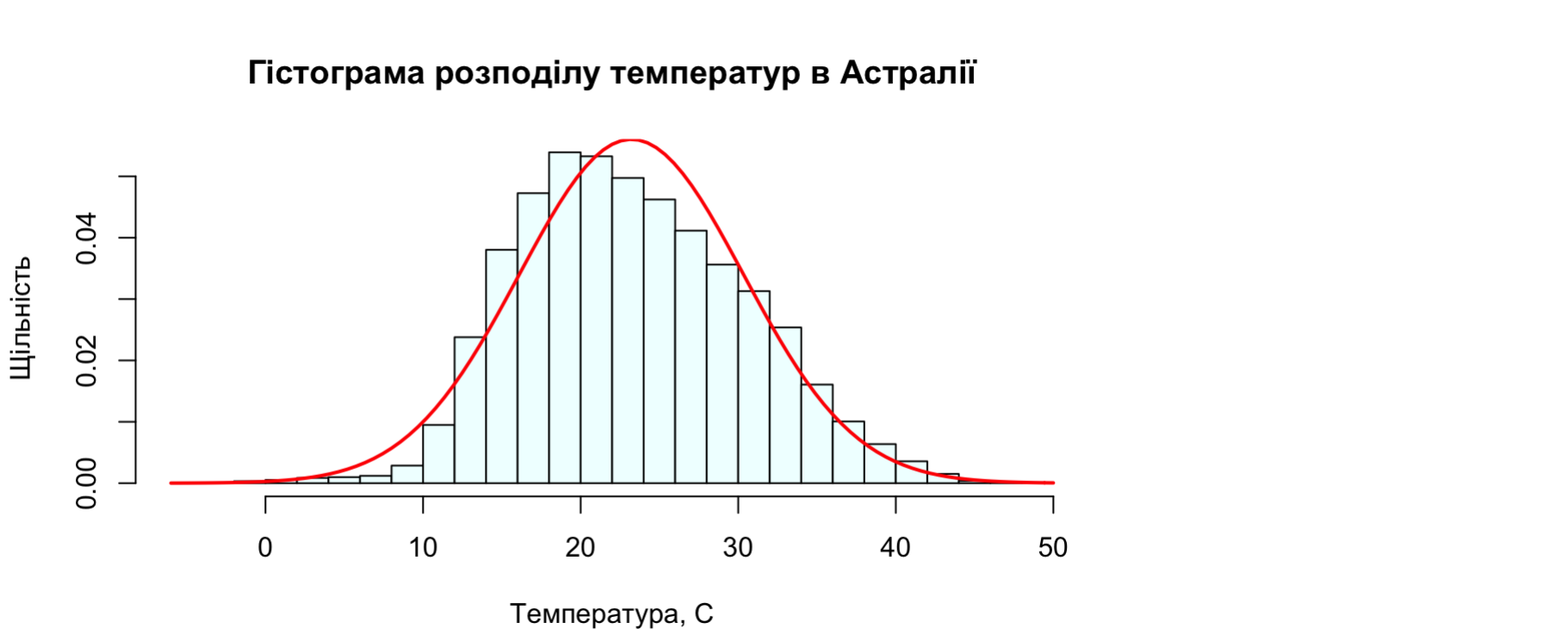
```
## [1] 0.2208371
```

```
kurtosis(X, na.rm = TRUE)
```

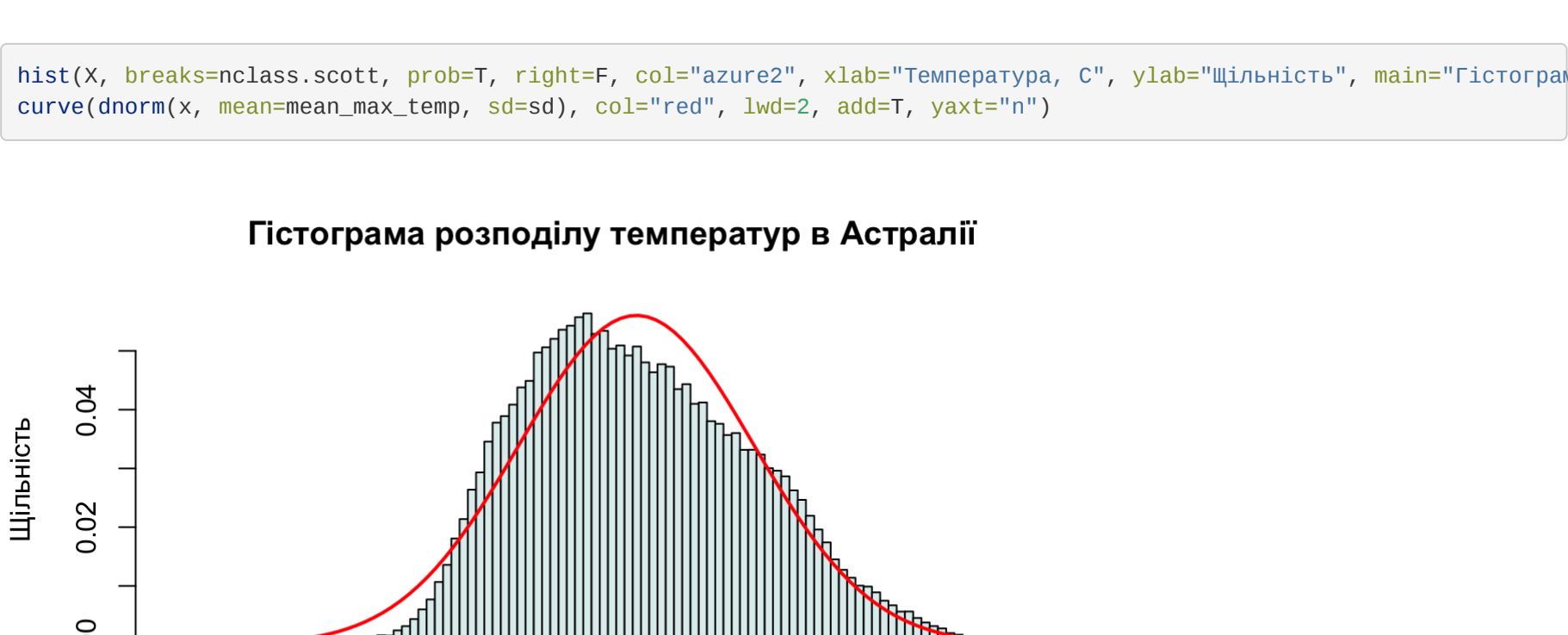
```
## [1] 2.775336
```

Побудуємо діаграми, використовуючи різні методи групування(базове правило, правило Скотта, правило Фрідмана-Діаконіса). На цьому ж графіку зобразимо оцінку щільності та графік щільності гіпотетичного розподілу.

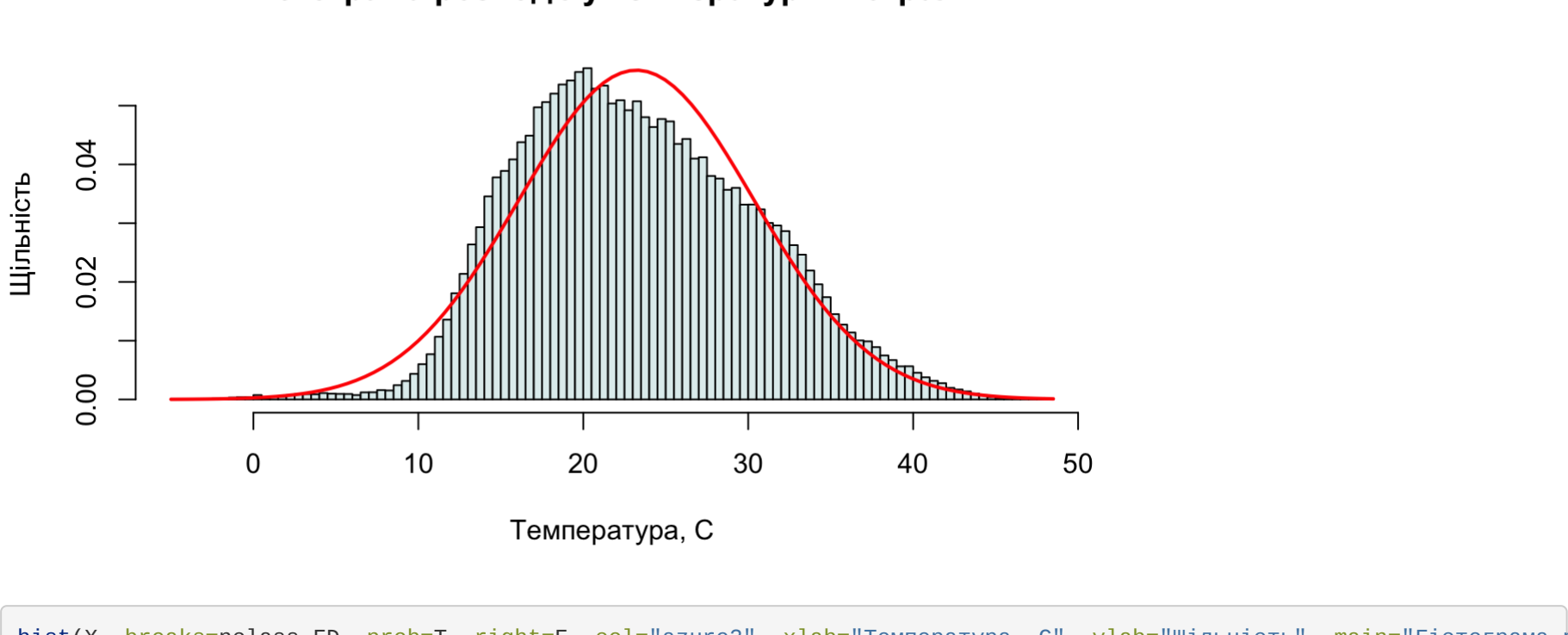
```
hist(X, breaks=nclass.Sturges, prob=T, right=F, col="azure", xlab="Температура, C", ylab="Щільність", main="Гістограма і
curve(dnorm(x, mean=mean_max_temp, sd=sd), col="red", lwd=2, add=T, yaxt="n")
```



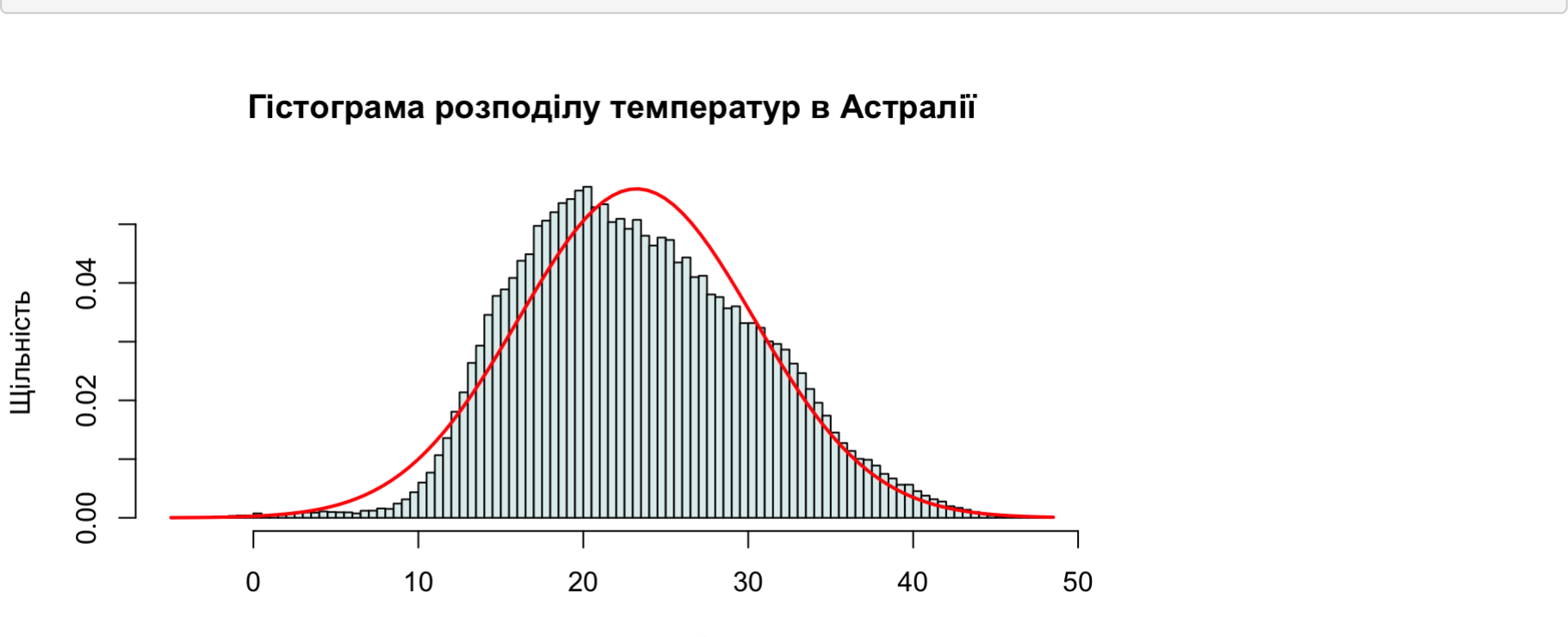
Гістограма розподілу температур в Астралії



Гістограма розподілу температур в Астралії



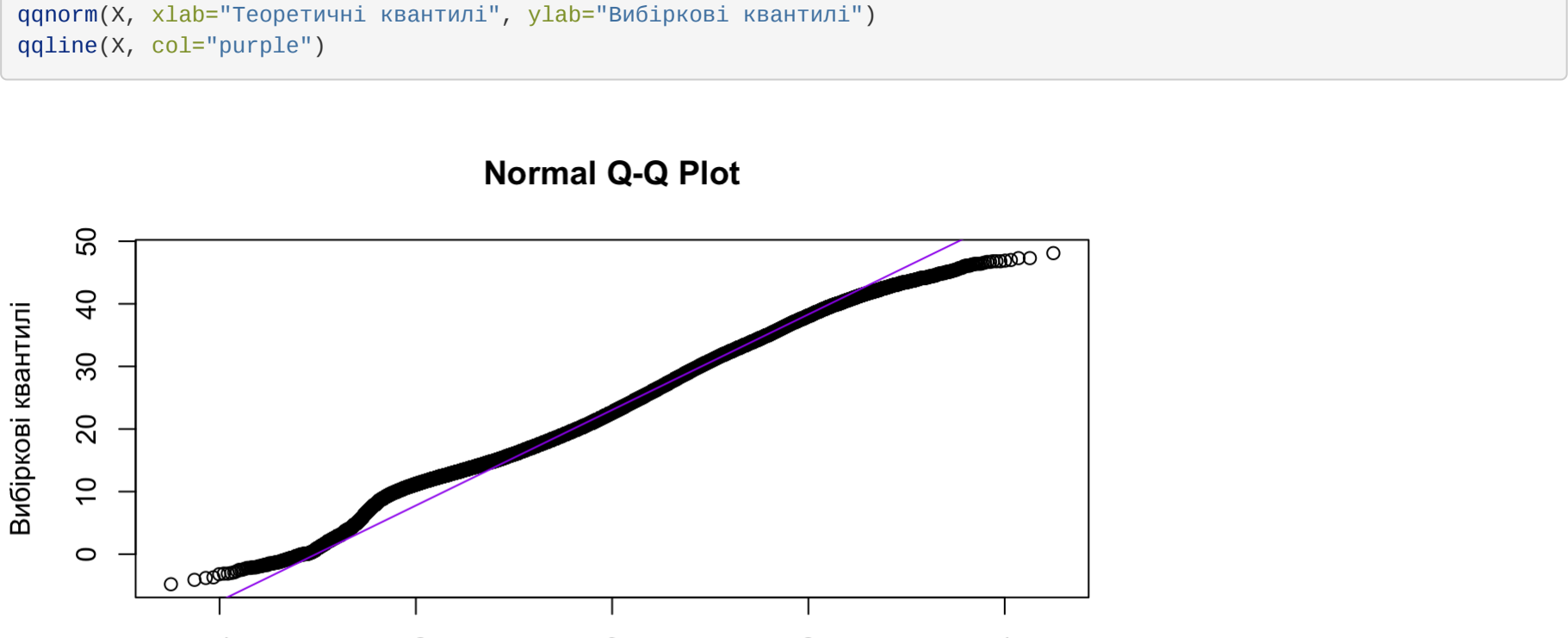
Гістограма розподілу температур в Астралії



Гістограма розподілу температур в Астралії

Зобразимо Q-Q діаграму для перевірки узгодженості з гіпотетичним розподілом.

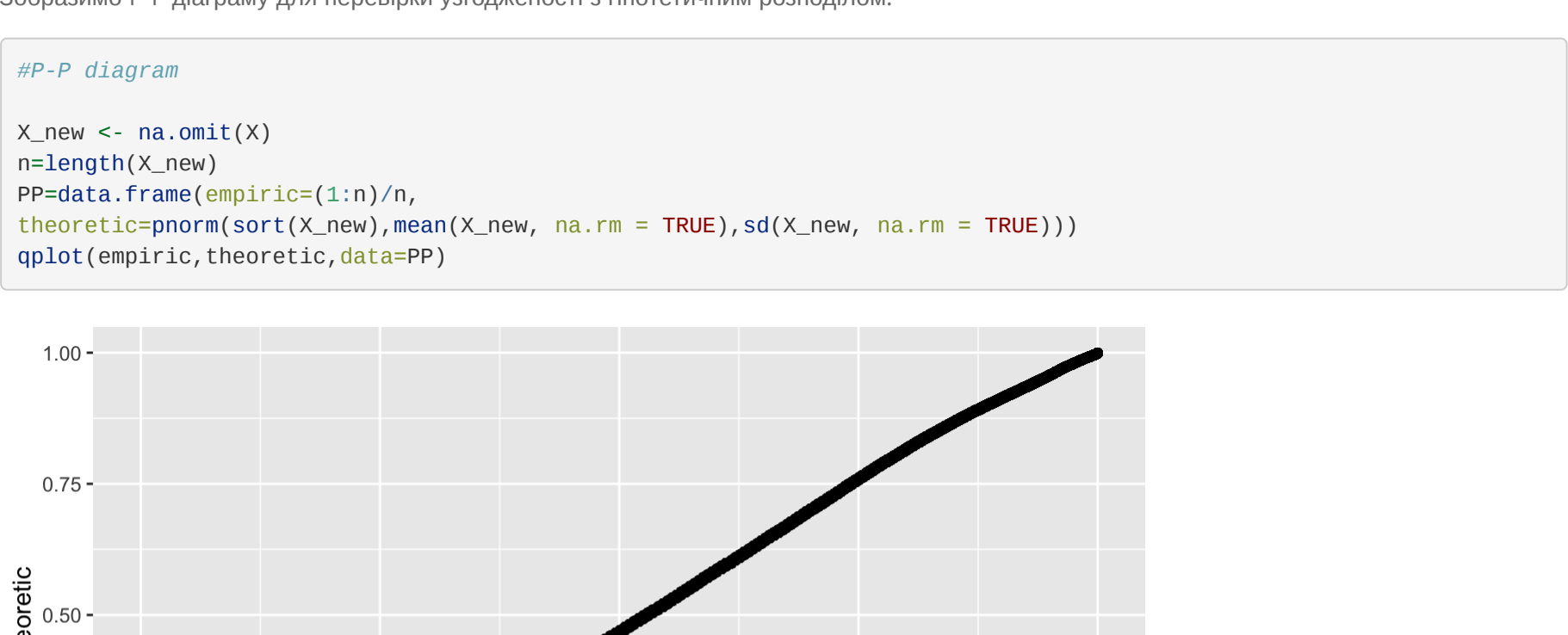
```
#Q-Q diagram
qqnorm(X, xlab="Теоретичні квантили", ylab="Вибіркові квантили")
qqline(X, col="purple")
```



Normal Q-Q Plot

Зобразимо P-P діаграму для перевірки узгодженості з гіпотетичним розподілом.

```
#P-P diagram
X_new <- na.omit(X)
n=length(X_new)
PP=data.frame(empiric=(1:n)/n,
               theoretic=pnorm(sort(X_new),mean(X_new, na.rm = TRUE),sd(X_new, na.rm = TRUE)))
plot(empiric,theoretic,data=PP)
```



За допомогою статистичного критерію перевіряємо згоду з гіпотетичним розподілом.

```
set.seed(23)
subsample <- sample(X, size=500, replace=TRUE)
shapiro.test(subsample)
```

```
##
## Shapiro-Wilk normality test
##
## data:  subsample
## W = 0.98811, p-value = 0.0004688
```