Statistics Homework

December 2023

Statistics Homework

Master of Science and Technology : Data Science for Business X-HEC

Solal Danan • Samuel Mesguiche • Ugo Benazra

Part 1 : Estimating parameters of a Poisson distribution to model the number of goals scored in football

We recall that the Poisson distribution with parameter $\theta > 0$ has a probability density function (pdf) given by $p(\theta, k)$, where $k \in \mathbb{N}$, with respect to the counting measure on $\mathbb{N}$:

$$p(\theta, k) = \frac{\exp(-\theta)\theta^k}{k!}$$

Question 1 : Is it a discrete or continuous distribution ? Can you give 3 examples of phenomena that could be modeled by such a distribution in statistics ?

Poisson distribution is a discrete probability distribution. It expresses the probability that an event will occur a certain number of time k (k ∈N) over fixed interval of time or space. Indeed, we can see that its density $p(\theta, .)$ is defined with respect to the counting measure on $\mathbb{N}$. In statistics, such a distribution can model such phenomena :

- **The number of incoming calls in an hour in a Call Center :** Assuming the call center has a steady average rate of calls per hour and the calls are independent of each other, the Poisson distribution can be used to predict the probability of receiving a certain number of calls within a given hour.
- **The number of emails a person receives in an hour :** If a person or an organization receives emails at a steady average rate, and the arrival of each email is independent of the others, then the Poisson distribution can be used to model the number of emails received in a set period, like an hour.
- **The number of equipment failures in a month :** The failure of equipment like light bulbs in a building can be thought of as random events, especially if the bulbs are of the same type and used in similar conditions. If we know the average failure rate, say 2 bulbs per month in a large office, we can use the Poisson distribution to estimate the probability of different numbers of failures over the next month.

Question 2 : Compute the mean and the variance of this distribution as a function of θ.

Let X be a random variable following a Poisson distribution with a parameter θ.

Espérance (μ) :

$$\begin{aligned}
\mu &= E(X) \\
&= \sum_{k=0}^{\infty} k \cdot P(X=k) \\
&= \sum_{k=0}^{\infty} k \cdot \frac{e^{-\theta}\theta^k}{k!} \\
&= \theta \sum_{k=1}^{\infty} \frac{e^{-\theta}\theta^{k-1}}{(k-1)!} \\
&= \theta \sum_{j=0}^{\infty} \frac{e^{-\theta}\theta^{j}}{j!} \quad \text{(setting } j = k-1\text{)} \\
&= \theta \cdot e^{-\theta} \sum_{j=0}^{\infty} \frac{\theta^{j}}{j!} \\
&= \theta \cdot e^{-\theta} \cdot e^{\theta} \quad \text{(the sum is the exponential series)} \\
&= \theta
\end{aligned}$$

Variance (σ^2) :

$$\begin{aligned}
\sigma^2 &= \mathrm{Var}(X) \\
&= E(X^2) - [E(X)]^2 \\
&= \sum_{k=0}^{\infty} k^2 \cdot P(X=k) - \theta^2 \\
&= \sum_{k=0}^{\infty} k^2 \cdot \frac{e^{-\theta}\theta^k}{k!} - \theta^2 \\
&= \sum_{k=1}^{\infty} (k(k-1)+k) \cdot \frac{e^{-\theta}\theta^k}{k!} - \theta^2 \\
&= \theta^2 \cdot e^{-\theta} \sum_{j=0}^{\infty} \frac{\theta^{j}}{j!} + \theta \cdot e^{-\theta} \sum_{j=0}^{\infty} \frac{\theta^{j}}{j!} - \theta^2 \\
&= \theta^2 \cdot e^{-\theta} \cdot e^{\theta} + \theta \cdot e^{-\theta} \cdot e^{\theta} - \theta^2 \quad \text{(the sum is the exponential series)} \\
&= \theta^2 + \theta - \theta^2 \\
&= \theta
\end{aligned}$$

Question 3 : What are our observations ? What distribution do they follow ? Write the corresponding statistical model. What parameter are we trying to estimate ?

We denote our n independent observations as $X_1, X_2, \ldots, X_n$. Each X_i follows a Poisson distribution and we denote $x_1, x_2, \ldots, x_n$ the realizations of $X_1, X_2, \ldots, X_n$.

Statistical Model : Our statistical model is described by the set of probability distributions $\{P_\theta : \theta \in \Theta\}$, where $\Theta = \mathbb{R}^{+*}$ represents the parameter space and $\{P_\theta\}$ represents the Poisson distribution.

Parameter to Estimate : The parameter we seek to estimate is denoted as θ, such that $\theta \in \mathbb{R}^{+*}$, which serves as a representation of the rate parameter in the Poisson distribution.

Question 4 : What is the likelihood function ? Compute the Maximum Likelihood Estimator $\hat{\theta}_{\text{MLE}}$.

$$X_1, ..., X_n \sim P(\theta) \text{ with } \theta > 0$$

The likelihood function for the realization $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$ is :

$$L_n(\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = e^{-\theta n} \cdot \theta^{\sum_{i=1}^{n} x_i} \cdot \prod_{i=1}^{n} \frac{1}{x_i!}$$

We use the log-likelihood $l(\theta)$:

$$l(\theta) = \log(L_n(\theta)) = -\theta n + \sum_{i=1}^{n} x_i \cdot \log(\theta) + \text{cst}$$

where $\text{cst} = \prod_{i=1}^{n} \frac{1}{x_i!}$.

We find a stationary point θ of $l(\theta)$ such that $\frac{\partial l(\theta)}{\partial \theta} = 0$:

$$\frac{\partial l(\theta)}{\partial \theta} = 0 \Leftrightarrow -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i = 0 \Leftrightarrow \theta = \frac{\sum_{i=1}^{n} x_i}{n}$$

To be sure that this is a maximum, we must show that l is concave :

$$\frac{\partial^2 l(\theta)}{\partial \theta^2} = \frac{-\sum_{i=1}^{n} x_i}{\theta^2}$$

$$\frac{\partial^2 l(\theta)}{\partial \theta^2} \text{ is then negative for all } \theta \text{ as } x_i \geq 0 \text{ for all } i \in \{1, \ldots, n\}$$

Then the MLE is given by $\hat{\theta}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Question 5 : Prove that $\sqrt{n}$ ($\hat{\theta}_{\text{MLE}} - \theta$) converges in distribution as n goes to infinity.

$X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables following a Poisson distribution with parameter θ.

For all $i \in \{1, \ldots, n\}$, $E\left[X_i^2\right] < +\infty$ and $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n \to +\infty]{a.s} E\left[X_1\right]$ by the strong law of large numbers (L.L.N.).

Then, by the Central Limit Theorem (T.C.L) :

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - E\left[X_1\right]\right) \xrightarrow[n \to +\infty]{d} N\left(0, \text{Var}\left(X_1\right)\right)$$

Therefore :

$$\sqrt{n}\left(\hat{\theta}_{\text{ML}} - \theta\right) \xrightarrow[n \to +\infty]{d} N(0, \theta)$$

Question 6 : Prove that $\frac{\sqrt{n}}{\sqrt{\hat{\theta}_{ML}}}$ ($\hat{\theta}_{\text{MLE}} - \theta$) converges in distribution as n goes to infinity.

Using the result of the last question :

$$\frac{\sqrt{n}}{\sqrt{\theta}}\left(\hat{\theta}_{ML} - \theta\right) \xrightarrow[n \to +\infty]{d} N(0, 1)$$

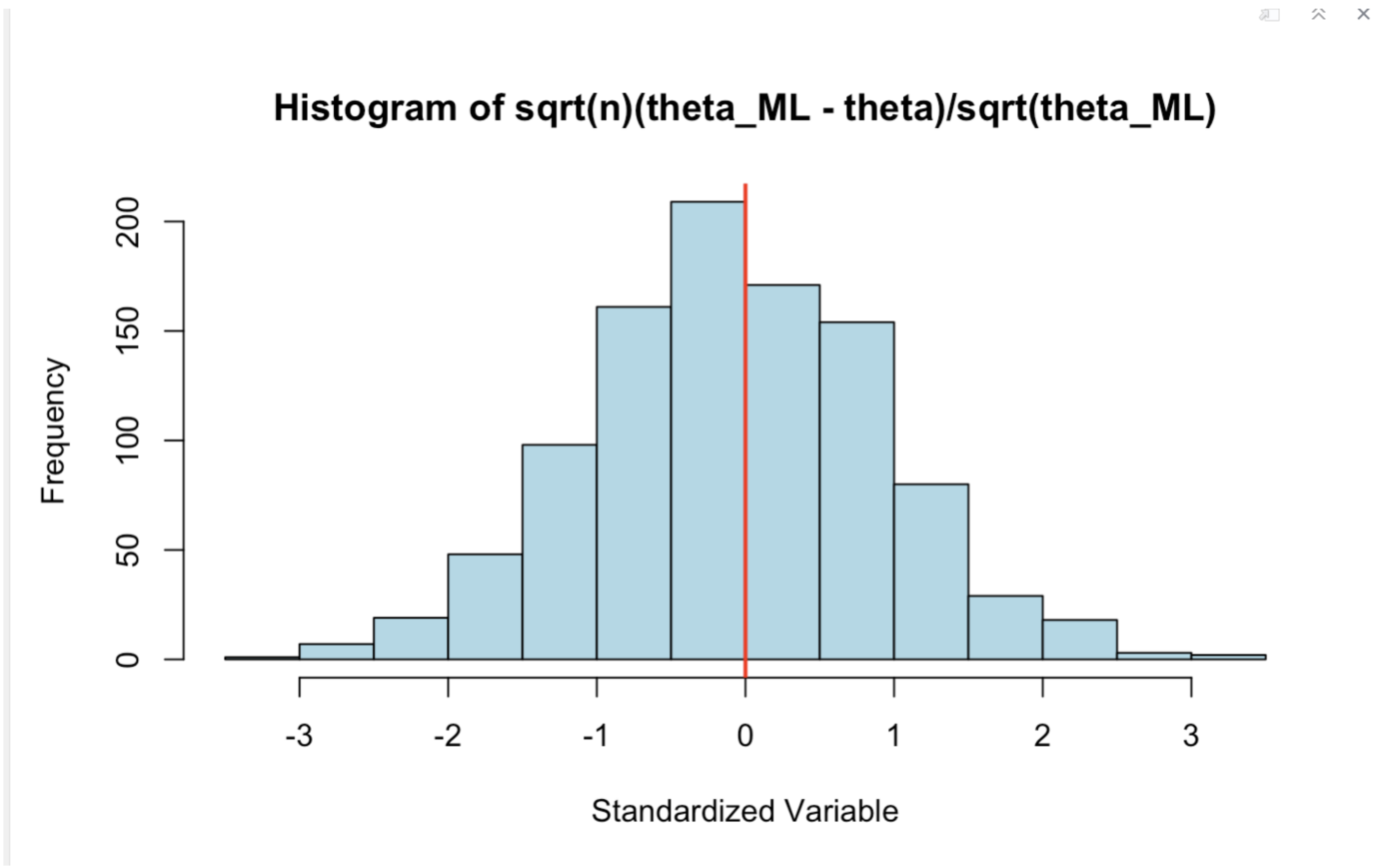
Also, $\frac{\hat{\theta}_{ML}}{\theta} \xrightarrow[n \to +\infty]{P} 1$, using again the L.L.N.

Using the continuity of the function $g : x \to \frac{1}{\sqrt{x}}$, we have, by the Slutsky theorem :

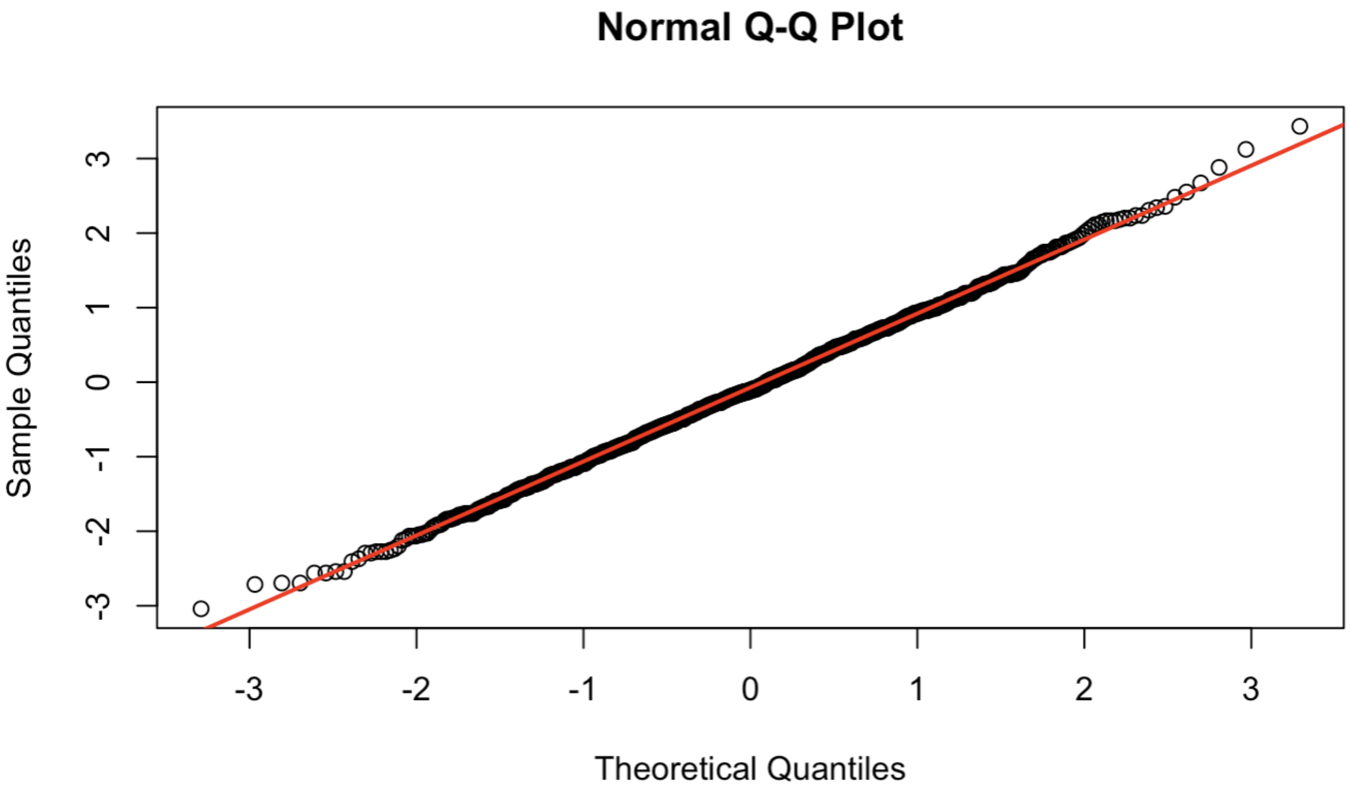
$$\frac{\sqrt{n}}{\sqrt{\hat{\theta}_{\text{ML}}}} \left(\hat{\theta}_{\text{ML}} - \theta\right) \xrightarrow[n \to +\infty]{d} N(0, 1)$$

On R, verify that the distribution of the random variable n ML is what you found theoretically, through a histogram and a QQ-plot (compute Nattempts = 1000 times the random variable n ^ML from a sample of size n of simulated Poisson data, with = 3, like in PC2).

```
# Set parameters
theta_true <- 3
n <- 1000  # sample size
N_attempts <- 1000  # number of simulation attempts

# Function to calculate MLE for Poisson distribution
mle_poisson <- function(data) {
  mean(data)
}

# Initialize vector for results
results <- numeric(N_attempts)

# Simulate and calculate MLE multiple times
for (i in 1:N_attempts) {
  simulated_data <- rpois(n, lambda = theta_true)
  theta_ML <- mle_poisson(simulated_data)
  results[i] <- sqrt(n) * (theta_ML - theta_true)/sqrt(theta_ML)
}

# Plot histogram
hist(results, main = "Histogram of sqrt(n)(theta_ML - theta)/sqrt(theta_ML)", col = "lightblue", xla
abline(v = 0, col = "red", lwd = 2)
```



```r
# QQ-Plot against a standardized normal distribution
qqnorm(results)
qqline(results, col = "red", lwd = 2)
```



```r
# Optional: Conduct a Shapiro-Wilk test for normality
shapiro.test(results)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  results
## W = 0.99903, p-value = 0.8895
```

Since the p-value is greater than the commonly chosen significance level of 0.05, you would not reject the null hypothesis. This means that, based on the Shapiro-Wilk test, there is no strong evidence to suggest that the data significantly deviates from a normal distribution.

The p-value obtained in this analysis exceeds the commonly accepted threshold of 0.05, a standard significance level in statistical tests. Consequently, this leads us to retain the null hypothesis. The implication of not rejecting the null hypothesis, particularly in the context of the Shapiro-Wilk test, is significant. It indicates that the evidence is not strong enough to conclude that the data deviate significantly from a normal distribution.

In simpler terms, as the p-value is higher than 0.05, it suggests that the observed data are consistent enough with a normal distribution.

Question 7 : For $\alpha \in$ (0, 1), give an asymptotic confidence interval of level for θ

For $\alpha \in (0,1)$, give an asymptotic confidence interval of level α, that is an interval $[a_n(\alpha,(X_i)_{i\in\{1,\ldots,n\}}); b_n(\alpha,(X_i)_{i\in\{1,\ldots,n\}})]$ such that :,

$$\lim_{n\to\infty} P(\theta \in [a_n(\alpha,(X_i)_{i\in\{1,\ldots,n\}}); b_n(\alpha,(X_i)_{i\in\{1,\ldots,n\}})]) \geq 1-\alpha$$

Let :

- $Z = \sqrt{n}\frac{\hat{\theta}_{ML}-\theta}{\sqrt{\hat{\theta}_{ML}}} \xrightarrow[n\to\infty]{\text{d}} N(0;1)$
- $q_{1-\frac{\alpha}{2}}$ be quantile of order $1-\frac{\alpha}{2}$ of a standard Gaussian

The symmetry of the standard Gaussian distribution implies that $q_{\frac{\alpha}{2}} = -q_{1-\frac{\alpha}{2}}$

$$\mathbb{P}\left(-q_{1-\frac{\alpha}{2}} \leq Z \leq q_{1-\frac{\alpha}{2}}\right) \xrightarrow[n\to\infty]{} 1-\alpha$$

$$\Rightarrow \mathbb{P}\left(-q_{1-\frac{\alpha}{2}} \leqslant \sqrt{n}\frac{\hat{\theta}_{ML}-\theta}{\sqrt{\hat{\theta}_{ML}}} \leqslant q_{1-\frac{\alpha}{2}}\right) \xrightarrow[n\to\infty]{} 1-\alpha$$

$$\Rightarrow \mathbb{P}\left(-q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}_{ML}}{n}} \leqslant \hat{\theta}_{ML}-\theta \leqslant q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}_{ML}}{n}}\right) \xrightarrow[n\to\infty]{} 1-\alpha$$

$$\Rightarrow \mathbb{P}\left(\hat{\theta}_{ML}-q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}_{ML}}{n}} \leqslant \theta \leqslant \hat{\theta}_{ML}+q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}_{ML}}{n}}\right) \xrightarrow[n\to\infty]{} 1-\alpha$$

$$\Rightarrow \mathbb{P}\left(\theta \in \left[\hat{\theta}_{ML}-q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}_{ML}}{n}}; \hat{\theta}_{ML}+q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}_{ML}}{n}}\right]\right) \xrightarrow[n\to\infty]{} 1-\alpha$$

For $\alpha \in (0;1)$ an asymptotic confidence interval of level α for θ is the interval :

$$\left[\hat{\theta}_{ML}-q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}_{ML}}{n}}; \hat{\theta}_{ML}+q_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{\theta}_{ML}}{n}}\right]$$

Question 8 :

— Propose two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of θ based on the first and second moments of a Poisson distribution.

On the one hand :

$$\mathbb{E}[X_1] = \theta$$

The random variables $(X_i)_{i=1,\ldots,n}$ are iid and integrable. Therefore, we can apply the Strong Law of Large Numbers. It implies that :

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s}} \mathbb{E}[X_1] = \theta$$

Let $\hat{\theta}_1 = \overline{X_n}$ our first estimator of θ.

What can you say about $\hat{\theta}_1$?

$\hat{\theta}_1$ is unbiased as $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \theta$

On the other hand :

$$\begin{aligned} &\mathbb{E}[X_1^2] = \theta^2 + \theta \\ \Rightarrow &\mathbb{E}[X_1^2] - \theta^2 = \theta \\ \Rightarrow &\mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \theta \end{aligned}$$

The random variables $(X_i)_{i=1,\ldots,n}$ are iid. Moreover, $\mathbb{E}[X_1^2] = \theta^2 + \theta < \infty$

Therefore, we can apply the Strong Law of Large Numbers. It implies that :

$$\begin{aligned} &\overline{X_n^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{\text{a.s}} \mathbb{E}[X_1^2] = \theta^2 + \theta \\ \Rightarrow &\overline{X_n^2} - \overline{X_n}^2 \xrightarrow{\text{a.s}} \theta \end{aligned}$$

Moreover,

$$\begin{aligned} \overline{X_n^2} - \overline{X_n}^2 &= \overline{X_n^2} - 2\overline{X_n}^2 + \overline{X_n}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\overline{X_n}\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} \overline{X_n}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2\overline{X_n}X_i + \overline{X_n}^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 \end{aligned}$$

Let $\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$ our second estimator of θ,

Question 9 : Compute the bias, the variance and the quadratic risk of $\hat{\theta}_{ML}$

$$\begin{aligned}
b(\hat{\theta}_{ML}) &= \mathbb{E}[\hat{\theta}_{\mathrm{ML}}] - \theta \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] - \theta \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] - \theta \text{ using the linearity of the expectation} \\
&= \frac{1}{n}\sum_{i=1}^{n} \theta - \theta \\
&= \frac{n}{n}\theta - \theta \\
&= 0
\end{aligned}$$

We show above that our estimator $\hat{\theta}_{\mathrm{ML}}$ is unbiased

We compute the variance of our estimator $\hat{\theta}_{\mathrm{ML}}$:

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}_{ML}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) \quad \text{because the random variables } (X_i)_{i=1,\ldots,n} \text{ are iid} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \theta \\
&= \frac{\theta}{n}.
\end{aligned}$$

We compute the quadratic risk of our estimator $\hat{\theta}_{\mathrm{ML}}$:

$$\begin{aligned}
\mathbb{E}[(\hat{\theta}_{\mathrm{ML}} - \theta)^2] &= \mathbb{E}[\hat{\theta}_{\mathrm{ML}}^2 - 2\hat{\theta}_{\mathrm{ML}}\theta + \theta^2] \\
&= \mathbb{E}[\hat{\theta}_{\mathrm{ML}}^2] - 2\mathbb{E}[\hat{\theta}_{\mathrm{ML}}]\theta + \theta^2 \\
&= \mathbb{E}[\hat{\theta}_{\mathrm{ML}}^2] - \mathbb{E}[\hat{\theta}_{\mathrm{ML}}]^2 - \theta^2 + \theta^2 \qquad \text{because } \mathbb{E}[\hat{\theta}_{\mathrm{ML}}] = \theta \\
&= \mathrm{Var}(\hat{\theta}_{ML}) + 0 \\
&= \mathrm{Var}(\hat{\theta}_{ML}) \\
&= \frac{\theta}{n}
\end{aligned}$$

Question 10 : Let $\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$, with $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Show that :

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2 - \left(\theta - \overline{X_n}\right)^2$$

We start from the right hand side of the equality and retrieve the first formula of $\hat{\theta}_2$:

$$
\begin{aligned}
\hat{\theta}_2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\theta)^2 - (\theta - \overline{X_n})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\theta + \theta^2) - (\theta^2 - 2\ \overline{X_n}\theta + \overline{X_n}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\theta\frac{1}{n}\sum_{i=1}^{n}X_i + \theta^2 - \theta^2 + 2\theta\overline{X_n} - (\overline{X_n})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\theta\overline{X_n} + 2\theta\overline{X_n} - (\overline{X_n})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X_n}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\overline{X_n}^2 + \overline{X_n}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\overline{X_n}\frac{1}{n}\sum_{i=1}^{n}X_i + \frac{1}{n}\sum_{i=1}^{n}\overline{X_n}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2\overline{X_n}X_i + \overline{X_n}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2
\end{aligned}
$$

Question 11 : Compute $\mathbb{E}[(\theta - \overline{X_n})^2]$.

Prove that $\hat{\theta}_2$ is a biased estimator of θ and give the bias. How can we get an unbiased estimator ?

$$
\begin{aligned}
\mathbb{E}[(\theta - \overline{X_n})^2] &= Var(\theta - \overline{X_n}) + \mathbb{E}[\theta - \overline{X_n}]^2 \\
&= Var(\overline{X_n}) + (\theta - \mathbb{E}[\overline{X_n}])^2 \\
&= Var(\overline{X_n}) + (\theta - \theta)^2 \\
&= Var(\overline{X_n}) \\
&= \frac{\theta}{n}
\end{aligned}
$$

Then :

$$
\begin{aligned}
\mathbb{E}[\hat{\theta}_2] &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i-\theta)^2 - (\theta - \overline{X}_n)^2] \qquad \text{(cf. Question 10)} \\
&= \mathbb{E}[(X_1-\theta)^2] - \mathbb{E}[(\theta - \overline{X}_n)^2] \qquad \text{because the random variables } (X_i)_{i=1,\dots,n} \text{ are iid} \\
&= \mathbb{E}[X_1^2] - 2\theta\mathbb{E}[X_1] + \theta^2 - \frac{\theta}{n} \qquad \text{(cf. higher)} \\
&= \theta^2 + \theta - \theta^2 - \frac{\theta}{n} \\
&= \frac{n-1}{n}\theta
\end{aligned}
$$

We can now compute the bias of $\hat{\theta}_2$:

$$
\begin{aligned}
b(\hat{\theta}_2) &= \mathbb{E}[\hat{\theta}_2] - \theta \\
&= -\frac{\theta}{n}
\end{aligned}
$$

To conclude, in order to have an unbiased estimator, we should take :

$$\hat{\theta}_3 = \frac{n}{n-1}\hat{\theta}_2$$

Part 2 : Application to Premier League scores.

QUESTION 1 :

Load the season-1718_csv file and describe what it contains. What do the variables FTHG, FTAG, FTR correspond to ?

```
data <- read.csv("season-1718_csv.csv", header = T,  sep = ",")
head(data)
```

```
##   Div     Date       HomeTeam     AwayTeam FTHG FTAG FTR HTHG HTAG HTR     Referee HS AS HST AST
## 1  E0 11/08/17        Arsenal    Leicester    4    3   H    2    2   D      M Dean 27  6  10   3
## 2  E0 12/08/17       Brighton     Man City    0    2   A    0    0   D    M Oliver  6 14   2   4
## 3  E0 12/08/17        Chelsea      Burnley    2    3   A    0    3   A    C Pawson 19 10   6   5
## 4  E0 12/08/17 Crystal Palace Huddersfield    0    3   A    0    2   A      J Moss 14  8   4   6
## 5  E0 12/08/17        Everton        Stoke    1    0   H    1    0   H N Swarbrick  9  9   4   1
## 6  E0 12/08/17    Southampton      Swansea    0    0   D    0    0   D     M Jones 29  4   2   0
##   HF AF HC AC HY AY HR AR B365H B365D B365A   BWH  BWD   BWA  IWH IWD   IWA   LBH  LBD   LBA   PS
## 1  9 12  9  4  0  1  0  0  1.53   4.5  6.50  1.50 4.60  6.75 1.47 4.5  6.50  1.44 4.40  6.50  1.5
## 2  6  9  3 10  0  2  0  0 11.00   5.5  1.33 11.00 5.25  1.30 8.00 5.3  1.35 10.00 5.00  1.30 10.9
## 3 16 11  8  5  3  3  2  0  1.25   6.5 15.00  1.22 6.50 12.50 1.22 6.2 13.50  1.25 5.75 15.00  1.2
## 4  7 19 12  9  1  3  0  0  1.83   3.6  5.00  1.80 3.50  4.75 1.85 3.5  4.30  1.80 3.40  4.60  1.8
## 5 13 10  6  7  1  1  0  0  1.70   3.8  5.75  1.70 3.60  5.50 1.70 3.7  5.00  1.67 3.60  5.25  1.7
## 6 10 13 13  0  2  1  0  0  1.62   4.0  6.50  1.57 4.00  6.00 1.65 3.8  5.30  1.60 3.70  6.00  1.6
##    PSD   PSA   WHH WHD   WHA   VCH  VCD   VCA Bb1X2 BbMxH BbAvH BbMxD BbAvD BbMxA BbAvA BbOU
## 1 4.55  6.85  1.53 4.2  6.00  1.53 4.50  6.50    41  1.55  1.51  4.60  4.43  6.89  6.44   37
## 2 5.55  1.34 10.00 4.8  1.33 10.00 5.50  1.33    40 11.50 10.10  5.60  5.25  1.36  1.32   35
## 3 6.30 15.25  1.25 5.5 13.00  1.25 6.25 15.00    41  1.27  1.24  6.55  6.06 15.50 13.67   36
## 4 3.58  5.11  1.80 3.3  5.00  1.83 3.60  5.00    41  1.86  1.81  3.65  3.50  5.11  4.82   36
## 5 3.83  5.81  1.70 3.5  5.50  1.70 3.80  5.75    40  1.71  1.69  3.85  3.69  6.00  5.50   35
## 6 3.94  6.35  1.62 3.6  6.00  1.65 4.00  5.50    41  1.66  1.61  4.05  3.84  6.50  5.98   36
##   BbMx.2.5 BbAv.2.5 BbMx.2.5.1 BbAv.2.5.1 BbAH BbAHh BbMxAHH BbAvAHH BbMxAHA BbAvAHA  PSCH PSCD
## 1     1.65     1.61       2.43       2.32   21 -1.00    1.91    1.85    2.10    2.02  1.49 4.73
## 2     1.70     1.63       2.40       2.27   20  1.50    1.95    1.91    2.01    1.96 11.75 6.15
## 3     1.71     1.66       2.33       2.23   20 -1.75    2.03    1.97    1.95    1.90  1.33 5.40
## 4     2.19     2.11       1.79       1.72   18 -0.75    2.10    2.05    1.86    1.83  1.79 3.56
## 5     2.17     2.08       1.80       1.76   19 -0.75    1.94    1.90    2.01    1.98  1.82 3.49
## 6     2.17     2.08       1.80       1.75   19 -0.75    1.83    1.78    2.16    2.10  1.56 4.25
##    PSCA
## 1  7.25
## 2  1.29
## 3 12.25
## 4  5.51
## 5  5.42
## 6  6.85
```

```
str(data)
```

```
## 'data.frame':    380 obs. of  65 variables:
##  $ Div       : chr  "E0" "E0" "E0" "E0" ...
##  $ Date      : chr  "11/08/17" "12/08/17" "12/08/17" "12/08/17" ...
##  $ HomeTeam  : chr  "Arsenal" "Brighton" "Chelsea" "Crystal Palace" ...
##  $ AwayTeam  : chr  "Leicester" "Man City" "Burnley" "Huddersfield" ...
##  $ FTHG      : int  4 0 2 0 1 0 3 1 4 0 ...
##  $ FTAG      : int  3 2 3 3 0 0 3 0 0 2 ...
##  $ FTR       : chr  "H" "A" "A" "A" ...
##  $ HTHG      : int  2 0 0 0 1 0 2 1 1 0 ...
##  $ HTAG      : int  2 0 3 2 0 0 1 0 0 0 ...
##  $ HTR       : chr  "D" "D" "A" "A" ...
##  $ Referee   : chr  "M Dean" "M Oliver" "C Pawson" "J Moss" ...
##  $ HS        : int  27 6 19 14 9 29 9 16 22 6 ...
##  $ AS        : int  6 14 10 8 9 4 14 9 9 18 ...
##  $ HST       : int  10 2 6 4 4 2 4 6 6 3 ...
##  $ AST       : int  3 4 5 6 1 0 5 2 1 6 ...
##  $ HF        : int  9 6 16 7 13 10 14 15 19 6 ...
##  $ AF        : int  12 9 11 19 10 13 8 3 7 10 ...
##  $ HC        : int  9 3 8 12 6 13 3 8 11 5 ...
##  $ AC        : int  4 10 5 9 7 0 3 2 1 7 ...
##  $ HY        : int  0 0 3 1 1 2 0 3 2 1 ...
##  $ AY        : int  1 2 3 3 1 1 3 1 2 2 ...
##  $ HR        : int  0 0 2 0 0 0 0 0 0 1 ...
##  $ AR        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ B365H     : num  1.53 11 1.25 1.83 1.7 1.62 6 2.4 1.3 5.5 ...
##  $ B365D     : num  4.5 5.5 6.5 3.6 3.8 4 4.2 3.3 5.75 4 ...
##  $ B365A     : num  6.5 1.33 15 5 5.75 6.5 1.62 3.3 12 1.7 ...
##  $ BWH       : num  1.5 11 1.22 1.8 1.7 1.57 6 2.4 1.28 5.25 ...
##  $ BWD       : num  4.6 5.25 6.5 3.5 3.6 4 4.2 3.2 5.5 3.8 ...
##  $ BWA       : num  6.75 1.3 12.5 4.75 5.5 6 1.55 3.1 11 1.67 ...
##  $ IWH       : num  1.47 8 1.22 1.85 1.7 1.65 5.5 2.3 1.33 4.8 ...
##  $ IWD       : num  4.5 5.3 6.2 3.5 3.7 3.8 4 3.3 5.3 3.6 ...
##  $ IWA       : num  6.5 1.35 13.5 4.3 5 5.3 1.6 3.15 8.7 1.75 ...
##  $ LBH       : num  1.44 10 1.25 1.8 1.67 1.6 5.8 2.4 1.33 5 ...
##  $ LBD       : num  4.4 5 5.75 3.4 3.6 3.7 4 3.1 5 3.8 ...
##  $ LBA       : num  6.5 1.3 15 4.6 5.25 6 1.57 3 10 1.67 ...
##  $ PSH       : num  1.53 10.95 1.26 1.83 1.7 ...
##  $ PSD       : num  4.55 5.55 6.3 3.58 3.83 3.94 4.29 3.25 5.68 4 ...
##  $ PSA       : num  6.85 1.34 15.25 5.11 5.81 ...
##  $ WHH       : num  1.53 10 1.25 1.8 1.7 1.62 5.5 2.5 1.3 5 ...
##  $ WHD       : num  4.2 4.8 5.5 3.3 3.5 3.6 3.8 3.1 5 3.75 ...
##  $ WHA       : num  6 1.33 13 5 5.5 6 1.62 3 11 1.7 ...
##  $ VCH       : num  1.53 10 1.25 1.83 1.7 1.65 6 2.5 1.3 5.5 ...
##  $ VCD       : num  4.5 5.5 6.25 3.6 3.8 4 4 3.3 5.5 4 ...
##  $ VCA       : num  6.5 1.33 15 5 5.75 5.5 1.65 3.13 11.5 1.7 ...
##  $ Bb1X2     : int  41 40 41 41 40 41 41 41 40 41 ...
##  $ BbMxH     : num  1.55 11.5 1.27 1.86 1.71 1.66 6.5 2.5 1.35 5.5 ...
##  $ BbAvH     : num  1.51 10.1 1.24 1.81 1.69 1.61 5.75 2.43 1.31 5.16 ...
##  $ BbMxD     : num  4.6 5.6 6.55 3.65 3.85 4.05 4.3 3.3 5.75 4 ...
##  $ BbAvD     : num  4.43 5.25 6.06 3.5 3.69 3.84 4.06 3.19 5.29 3.87 ...
##  $ BbMxA     : num  6.89 1.36 15.5 5.11 6 6.5 1.65 3.3 13 1.75 ...
##  $ BbAvA     : num  6.44 1.32 13.67 4.82 5.5 ...
##  $ BbOU      : int  37 35 36 36 35 36 37 36 35 36 ...
##  $ BbMx.2.5  : num  1.65 1.7 1.71 2.19 2.17 2.17 1.89 2.25 1.76 1.88 ...
##  $ BbAv.2.5  : num  1.61 1.63 1.66 2.11 2.08 2.08 1.82 2.16 1.7 1.81 ...
##  $ BbMx.2.5.1: num  2.43 2.4 2.33 1.79 1.8 1.8 2.08 1.75 2.25 2.07 ...
##  $ BbAv.2.5.1: num  2.32 2.27 2.23 1.72 1.76 1.75 1.99 1.7 2.16 2.01 ...
##  $ BbAH      : int  21 20 20 18 19 19 21 22 19 20 ...
##  $ BbAHh     : num  -1 1.5 -1.75 -0.75 -0.75 -0.75 1 -0.25 -1.5 0.75 ...
##  $ BbMxAHH   : num  1.91 1.95 2.03 2.1 1.94 1.83 1.9 2.12 2.01 2.06 ...
##  $ BbAvAHH   : num  1.85 1.91 1.97 2.05 1.9 1.78 1.84 2.08 1.96 2 ...
##  $ BbMxAHA   : num  2.1 2.01 1.95 1.86 2.01 2.16 2.13 1.85 1.95 1.92 ...
##  $ BbAvAHA   : num  2.02 1.96 1.9 1.83 1.98 2.1 2.04 1.81 1.92 1.87 ...
##  $ PSCH      : num  1.49 11.75 1.33 1.79 1.82 ...
##  $ PSCD      : num  4.73 6.15 5.4 3.56 3.49 4.25 4.27 3.21 5.79 3.9 ...
##  $ PSCA      : num  7.25 1.29 12.25 5.51 5.42 ...
```

Description of the data : The data corresponds to all the games of the English Premier League during the 2017-2018 season. It describes different data of each games : The division is always the same as it is premier league(E0), the date of each game, the name of the home and the away teams with their respective goals score and the result of the game. FTHG= Full Time Home Goals : the number of goals scored by the home team at the end of the game. FTAG = Full Time Away Goals : the number of goals scored by the away team at the end of the game. FTR = Full Time Results : the result at the end of the game. Then there are other data for each game such as the half time number of goals, the referee name, the number of strikes, strikes on target, fouls, corners, yellow and red card corresponding to each team. There are at the end a lot of bets data with the odds of betting for home team, draw or away from different betting websites. The max and the average odds for each type of results.

The dataset is composed of 7 characters columns, 19 integer columns, 39 float columns.

QUESTION 2 :

Compute the number of points over the season of each team (victory = 3 points, draw= 1 point), the number of points in "home" matches, the number of points in "away" matches.

```
Home_Points <- ifelse(data$FTR == "H" , 3 , ifelse(data$FTR == "D", 1, 0))

Away_Points <- ifelse(data$FTR == "A", 3, ifelse(data$FTR == "D", 1 , 0))

data <- cbind(data, Home_Points, Away_Points)
```

```
home_points <- aggregate(Home_Points ~ HomeTeam, data = data, FUN = sum)
names(home_points)[names(home_points) == "Home_Points"] <- "TotalHomePoints"

away_points <- aggregate(Away_Points ~ AwayTeam, data = data, FUN = sum)
names(away_points)[names(away_points) == "Away_Points"] <- "TotalAwayPoints"

total_points <- merge(home_points, away_points, by.x = "HomeTeam", by.y = "AwayTeam")

total_points$TotalPoints <- total_points$TotalHomePoints + total_points$TotalAwayPoints

summary(total_points)
```

```
##    HomeTeam         TotalHomePoints TotalAwayPoints  TotalPoints
##  Length:20          Min.   :18.00   Min.   :11.00   Min.   : 31.00
##  Class :character   1st Qu.:25.25   1st Qu.:14.00   1st Qu.: 39.25
##  Mode  :character   Median :27.00   Median :16.50   Median : 44.00
##                     Mean   :30.90   Mean   :21.15   Mean   : 52.05
##                     3rd Qu.:38.50   3rd Qu.:29.00   3rd Qu.: 64.75
##                     Max.   :50.00   Max.   :50.00   Max.   :100.00
```

```
total_points_sorted <- total_points[order(-total_points$TotalPoints),]
print(total_points_sorted)
```

```
##          HomeTeam TotalHomePoints TotalAwayPoints TotalPoints
## 11       Man City              50              50         100
## 12     Man United              47              34          81
## 17      Tottenham              43              34          77
## 10      Liverpool              43              32          75
## 5         Chelsea              37              33          70
## 1         Arsenal                    47               16          63
## 4         Burnley                    26               28          54
## 7         Everton                    34               15          49
## 9       Leicester                    27               20          47
## 2     Bournemouth                    26               18          44
## 6  Crystal Palace                    26               18          44
## 13      Newcastle                    28               16          44
## 20       West Ham                    27               15          42
## 18        Watford                    27               14          41
## 3        Brighton                    29               11          40
## 8    Huddersfield                    23               14          37
## 14    Southampton                    19               17          36
## 15          Stoke                    20               13          33
## 16        Swansea                    21               12          33
## 19      West Brom                    18               13          31
```

How many points did Arsenal score

```r
arsenal_points <- total_points_sorted[total_points_sorted$HomeTeam == "Arsenal",]$TotalPoints
print(paste("Arsenal's total points: ", arsenal_points))
```

```
## [1] "Arsenal's total points:  63"
```

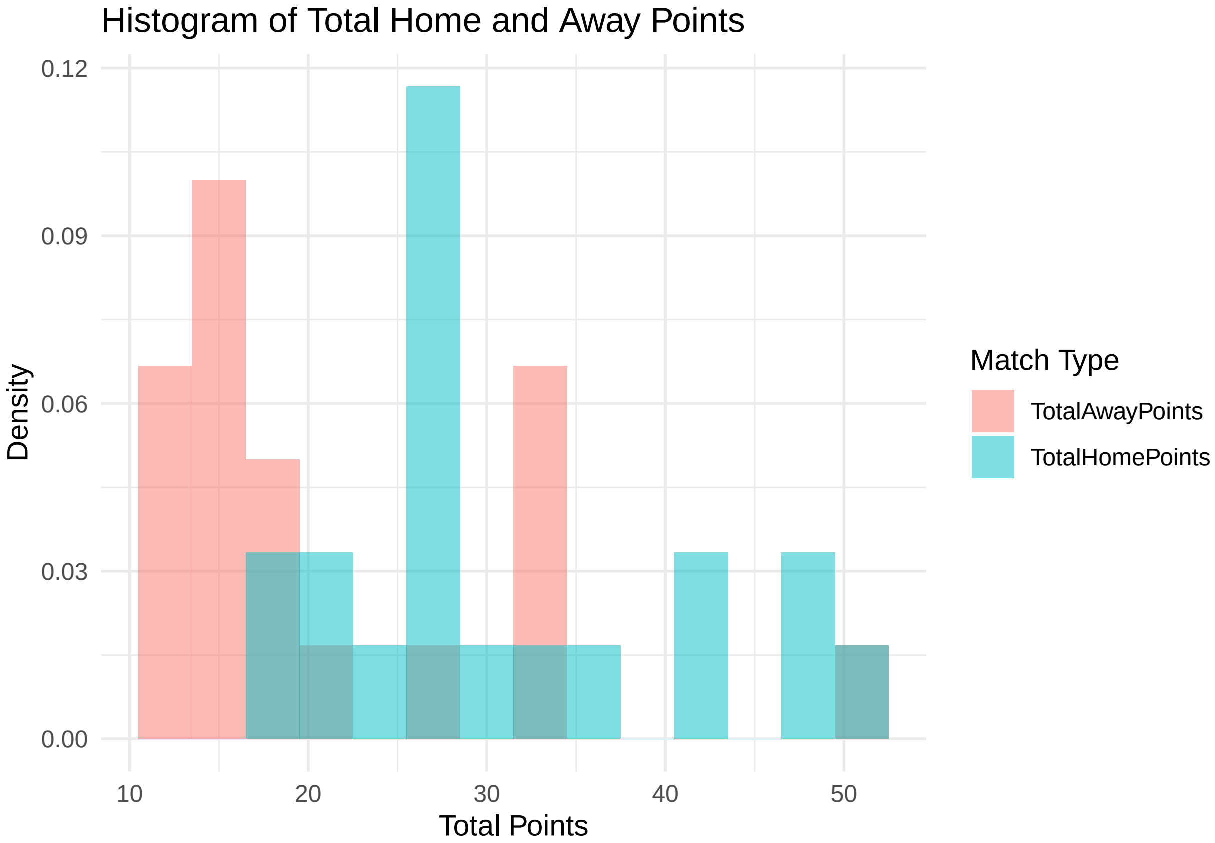
What was Arsenal's rank

```r
arsenal_rank <- which(total_points_sorted$HomeTeam == "Arsenal")
print(paste("Arsenal's rank: ", arsenal_rank))
```

```
## [1] "Arsenal's rank:  6"
```

Compare the histogram of the total number of points at home and away

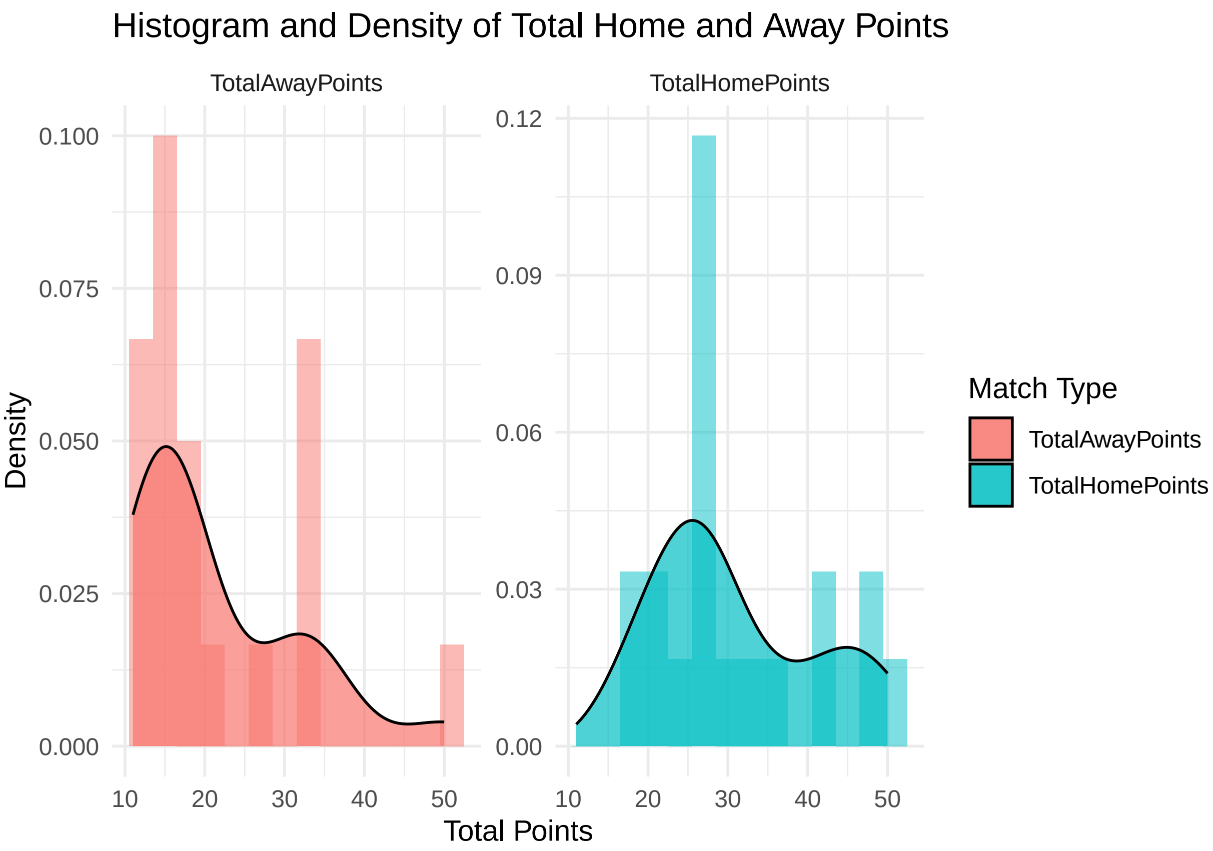
```r
library(tidyr)
library(ggplot2)
total_points_long <- gather(total_points, key = "Type", value = "Points", TotalHomePoints, TotalAway

ggplot(total_points_long, aes(x = Points, fill = Type)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 3, alpha = 0.5, position = 'identity') +
  labs(x = "Total Points", y = "Density", fill = "Match Type") +
  ggtitle("Histogram of Total Home and Away Points") +
  theme_minimal()
```



Comparing the histogram of the total number of points at home and away, we can see that 8 of 20 teams took between 25 and 30 points during home games, then there are 5 teams under 25 points and 7 teams over 30 points with 3 teams that took between 45 and 50 points at home. On the other side, most of the teams took less than 20 points playing away games with 14 teams between 10 and 20 points. 5 teams are around 30 points and only 1 teams at 50 points (not surprising Man City). We can well see the impact of playing at home or away with these two histograms.

Fit a density to those histograms.

```
# Convert data to long format for ggplot
library(tidyr)
library(ggplot2)
total_points_long <- gather(total_points, key = "Type", value = "Points", TotalHomePoints, TotalAway

ggplot(total_points_long, aes(x = Points, fill = Type)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 3, alpha = 0.5, position = 'identity') +
  geom_density(alpha = 0.7, adjust = 1) +
  facet_wrap(~Type, scales = 'free_y') +
  labs(x = "Total Points", y = "Density", fill = "Match Type") +
  ggtitle("Histogram and Density of Total Home and Away Points") +
  theme_minimal()
```



QUESTION 3

Write the statistical model associated to the observation of n match results. Do you think it is a realistic model ?

Our statistical model is described by the set or probability distribution $\{P_\theta, \theta \in \Omega\}$ where $\Omega \in \mathbb{R}_+^*$ represents the parameter space and $\{P_\theta\}$ the Poisson distribution. Let $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ be our n observations of the $X_1, \ldots, X_n$ (FTHG) and $Y_1, \ldots, Y_n$ (FTAG) random variables. For any $i = 1, \ldots, n$, X_i follows a Poisson(λ) and Y_i follows a Poisson(μ).

Model Realistic ? The Poisson law is a discrete distribution that describes the behavior of the number of events that happen within a certain amount of time. For that reason, this model is realistic since we are looking at the number of goals within 90min. However, we should maybe focus on a specific team because each team has its own level and its own probability to score.

In order to estimate λ and μ we can use the MLE estimator as seen in Part 1.

QUESTION 4

```r
# to execute just once
attach(data)
```

Compute the empirical mean and variance of the number of goals of 1)the visiting team 2) the home team. Compute the MLE estimators (of ,) for the Poisson model.

```r
# empirical mean
empirical_mean_goals_home_team <- mean(FTHG)
empirical_mean_goals_away_team <- mean(FTAG)

# variance
variance_goals_home_team <- var(FTHG)
variance_goals_away_team <- var(FTAG)

# MLE Estimators
```

```
lambda_hat <- empirical_mean_goals_home_team
mu_hat <- empirical_mean_goals_away_team

cat("empircal mean for goals of home teams :", empirical_mean_goals_home_team, "\n")
```

```
## empircal mean for goals of home teams : 1.531579
```

```
cat("empircal mean for goals of away teams :", empirical_mean_goals_away_team, "\n\n")
```

```
## empircal mean for goals of away teams : 1.147368
```

```
cat("variance for goals of home teams :", variance_goals_home_team, "\n")
```

```
## variance for goals of home teams : 1.795834
```

```
cat("variance for goals of away teams :", variance_goals_away_team, "\n\n")
```

```
## variance for goals of away teams : 1.387196
```

```
cat("MLE estimator lambda_hat of the Poisson model for FTHG :", lambda_hat, "\n")
```

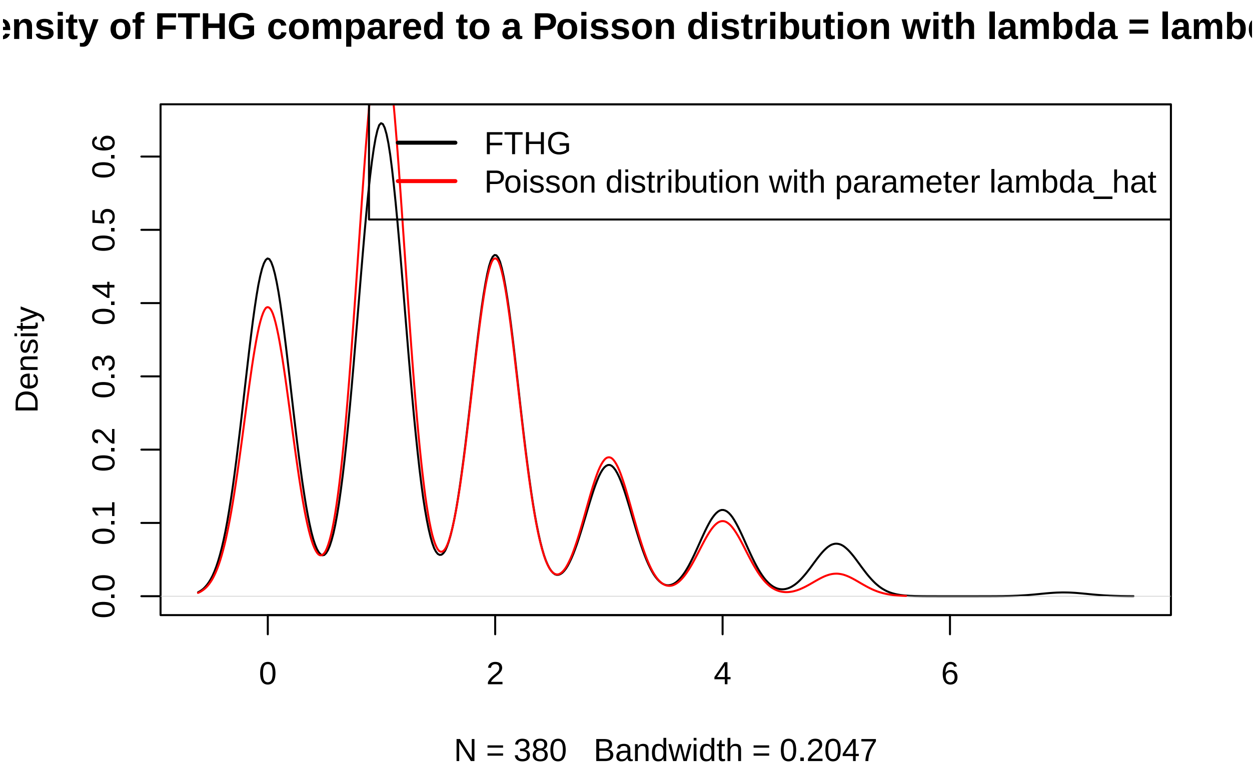
```
## MLE estimator lambda_hat of the Poisson model for FTHG : 1.531579
```

```
cat("MLE estimator mu_hat of the Poisson model for FTAG :", mu_hat, "\n\n")
```

```
## MLE estimator mu_hat of the Poisson model for FTAG : 1.147368
```

Does the Poisson assumption look correct ?

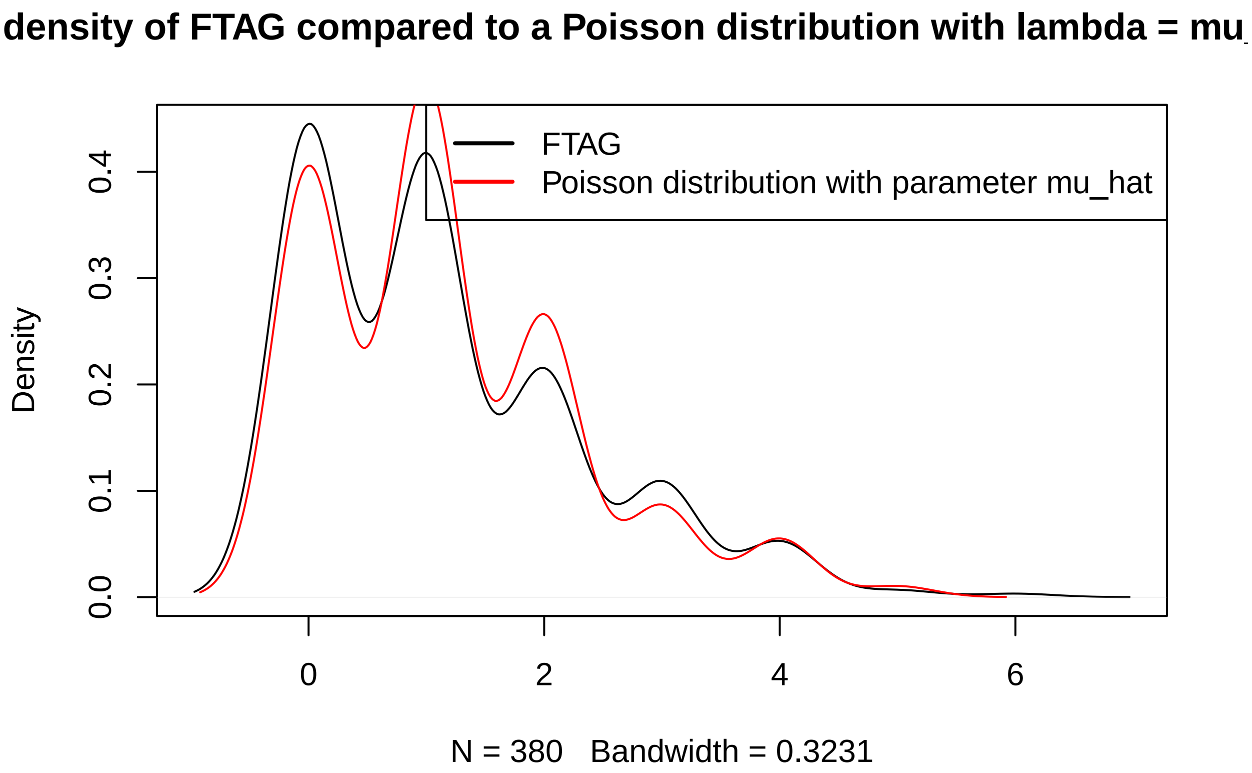
```
# We compare FTHG and the estimated Poisson model : Pois(lambda_hat)

# comparison on the density
simulated_data <- rpois(length(FTHG), lambda_hat)
plot(density(FTHG), main = "Density of FTHG compared to a Poisson distribution with lambda = lambda_
lines(density(simulated_data), col='red')
legend("topright", legend = c("FTHG", "Poisson distribution with parameter lambda_hat"), col = c("bl
```



As we see above, the density of our data FTHG is similar to a poisson distribution with parameter

$$\lambda = \hat{\lambda}$$

```
# We compare FTAG and the estimated Poisson model : Pois(mu_hat)

# comparison on the density
simulated_data <- rpois(length(FTAG), mu_hat)
plot(density(FTAG), main = "density of FTAG compared to a Poisson distribution with lambda = mu_hat'
lines(density(simulated_data), col='red')
legend("topright", legend = c("FTAG", "Poisson distribution with parameter mu_hat"), col = c("black'
```



As we see above, the density of our data FTAG is similar to a poisson distribution with parameter

$$\lambda = \hat{\mu}$$

QUESTION 5

Compute the confidence intervals for derived at question 7 (part 1), and a similar confidence interval for μ.

```
conf_int_student <- function(x, alpha){
  n <- length(x)
  q_alpha <- qnorm(1-alpha/2)
  xbar = mean(x)
  Sn <- xbar/n
  CI <- c(xbar - sqrt(Sn)*q_alpha, xbar + sqrt(Sn)*q_alpha)
  return(CI)
}

cat("confidence interval at a 95 level for lambda : ", conf_int_student(FTHG, 0.05),  "\n")
```

```
## confidence interval at a 95 level for lambda :  1.407149 1.656009
```

```
cat("confidence interval at a 95 level for mu : ", conf_int_student(FTAG, 0.05),  "\n")
```

```
## confidence interval at a 95 level for mu :  1.03967 1.255066
```

Do you think the distribution of the number of goals scored by the home team and the visiting team is the same ?

The distribution of the number of goals scored by the home team and the visiting team are clearly not the same since the intersection of the confidence intervals for their parameters lambda and mu is null.

QUESTION 6

What would be the best approach to answer the previous question ? Formalize the problem as a testing problem.

To formally test whether the distribution of the number of goals scored by the home team is the same as that of the visiting team, you can set up a hypothesis testing framework. Here's how we can set up the testing problem :

Hypotheses :

Null Hypothesis (H_0) : The mean goal rate for the home team (λ) is equal to the mean goal rate for the visiting team (μ).

Alternative Hypothesis (H_1) : The mean goal rate for the home team (λ) is not equal to the mean goal rate for the visiting team (μ).

$$H_0 : \lambda = \mu \quad \text{vs} \quad H_1 : \lambda \neq \mu$$

Therefore, the best approach to answer the previous question is to do a two-sample t-test, which is a method used to test whether the unknown population means of two groups are equal or not.

Use a t.test to give a more precise answer.

```
t_test_result <- t.test(FTHG, FTAG)

# Print the results
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  FTHG and FTAG
## t = 4.198, df = 745.71, p-value = 3.018e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2045378 0.5638832
## sample estimates:
## mean of x mean of y
##  1.531579  1.147368
```

Comment on the assumptions of such a test. Are they valid, "nearly valid", or problematic ?

The Two-sample t-test decides if the population means for two different groups are equal or not. The Two-sample t-test assumptions are :

- Homogeneity of variance. The distribution is approximately normal.
- Independence of the samples.

```
var(data$FTHG)
```

```
## [1] 1.795834
```

```
var(data$FTAG)
```

```
## [1] 1.387196
```

```r
shapiro.test(data$FTHG)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$FTHG
## W = 0.8753, p-value < 2.2e-16
```

```r
shapiro.test(data$FTAG)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$FTAG
## W = 0.83896, p-value < 2.2e-16
```

We see that the two samples have similar variance, we can consider that theso the first assumption on the homogeneity of variancesis "valid" or "nearly valid." We can see with the Shapiro-Wilk test, both p-value are very small (smaller than 2.2e-16), this suggests that the data significantly deviates from a normal distribution. We can thus say that this assumption is problematic. However, the t-test tends to be robust against moderate deviations from normality when sample sizes are big enough. The independance is valid as in football games, the goals scored in different matches are independent of each other.

What is the p-value of the test ? What does that mean ?

The p-value of the t-test is 3.018e-05. It means that the probability to observe under HO a more extreme value of the statistic (used in the t-test)

If you want a test of level = 0.05, do you accept or reject the null hypothesis ?

Here the p-value is very small (3.018×10^{-5}). Here, the p-value is such that $p < \alpha$ (for $\alpha = 0.05$). This indicates that the mean goal rates for the home and visiting teams are significantly different and that we should reject H_0. It suggests that the distributions of the number of goals scored by the home team and the visiting team are not the same.

QUESTION 7

Create two vectors ManCity and Liverpool containing the goals scored during the season, both away and home.

```r
ManCity <- c(data$FTHG[data$HomeTeam == "Man City"], data$FTAG[data$AwayTeam == "Man City"])
Liverpool <- c(data$FTHG[data$HomeTeam == "Liverpool"], data$FTAG[data$AwayTeam == "Liverpool"])
```

Formalise the previous question as a testing problem and use a t.test to answer it.

Null Hypothesis ($\mathbf{H_0}$) : The mean goal rate for Manchester City (μ_{ManCity}) is equal to the mean goal rate for Liverpool ($\mu_{\text{Liverpool}}$).

Alternative Hypothesis ($\mathbf{H_1}$) : The mean goal rate for Manchester City (μ_{ManCity}) is not equal to the mean goal rate for Liverpool ($\mu_{\text{Liverpool}}$).

$$H_0 : \mu_{\text{ManCity}} = \mu_{\text{Liverpool}} \quad \text{vs} \quad H_1 : \mu_{\text{ManCity}} \neq \mu_{\text{Liverpool}}$$

```r
# Perform the t-test
t_test_result <- t.test(ManCity, Liverpool)
t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  ManCity and Liverpool
## t = 1.5891, df = 73.992, p-value = 0.1163
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1470057  1.3049004
## sample estimates:
## mean of x mean of y
##  2.789474  2.210526
```

We set a significance level ($\alpha = 0.05$) to decide whether to reject the null hypothesis. In our case, the p-value is greater than 0.05, leading us to fail to reject the null hypothesis. This means we don't have strong evidences to claim a difference in the mean number of goals between Manchester City and Liverpool. However, this season Manchester city has scored 106 goals in total while Liverpool scored 84. Both number of goals is high but 22 goals difference in 38 games is still a big difference. Is it enough to say that Man City has the best offence ? Probably not for the t-test but this season they were unplayable and this 17-18 Man City is the record breaker of the most goals scored by a team during a Premier League season.