MAP531 Homework - Charles Cazals & Thomas de Mareuil

Problem 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 pdf given by $(p(\theta, k), k \in \mathbb{N})$ w.r.t the counting measure on $\mathbb{N}$:

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

Question 1

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

It's a discrete distribution, as it is defined for $k \in \mathbb{N}$.

We could use Poisson for the following predictions:

- The number of birthdays among all your Facebook friends on a given day
- The number of spelling mistakes on a given page in a published novel
- The number of car-accident deaths in a given day in France.

Question 2

Compute the mean and the variance of this distribution as a function of λ.

Remark: that if X_1 and X_2 are two independent random variables following a Poisson distribution with respective parameters $\lambda_1 > 0$ and $\lambda_2 > 0$, then $X_1 + X_2$ has a Poisson distribution of parameter $\lambda_1 + \lambda_2$. You do not need to prove this result.

Mean:

$$\begin{aligned}
\mu &= \sum_{k\in\mathbb{N}} k \times p(k, \lambda) \\
&= \sum_{k=0}^{\infty} k \times e^{-\lambda}\frac{\lambda^k}{k!} \\
&= e^{-\lambda}\sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \\
&= \lambda e^{-\lambda}\sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \qquad \text{with } j = k-1 \\
&= \lambda e^{-\lambda}e^{\lambda} \\
&\boxed{\mu = \lambda}
\end{aligned}$$

Variance:

$$
\begin{aligned}
\sigma^2 &= \sum_{k\in\mathbb{N}} (k-\mu)^2 \times p(k,\lambda) \\
&= \sum_{k=0}^{\infty} (k-\lambda)^2 \times e^{-\lambda}\frac{\lambda^k}{k!} \\
&= e^{-\lambda}\Big(\lambda^2\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} - 2\lambda\sum_{k=0}^{\infty}k\frac{\lambda^k}{k!} + \sum_{k=0}^{\infty}k^2\frac{\lambda^k}{k!}\Big) \\
&= e^{-\lambda}\Big(\lambda^2 e^{\lambda} - 2\lambda\times\lambda e^{\lambda} + \sum_{k=1}^{\infty}\frac{k\lambda^k}{(k-1)!}\Big) \\
&= -\lambda^2 + e^{-\lambda}\sum_{j=0}^{\infty}\frac{(j+1)\lambda\lambda^j}{j!} && \text{with } j=k-1 \\
&= -\lambda^2 + e^{-\lambda}\lambda\Big(\sum_{j=1}^{\infty}\frac{\lambda^j}{(j-1)!} + \sum_{j=0}^{\infty}\frac{\lambda^j}{j!}\Big) \\
&= -\lambda^2 + e^{-\lambda}\lambda\Big(\lambda\sum_{i=0}^{\infty}\frac{\lambda^i}{i!} + e^{\lambda}\Big) && \text{with } i=j-1 \\
&= -\lambda^2 + e^{-\lambda}\lambda\big(\lambda e^{\lambda} + e^{\lambda}\big) \\
&= -\lambda^2 + \lambda^2 + \lambda \\
&\boxed{\sigma^2 = \lambda}
\end{aligned}
$$

We are provided with n independent observations of a Poisson random variable of parameter $\theta \in \Theta = \mathbb{R}_+^*$.

Question 3

- **What are our observations? What distribution do they follow?**

The observations represent the number of occurences of a phenomenon for a given period, which appears on average θ times per period (with variance θ) following a Poisson distribution.

- **Write the corresponding statistical model.**

The statistical model is $\{p(\theta,k), k\in\mathbb{N}\}$, with $\forall i\in\{1,100\},\ \forall k\in\mathbb{N},\ p(x_i=k)=e^{-\theta}\frac{\theta^k}{k!}$.

- **What parameter are we trying to estimate?**

We are trying to estimate the parameter θ.

Question 4

- **What is the likelihood function?**

$$
\begin{aligned}
l(\theta) &= \mathbb{P}(x_1,...,x_n|\theta) \\
&= \prod_{i=1}^{n}\mathbb{P}(x_i|\theta) && \text{by independence} \\
&= \prod_{i=1}^{n} e^{-\theta}\frac{\theta^{x_i}}{x_i!} && \text{identically distributed}
\end{aligned}
$$

- **Compute the Maximum Likelihood Estimator $\hat{\theta}_{ML}$.**

To compute the MLE we will maximize the log-likelihood function:

$$\begin{aligned}
L(\theta) = log(l(\theta)) &= \sum_{i=1}^{n} log\big(e^{-\theta}\frac{\theta^{x_i}}{x_i!}\big) \\
&= \sum_{i=1}^{n} \big(-\theta + x_i log(\theta) - log(x_i!)\big) \\
&= -n\theta + log(\theta)\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} log(x_i!)
\end{aligned}$$

First-order condition:

$\frac{\partial L(\theta)}{\partial \theta} = -n + \frac{1}{\hat{\theta}_{ML}}\sum_{i=1}^{n} x_i = 0$, i.e. $\boxed{\hat{\theta}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}}$.

Question 5

Prove that $\sqrt{n}(\hat{\theta}_{ML} - \theta)$ converges in distribution as $n \to \infty$.

By Central Limit Theorem, we know that: $\sqrt{n}(\bar{x} - \mu) \xrightarrow{d} \mathcal{N}(0, \theta)$ as $n \to \infty$.

Since $\hat{\theta}_{ML} = \bar{x}$ and $\mu = \sigma^2 = \theta$, we have: $\boxed{\sqrt{n}(\hat{\theta}_{ML} - \theta) \xrightarrow{d} \mathcal{N}(0, \theta)}$ as $n \to \infty$.

Question 6

- **Prove that $\sqrt{n}\frac{(\hat{\theta}_{ML} - \theta)}{\sqrt{\hat{\theta}_{ML}}}$ converges in distribution as $n \to \infty$.**

By Law of Large Numbers, we know that: $\hat{\theta}_{ML} = \bar{x} \xrightarrow{a.s.} \mu = \theta$ as $n \to \infty$.

By Continuous Mapping Theorem, we thus have: $\sqrt{\hat{\theta}_{ML}} \xrightarrow{a.s.} \sqrt{\theta}$ as $n \to \infty$, since $\theta \in \mathbb{R}_+^*$ and $\sqrt{.}$ is continuous on $\mathbb{R}_+^*$.

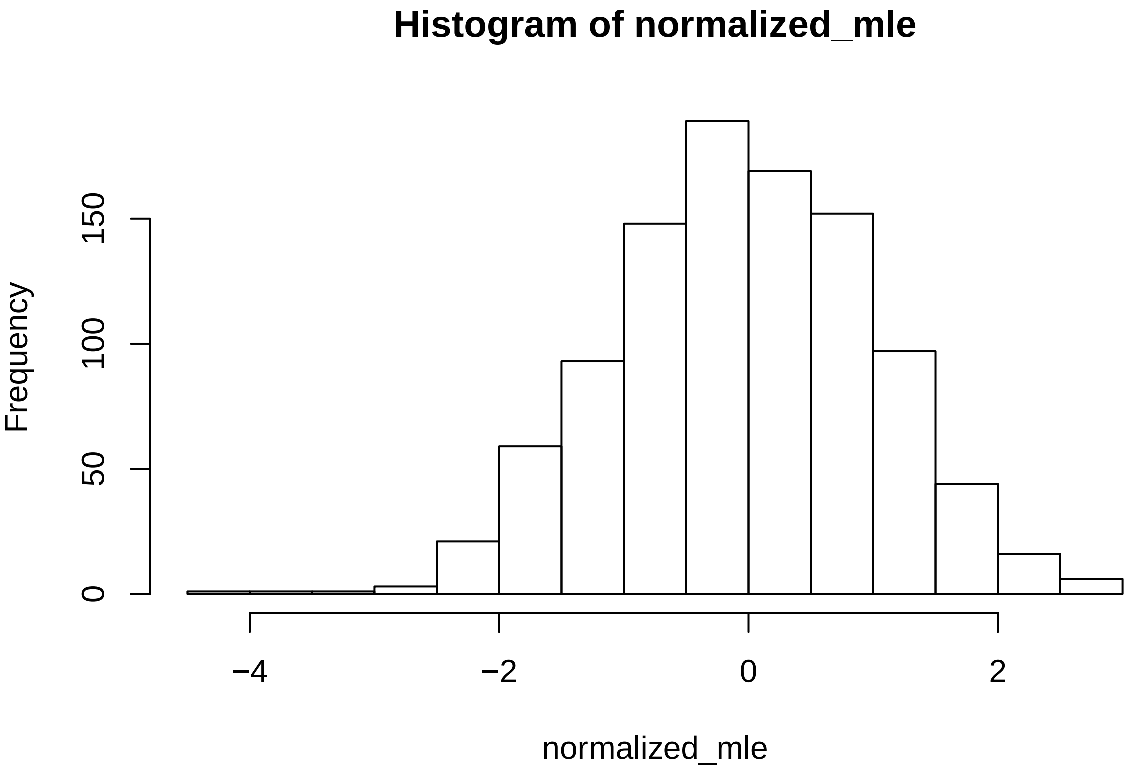
Finally, using Slutsky's lemma, we have:

$\sqrt{n}(\hat{\theta}_{ML} - \theta) \xrightarrow{d} \mathcal{N}(0, \theta)$ and $\sqrt{\hat{\theta}_{ML}} \xrightarrow{a.s.} \sqrt{\theta}$ constant, so $\boxed{\frac{\sqrt{n}(\hat{\theta}_{ML} - \theta)}{\sqrt{\hat{\theta}_{ML}}} \xrightarrow{d} \mathcal{N}(0, 1)}$.

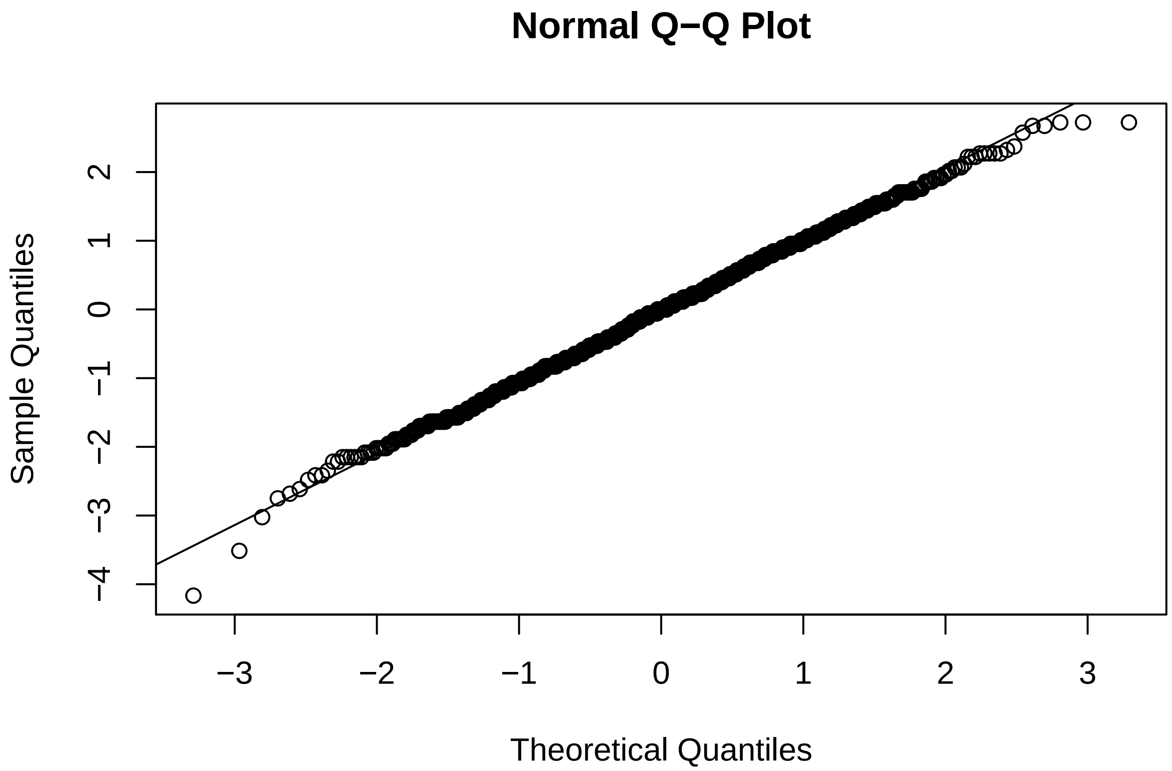
- **On R, verify that the distribution of the random variable $\sqrt{n}\frac{(\hat{\theta}_{ML} - \theta)}{\sqrt{\theta_{ML}}}$ is what you found theoretically, through a histogram and a QQ-plot (compute $N_{attempts} = 1000$ times the random variable $\sqrt{n}_{sample}\frac{(\hat{\theta}_{ML} - \theta)}{\sqrt{\theta_{ML}}}$ from a sample of size n_{sample} of simulated Poisson data, with $\theta = 3$, like in PC2).**

```
Nattempts <- 1000
nsample <- 100
theta <- 3

theta_mle <- c()
normalized_mle <- c()
```

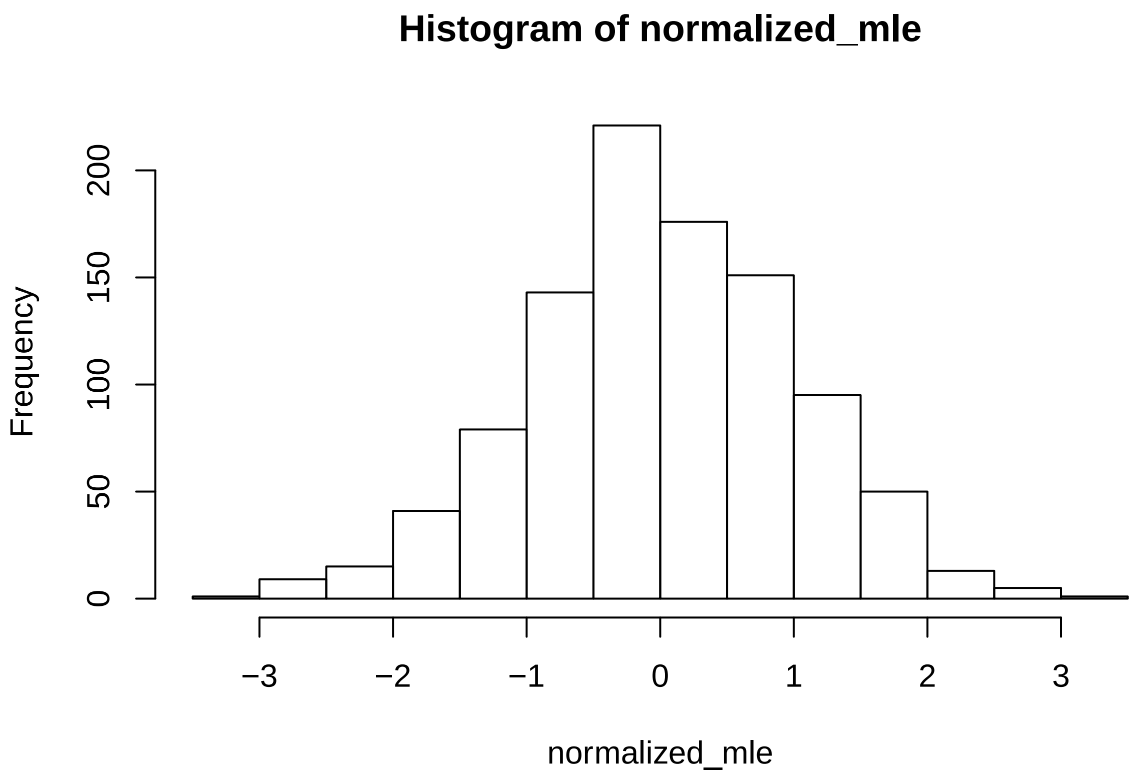
```r
for (i in 1:Nattempts) {
  poisson_sample <- rpois(nsample, theta)
  theta_mle[i] <- mean(poisson_sample)
  normalized_mle[i] <- sqrt(nsample)*(theta_mle[i]-theta)/sqrt(theta_mle[i]) }

hist(normalized_mle)
```



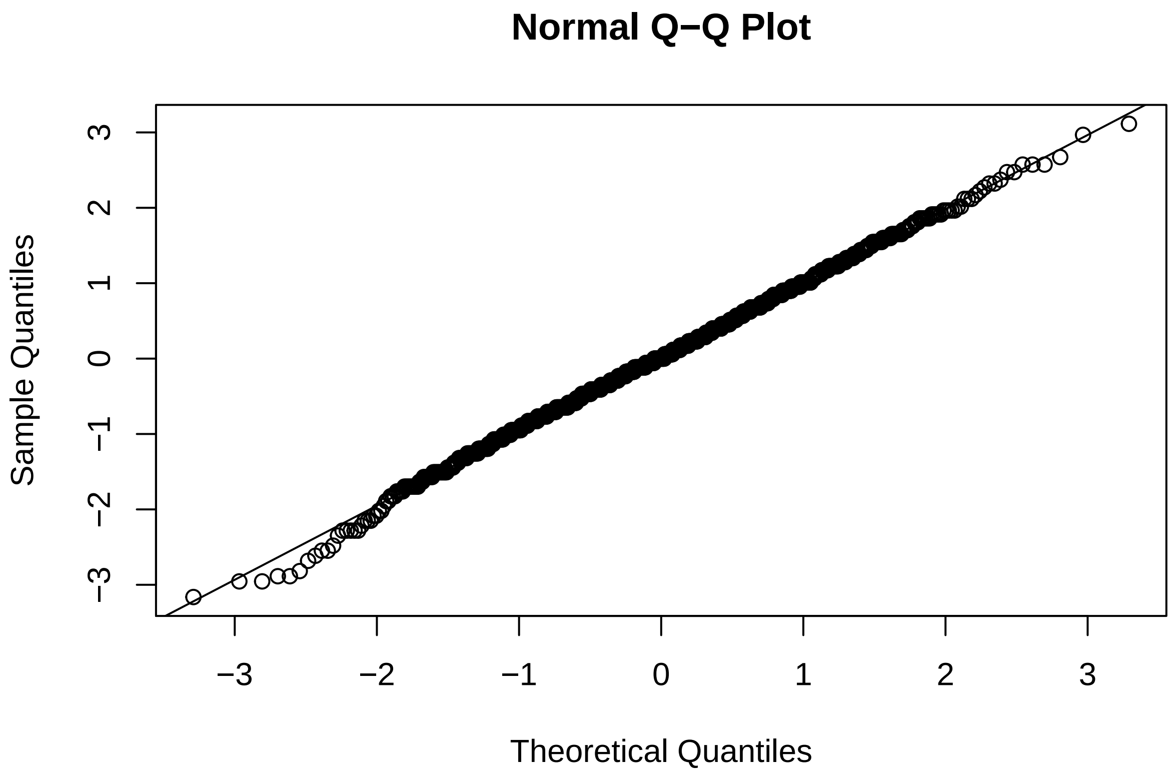
```r
qqnorm(normalized_mle)
qqline(normalized_mle)
```



```r
# Same thing without a for loop

theta_mle <- replicate(Nattempts, mean(rpois(nsample, theta)))
normalized_mle <- sqrt(nsample) * (theta_mle - theta) / sqrt(theta_mle)

hist(normalized_mle)
```



```r
qqnorm(normalized_mle)
qqline(normalized_mle)
```



Question 7

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

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

As $n \to \infty$, we have seen that $\frac{\sqrt{n}(\hat{\theta}_{ML}-\theta)}{\sqrt{\theta_{ML}}} \xrightarrow{d} \mathcal{N}(0,1)$. Therefore:

$$\lim_{n\to\infty} \mathbb{P}\Big(z_{\frac{\alpha}{2}} \leq \sqrt{n}\frac{\hat{\theta}_{ML}-\theta}{\sqrt{\theta_{ML}}} \leq z_{1-\frac{\alpha}{2}}\Big) = 1-\alpha$$

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{\sqrt{\hat{\theta}_{ML}}\cdot z_{\frac{\alpha}{2}}}{\sqrt{n}} - \hat{\theta}_{ML} \leq -\theta \leq \frac{\sqrt{\hat{\theta}_{ML}}\cdot z_{1-\frac{\alpha}{2}}}{\sqrt{n}} - \hat{\theta}_{ML}\Big) = 1-\alpha$$

$$\lim_{n\to\infty} \mathbb{P}\Big(\hat{\theta}_{ML} - \frac{\sqrt{\hat{\theta}_{ML}}\cdot z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \leq \theta \leq \hat{\theta}_{ML} - \frac{\sqrt{\hat{\theta}_{ML}}\cdot z_{\frac{\alpha}{2}}}{\sqrt{n}}\Big) = 1-\alpha$$

Since $\hat{\theta}_{ML} = \bar{x}$ and $z_{\frac{\alpha}{2}} = z_{1-\frac{\alpha}{2}}$ by symmetry of the normal distribution, we have:

$$\boxed{\lim_{n\to\infty} \mathbb{P}\Big(\theta \in \bar{x} \pm \frac{\sqrt{\bar{x}}\cdot z_{\frac{\alpha}{2}}}{\sqrt{n}}\Big) = 1-\alpha}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Note: any interval $\big[\bar{x} - \frac{\hat{\theta}_{ML}\cdot z_\epsilon}{\sqrt{n}}; \bar{x} + \frac{\hat{\theta}_{ML}\cdot z_{1-\alpha-\epsilon}}{\sqrt{n}}\big]$ also works. Here we have taken the symmetric interval around zero, choosing $\epsilon = \frac{\alpha}{2}$.

Question 8*

Using δ-method (seen during refreshers), prove that $\sqrt{n}(2\sqrt{\hat{\theta}_{ML}} - 2\sqrt{\theta}))$ converges in distribtion as $n \to \infty$.

Let $g(x) = 2\sqrt{x}$ be a function defined on $\mathbb{R}_+$. g is continuous and differentiable on $\mathbb{R}_+^*$ and $g'(x) = \frac{1}{\sqrt{x}}$.

Applying the δ-method we have: $\sqrt{n}\big(g(\bar{x}) - g(\mu)\big) \xrightarrow{d} \mathcal{N}\big(0, g(\mu)^2\sigma^2\big)$ when $n \to \infty$. With $\mu = \theta$, $\sigma^2 = \theta$ and $\bar{x} = \hat{\theta}_{ML}$, we can re-write it as:

$$\sqrt{n}\Big(2\sqrt{\hat{\theta}_{ML}} - 2\sqrt{\theta}\Big) \xrightarrow{d} \mathcal{N}\Big(0, (\frac{1}{\sqrt{\theta}})^2\cdot\theta\Big) = \mathcal{N}(0,1)$$

when $n \to \infty$.

Question 9* (another CI)

Give another asymptotic confidence interval of level α, based on question 9, that is an interval $[c_n(\alpha,(X_i)_{i\in 1,\dots,n}); d_n(\alpha,(X_i)_{i\in 1,\dots,n})]$, such that:

$$\lim_{n\to\infty} \mathbb{P}\Big(\theta \in \big[c_n(\alpha,(X_i)_{i\in 1,\dots,n}); d_n(\alpha,(X_i)_{i\in 1,\dots,n})\big]\Big) \geq 1-\alpha.$$

From the last question, we know that:

$$\lim_{n\to\infty} \mathbb{P}\Big(z_{\frac{\alpha}{2}} \leq 2\sqrt{n}(\sqrt{\hat{\theta}_{ML}} - \sqrt{\theta} \leq z_{1-\frac{\alpha}{2}} \Big) = 1 - \alpha$$

$$\lim_{n\to\infty} \mathbb{P}\Big(\sqrt{\hat{\theta}_{ML}} - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}} \leq \sqrt{\theta} \leq \sqrt{\hat{\theta}_{ML}} + \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}} \Big) = 1 - \alpha$$

$$\boxed{\lim_{n\to\infty} \mathbb{P}\Big(\theta \in (\sqrt{\bar{x}} \pm \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}})^2 = 1 - \alpha} \qquad \text{using } z_{\frac{\alpha}{2}} = z_{1-\frac{\alpha}{2}} \text{ and } \hat{\theta}_{ML} = \bar{x}.$$

Question 10

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

1st moment condition: $\mathbb{E}(X) = \mu = \frac{1}{n}\sum_{i=1}^{n}(x_i)$

which gives $\boxed{\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i) = \bar{x}}$.

2nd moment condition: $\mathbb{E}(X^2) = \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2)$ i.e. $\hat{\theta}_2 + \hat{\theta}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2)$

which gives $\boxed{\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2) - \bar{x}^2}$.

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

$\hat{\theta}_1 = \bar{x}$ is the same as the maximum likelihood estimator.

Question 11

Compute the Bias, the Variance, and the quadratic risk of $\hat{\theta}_{ML}$.

$$\begin{aligned}
\mathbb{E}(\hat{\theta}_{ML}) &= \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n}(x_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}(x_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\theta) \qquad \text{since } x_i s \text{ are Poisson iid} \\
&= \theta
\end{aligned}$$

so $\boxed{Bias(\hat{\theta}_{ML}) = 0}$.

$$\begin{aligned}
Var(\hat{\theta}_{ML}) = Var(\frac{1}{n}\sum_{i=1}^{n} x_i) &= \frac{1}{n^2}Var(\sum_{i=1}^{n} x_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) \quad \text{by independence} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \theta \qquad \text{identically Poisson-distributed} \\
&= \frac{\theta}{n}
\end{aligned}$$

$$\begin{aligned}
Risk(\theta, \hat{\theta}_{ML}) &= \mathbb{E}_\theta[(\hat{\theta}_{ML} - \theta)^2] \\
&= \mathbb{E}_\theta[(\hat{\theta}_{ML} - \mathbb{E}(\hat{\theta}_{ML}))^2] \quad \text{since } \mathbb{E}(\hat{\theta}_{ML}) = \theta \\
&= Var_\theta(\hat{\theta}_{ML}) \\
&= \frac{\theta}{n}
\end{aligned}$$

Question 12*

Compute the Cramer Rao bound. What do you conclude about $\hat{\theta}_{ML}$?

The Cramer Rao bound for the unbiased estimator $\hat{\theta}$ of θ is :

$Var(\hat{\theta}) \geq \frac{1}{I_n(\theta)}$

where $I_n = \mathbb{E}[-L_n''(\theta)]$.

Recall that $L_n'(\theta) = -n + \frac{1}{\theta}\sum_{i=1}^n x_i$

hence $L_n''(\theta) = -\frac{1}{\theta^2}\sum_{i=1}^n x_i$

so $I_n = \mathbb{E}[\frac{1}{\theta^2}\sum_{i=1}^n x_i] = \frac{1}{\theta^2}\sum_{i=1}^n \mathbb{E}[x_i] = \frac{n}{\theta}$

i.e. $\boxed{Var(\hat{\theta}) \geq \dfrac{\theta}{n}}$.

Since $Var(\hat{\theta}_{ML}) = \frac{\theta}{n}$ is equal to that lower bound, we can conclude that the Maximum Likelihood estimator is **efficient** in this case.

Question 13

Let $\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$, with $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n (X_i)$. Show that:

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^n (X_i - \theta)^2 - (\theta - \bar{X}_n)^2$$

$$\begin{aligned}
\hat{\theta}_2 &= \frac{1}{n}\sum_{i=1}^n (X_i - X_n)^2 \\
&= \frac{1}{n}\sum_{i=1}^n [(X_i - \theta) - (\bar{X}_n - \theta)]^2 \\
&= \frac{1}{n}\sum_{i=1}^n \Big((X_i - \theta)^2 - 2(X_i - \theta)(\bar{X}_n - \theta) + (\bar{X}_n - \theta)^2\Big) \\
&= \frac{1}{n}\sum_{i=1}^n (X_i - \theta)^2 - \frac{2}{n}(\bar{X}_n - \theta)\sum_{i=1}^n (X_i - \theta) + \frac{1}{n}\sum_{i=1}^n (\bar{X}_n - \theta)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (X_i - \theta)^2 - \frac{2}{n}(\bar{X}_n - \theta)(n\bar{X}_n - n\theta) + \frac{1}{n}n(\bar{X}_n - \theta)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (X_i - \theta)^2 - 2(\bar{X}_n - \theta)^2 + (\bar{X}_n - \theta)^2 \\
\hat{\theta}_2 &= \frac{1}{n}\sum_{i=1}^n (X_i - \theta)^2 - (\bar{X}_n - \theta)^2
\end{aligned}$$

Question 14

- **Compute $\mathbb{E}(\theta - \bar{X}_n)^2$.**

$$\begin{aligned}\mathbb{E}(\theta - \bar{X}_n)^2 &= \mathbb{E}(\theta^2 - 2\theta\bar{X}_n + \bar{X}_n^2)\\ &= \theta^2 - \theta\mathbb{E}(\bar{X}_n) + \mathbb{E}(\bar{X}_n^2)\\ &= \theta^2 - \theta\cdot\theta + Var(\bar{X}_n) + \mathbb{E}(\bar{X}_n)^2\\ &= \theta^2 - 2\theta^2 + \frac{\theta}{n} + \theta^2\end{aligned}$$

$$\mathbb{E}(\theta - \bar{X}_n)^2 = \frac{\theta}{n}$$

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

$$\begin{aligned}\mathbb{E}(\hat{\theta}_2^2) &= \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2 - (\theta - \bar{X}_n)^2\Big)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i - \theta)^2 - \frac{\theta}{n}\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - 2\theta\mathbb{E}(X_i) + \theta^2 - \frac{\theta}{n}\\ &= \frac{1}{n}\sum_{i=1}^{n}Var(X_i) + \mathbb{E}(X_i)^2 - 2\theta^2 + \theta^2 - \frac{\theta}{n}\\ &= \frac{1}{n}\sum_{i=1}^{n}\theta + \theta^2 - 2\theta^2 + \theta^2 - \frac{\theta}{n}\end{aligned}$$

$$\mathbb{E}(\hat{\theta}_2^2) = \theta - \frac{\theta}{n}$$

So $\boxed{Bias(\hat{\theta}_2) = -\frac{\theta}{n}}$.

To obtain an unbiased estimator, we can change

$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

to be

$\hat{\theta}'_2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.

Then, we can show, as in question 13, that $\hat{\theta}'_2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \theta)^2 - \frac{n}{n-1}(\bar{X}_n - \theta)^2$

And, using $\mathbb{E}(\theta - \bar{X}_n)^2 = \frac{\theta}{n}$ as proven above, we indeed obtain:

$\mathbb{E}(\hat{\theta}'_2) = \frac{1}{n-1}(\sum_{i=1}^{n}\theta) - \frac{n}{n-1}\cdot\frac{\theta}{n}$

$\mathbb{E}(\hat{\theta}'_2) = \frac{n}{n-1}\theta - \frac{\theta}{n-1} = \frac{(n-1)\theta}{n-1} = \theta$

Hence $\boxed{Bias(\hat{\theta}'_2) = 0}$.

Question 15

- **Using the decomposition in Question 13, prove that $\sqrt{n}(\hat{\theta}_2 - \theta)$ converges in distribution, and give a third asymptotic confidence interval centered in $\hat{\theta}_2$. Comment on this third interval.**
- **You may use: $Var[(X_i - \theta)^2] = 2\theta^2 + \theta$.**
- **Compare the asymptotic variance to the one of $\hat{\theta}_{ML}$ and to the Cramer Rao bound. What can you say?**

Question 16*

- **Compute the probability generating function of the Poisson distribution given by $G_X(s) = \mathbb{E}[s^X]$ for $s \in \mathbb{R}$.**

$$\begin{aligned}\forall s \in \mathbb{R}, M_X(s) &= \mathbb{E}[e^{sX}] \\ &= \sum_{k=0}^{\infty} e^{sk}\mathbb{P}(X = k) \\ &= \sum_{k=0}^{\infty} e^{sk}e^{-\theta}\frac{\theta^k}{k!} \\ &= e^{-\theta}\sum_{k=0}^{\infty}\frac{(\theta e^s)^k}{k!} \\ &= e^{-\theta}e^{\theta e^s}\end{aligned}$$

Therefore we obtain: $\forall s \in \mathbb{R}$, $\boxed{M_X(s) = e^{\theta(e^s-1)}}$.

We can recover the mean:

$$\begin{aligned}M'(s) &= e^{\theta(e^s-1)}\theta e^s \\ &= \theta e^{\theta(e^s-1)+s}\end{aligned}$$

so $\mu = E[X] = M'(0) = \theta.e^0 = \theta$

And the variance:

$$M''(s) = \theta e^{\theta(e^s-1)+s}(\theta e^s + 1)$$

so

$$\begin{aligned}\sigma^2 = E[X^2] - (E[X])^2 \quad &= M''(0) - (M'(0)^2) \\ &= \theta e^0(\theta + 1) - \theta^2 \\ &= \theta^2 + \theta - \theta^2 \\ &= \theta\end{aligned}$$

- **Recover the result of question 2 and prove $Var[(X_i - \theta)^2] = 2\theta^2 + \theta$.**

$$\begin{aligned}Var[(X_i - \theta)^2] &= \mathbb{E}[(X_i - \theta)^4] - (\mathbb{E}[(X_i - \theta)^2])^2 \\ &= \mathbb{E}(X_i^4) - 4\theta\mathbb{E}(X_i^3) + 6\theta^3 + 3\theta^4 - \theta^2\end{aligned}$$

Using $\mathbb{E}(X_i) = \theta$ and $\mathbb{E}(X_i^2) = \theta + \theta^2$.

We can use the moment generating function to compute $\mathbb{E}(X_i^3)$ and $\mathbb{E}(X_i^4)$, and we obtain:

$M^{(3)}(s) = (\theta^3 e^{2s} + 3\theta^2 e^s + \theta)e^{\theta e^s - \theta + s}$
$M^{(4)}(s) = (\theta^4 e^{3s} + 6\theta^3 e^{2s} + \theta)e^{\theta e^s - \theta + s}$

so $\mathbb{E}(X_i^3) = M^{(3)}(0) = \theta^3 + 3\theta^2 + \theta$
and $\mathbb{E}(X_i^4) = M^{(4)}(0) = \theta^4 + 6\theta^3 + 7\theta^2 + \theta$.

Therefore,
$Var[(X_i - \theta)^2] = \theta^4 + 6\theta^3 + 7\theta^2 + \theta - 4\theta^4 - 12\theta^3 - 4\theta^2 + 6\theta^3 + 3\theta^4 - \theta^2$
$\boxed{Var[(X_i - \theta)^2] = 2\theta^2 + \theta}$

Problem 2: Analysis of the USJudgeRatings data set

This exercise is open. You are asked to use the tools we have seen together to analyze the USJudgeRatings data set. This data set is provided in the package datasets. Your analysis should be reported here and include:

- an introduction
- a general description of the data
- the use of descriptive statistics
- the use of all techniques we have seen together that might be relevant
- a conclusion

Overall, your analysis, including the graphs and the codes should not exceed 15 pages in pdf.

Loading and describing the dataset (variables, missing values, outliers, numerical summaries)

```
data(USJudgeRatings)
?USJudgeRatings
head(USJudgeRatings)
```

```
##                CONT INTG DMNR DILG CFMG DECI PREP FAMI ORAL WRIT PHYS RTEN
## AARONSON,L.H.   5.7  7.9  7.7  7.3  7.1  7.4  7.1  7.1  7.1  7.0  8.3  7.8
## ALEXANDER,J.M.  6.8  8.9  8.8  8.5  7.8  8.1  8.0  8.0  7.8  7.9  8.5  8.7
## ARMENTANO,A.J.  7.2  8.1  7.8  7.8  7.5  7.6  7.5  7.5  7.3  7.4  7.9  7.8
## BERDON,R.I.     6.8  8.8  8.5  8.8  8.3  8.5  8.7  8.7  8.4  8.5  8.8  8.7
## BRACKEN,J.J.    7.3  6.4  4.3  6.5  6.0  6.2  5.7  5.7  5.1  5.3  5.5  4.8
## BURNS,E.B.      6.2  8.8  8.7  8.5  7.9  8.0  8.1  8.0  8.0  8.0  8.6  8.6
```

```
str(USJudgeRatings)
```

```
## 'data.frame':    43 obs. of  12 variables:
##  $ CONT: num  5.7 6.8 7.2 6.8 7.3 6.2 10.6 7 7.3 8.2 ...
##  $ INTG: num  7.9 8.9 8.1 8.8 6.4 8.8 9 5.9 8.9 7.9 ...
##  $ DMNR: num  7.7 8.8 7.8 8.5 4.3 8.7 8.9 4.9 8.9 6.7 ...
##  $ DILG: num  7.3 8.5 7.8 8.8 6.5 8.5 8.7 5.1 8.7 8.1 ...
##  $ CFMG: num  7.1 7.8 7.5 8.3 6 7.9 8.5 5.4 8.6 7.9 ...
##  $ DECI: num  7.4 8.1 7.6 8.5 6.2 8 8.5 5.9 8.5 8 ...
##  $ PREP: num  7.1 8 7.5 8.7 5.7 8.1 8.5 4.8 8.4 7.9 ...
##  $ FAMI: num  7.1 8 7.5 8.7 5.7 8 8.5 5.1 8.4 8.1 ...
##  $ ORAL: num  7.1 7.8 7.3 8.4 5.1 8 8.6 4.7 8.4 7.7 ...
##  $ WRIT: num  7 7.9 7.4 8.5 5.3 8 8.4 4.9 8.5 7.8 ...
##  $ PHYS: num  8.3 8.5 7.9 8.8 5.5 8.6 9.1 6.8 8.8 8.5 ...
##  $ RTEN: num  7.8 8.7 7.8 8.7 4.8 8.6 9 5 8.8 7.9 ...
```

```
dim(USJudgeRatings)
```

```
## [1] 43 12
```

We are provided with 43 observations on 12 numeric (continuous) variables. The observations correspond to lawyers' ratings of state judges in the US Superior Court (New Haven Register, 14 January, 1977). Meaning of the variables' abbreviations is the following:

CONT Number of contacts of lawyer with judge. INTG Judicial integrity. DMNR Demeanor. DILG Diligence. CFMG Case flow managing. DECI Prompt decisions. PREP Preparation for trial. FAMI Familiarity with law. ORAL Sound oral rulings. WRIT Sound written rulings. PHYS Physical ability. RTEN Worthy of retention.

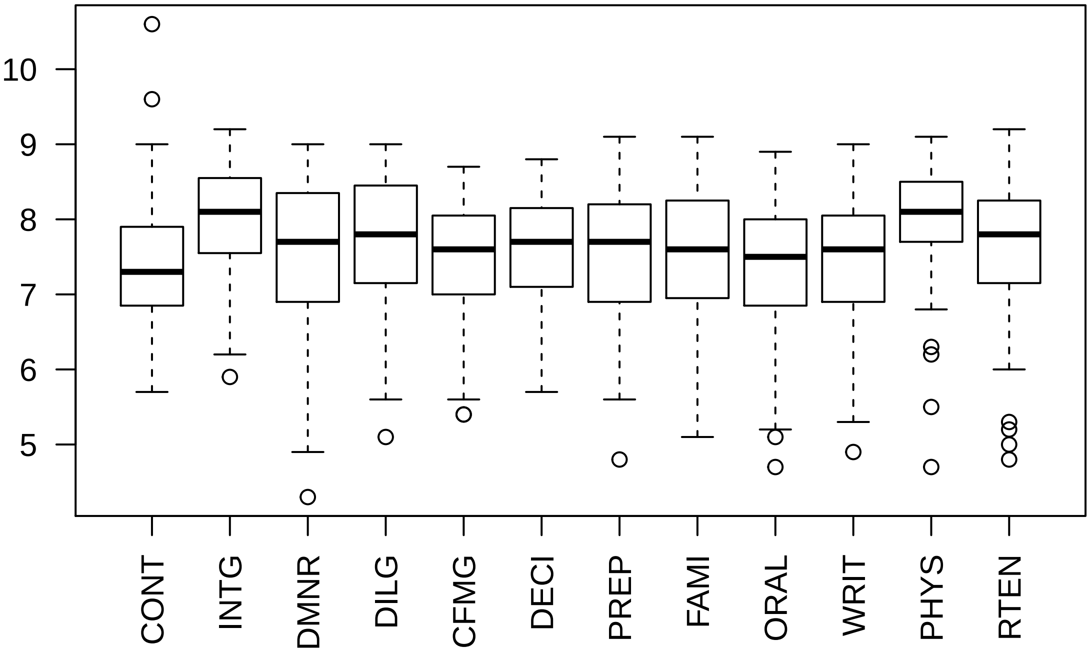
```
sum(is.na(USJudgeRatings))
```

```
## [1] 0
```

There are no missing values.

Let's now draw boxplots to study the presence of mistakes and/or outliers:

```
boxplot(USJudgeRatings, las = 2)
```



We observe the presence of a few outliers with low values for all variables except DECI (no outliers), FAMI (no outliers) and CONT (2 outliers with large values).

Let's point out for example the highest outlier for the CONT variable, corresponding to the lawyer with the highest number of contacts with State judges:

```
max(USJudgeRatings$CONT)
```

```
## [1] 10.6
```

```
which.max(USJudgeRatings$CONT)
```

```
## [1] 7
```

```
rownames(USJudgeRatings)[which.max(USJudgeRatings$CONT)]
```

```
## [1] "CALLAHAN,R.J."
```

The rating given by lawyer Callahan for the CONT criterion is the only rating above 10 in the whole dataset. It looks like a mistake. We will therefore remove this line:

```
data = USJudgeRatings[-7,]
dim(data)
```

```
## [1] 42 12
```

We are not provided with additional information and we cannot check whether the other outliers we observe correspond to mistakes. We will therefore assume they are *not* mistakes.

Computing numerical summaries

Let's now compute numerical summaries for each variable to study the location (empirical mean, median) and dispersion (interquartile range) of the variables. We will then compute the empirical variances and standard deviations in order to futher study the dispersion of the variables.

```
summary(data)
```

```
##       CONT            INTG            DMNR            DILG
##  Min.   :5.700   Min.   :5.900   Min.   :4.300   Min.   :5.100
##  1st Qu.:6.825   1st Qu.:7.525   1st Qu.:6.900   1st Qu.:7.125
##  Median :7.250   Median :8.100   Median :7.700   Median :7.800
##  Mean   :7.362   Mean   :7.998   Mean   :7.483   Mean   :7.669
##  3rd Qu.:7.775   3rd Qu.:8.500   3rd Qu.:8.275   3rd Qu.:8.375
##  Max.   :9.600   Max.   :9.200   Max.   :9.000   Max.   :9.000
##       CFMG            DECI            PREP            FAMI
##  Min.   :5.400   Min.   :5.700   Min.   :4.800   Min.   :5.100
##  1st Qu.:7.000   1st Qu.:7.100   1st Qu.:6.900   1st Qu.:6.925
##  Median :7.550   Median :7.700   Median :7.650   Median :7.550
##  Mean   :7.455   Mean   :7.543   Mean   :7.443   Mean   :7.464
##  3rd Qu.:8.000   3rd Qu.:8.100   3rd Qu.:8.175   3rd Qu.:8.175
##  Max.   :8.700   Max.   :8.800   Max.   :9.100   Max.   :9.100
##       ORAL            WRIT           PHYS            RTEN
##  Min.   :4.700   Min.   :4.90   Min.   :4.700   Min.   :4.800
##  1st Qu.:6.825   1st Qu.:6.85   1st Qu.:7.650   1st Qu.:7.125
##  Median :7.450   Median :7.50   Median :8.100   Median :7.800
##  Mean   :7.262   Mean   :7.36   Mean   :7.907   Mean   :7.569
##  3rd Qu.:7.975   3rd Qu.:8.00   3rd Qu.:8.500   3rd Qu.:8.200
##  Max.   :8.900   Max.   :9.00   Max.   :9.000   Max.   :9.200
```

```
# Variance of each variable
var = diag(var(USJudgeRatings))
round(var, 2)
```

```
## CONT INTG DMNR DILG CFMG DECI PREP FAMI ORAL WRIT PHYS RTEN
## 0.89 0.59 1.31 0.81 0.74 0.64 0.91 0.90 1.02 0.92 0.88 1.21
```

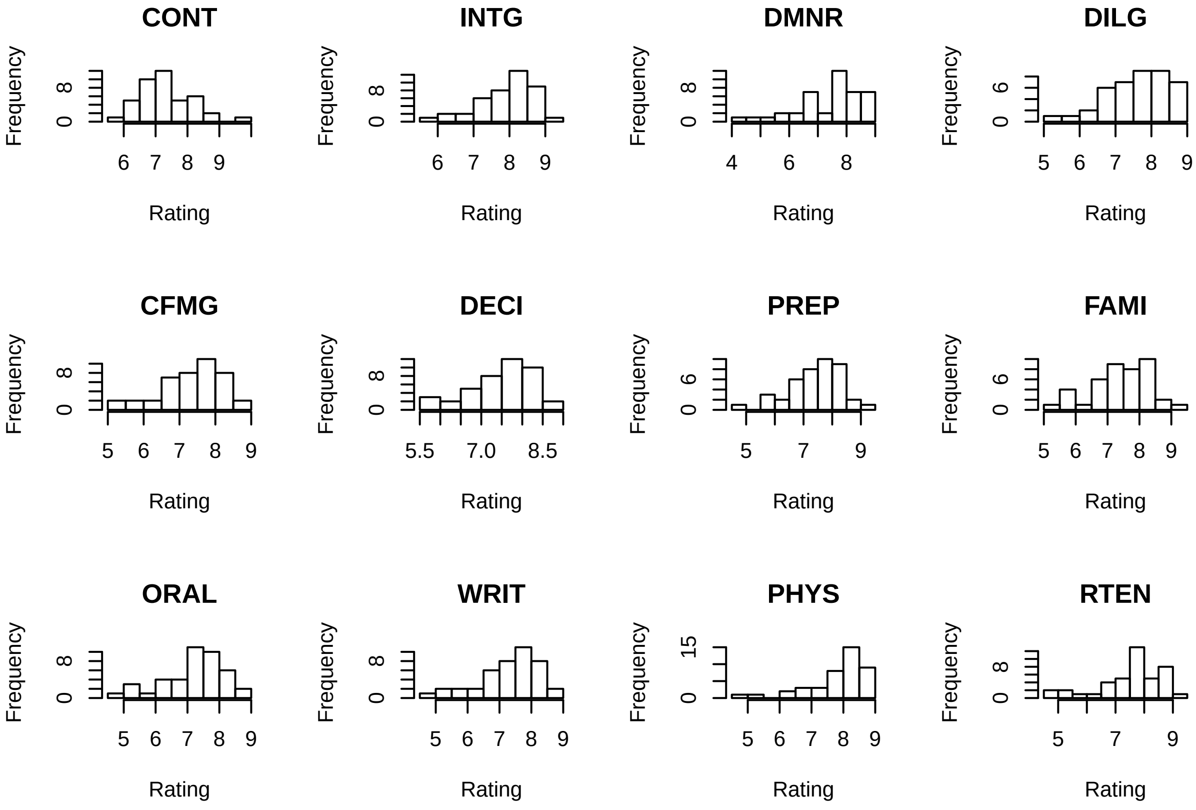
```
# Standard deviation of each variable
stdev = sqrt(diag(var(USJudgeRatings)))
round(stdev, 2)
```

```
## CONT INTG DMNR DILG CFMG DECI PREP FAMI ORAL WRIT PHYS RTEN
## 0.94 0.77 1.14 0.90 0.86 0.80 0.95 0.95 1.01 0.96 0.94 1.10
```

The lowest variability is observed for the DECI ("Prompt decisions") variable, meaning that this was the criterion on which the voting lawyers were the most aligned. Conversely, the largest variability is observed for the DMNR ("Demeanor") and RTEN ("Worthy of retention") variables, meaning that these criteria were those on which lawyers were the most unaligned. This can also be seen in the boxplots sizes (wider, i.e. more variability, for DMRN and RTEN, narrower for DECI).

Last, as a visual tool, histograms allow us to illustrate the dispersion and empirical laws of the variables:

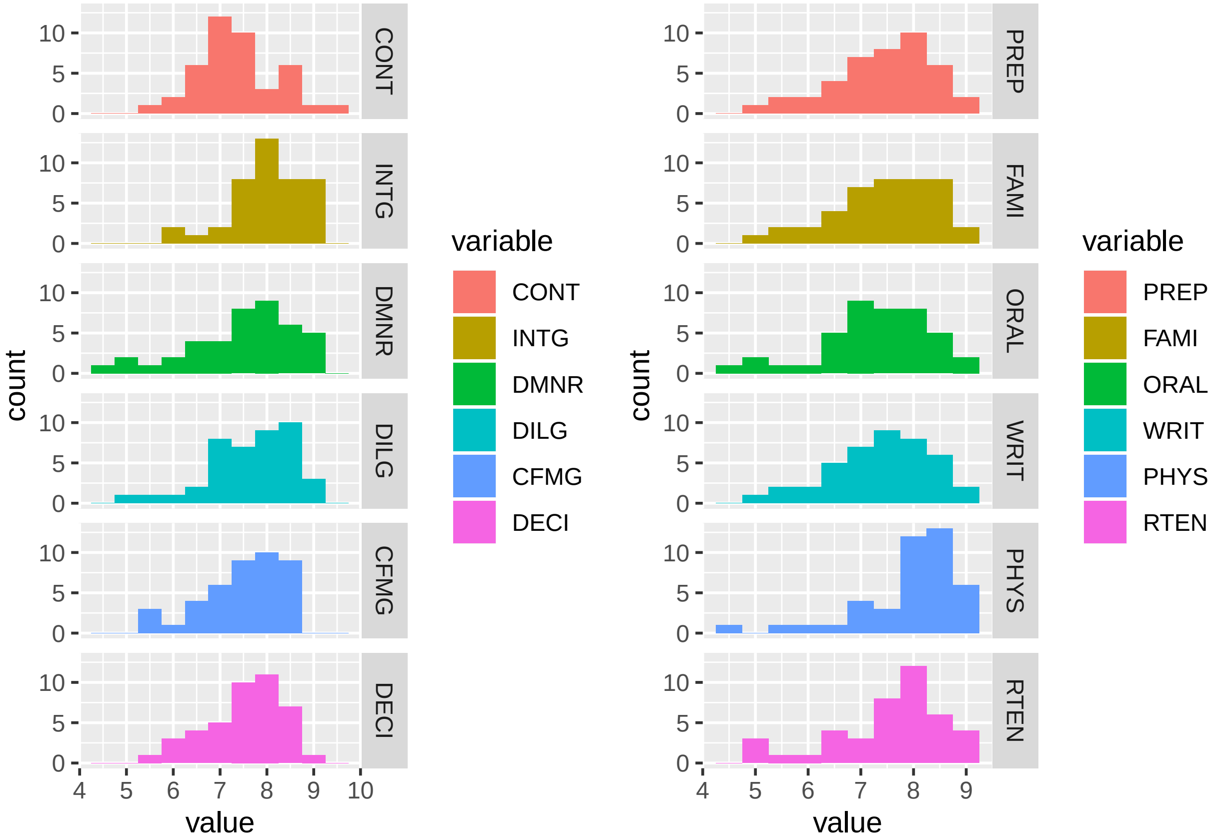
```
# Histograms using apply functions

par(mfrow = c(3,4))
invisible(mapply(FUN = hist, data, main = colnames(data), xlab = "Rating", breaks = 10))
```



```
## "Invisible" avoids to display an additional (and non-visual) console output
```

```
# Histograms using ggplot functions

library(reshape2)
library(ggplot2)
library(gridExtra)
```

```r
ggdata1 <- melt(data[,1:6]) ## We split data in 2 for better visual representation below
ggdata2 <- melt(data[,7:12])
plot1 <- ggplot(ggdata1, aes(x = value, fill = variable)) + geom_histogram(binwidth = 0.5) +
  facet_grid(variable~.)
plot2 <- ggplot(ggdata2, aes(x = value, fill = variable)) + geom_histogram(binwidth = 0.5) +
  facet_grid(variable~.)
grid.arrange(plot1, plot2, ncol = 2, nrow = 1)
```



Studying the relationship between the variables

Let's now study the (linear) relationship between the variables using the empirical correlation matrix, correlation plots and scatter plots:

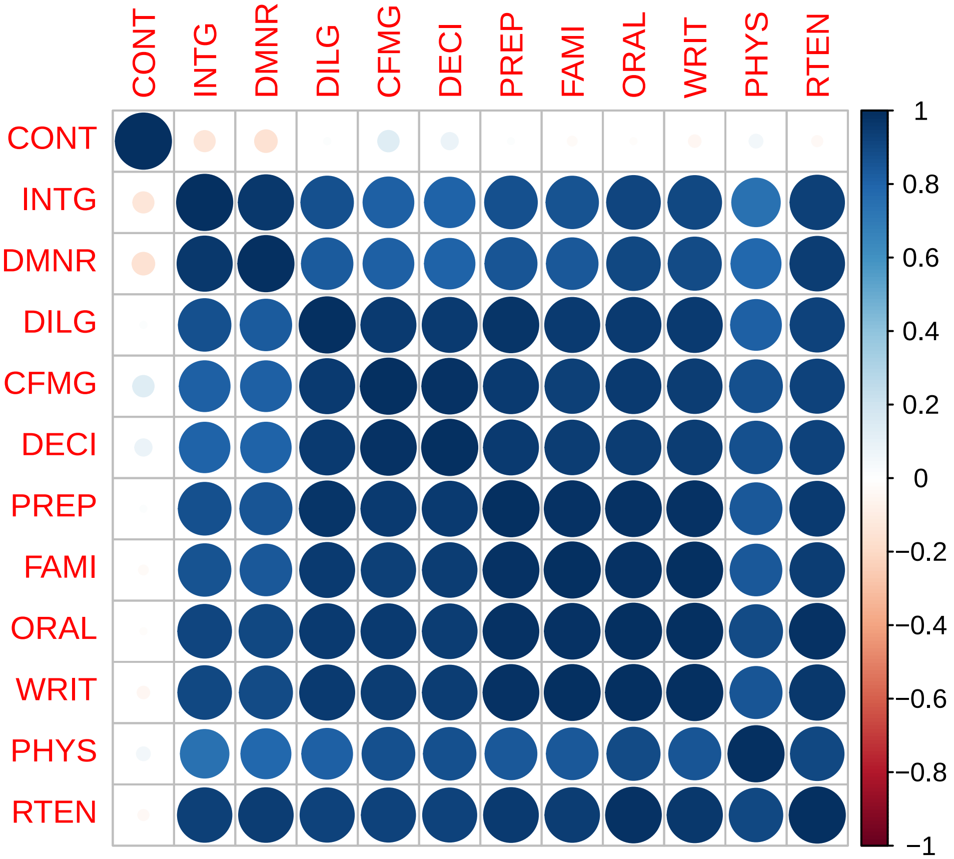
```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
cor = round(cor(USJudgeRatings), 2)
cor
```

```
##       CONT  INTG  DMNR DILG CFMG DECI PREP  FAMI  ORAL  WRIT PHYS  RTEN
## CONT  1.00 -0.13 -0.15 0.01 0.14 0.09 0.01 -0.03 -0.01 -0.04 0.05 -0.03
## INTG -0.13  1.00  0.96 0.87 0.81 0.80 0.88  0.87  0.91  0.91 0.74  0.94
## DMNR -0.15  0.96  1.00 0.84 0.81 0.80 0.86  0.84  0.91  0.89 0.79  0.94
## DILG  0.01  0.87  0.84 1.00 0.96 0.96 0.98  0.96  0.95  0.96 0.81  0.93
## CFMG  0.14  0.81  0.81 0.96 1.00 0.98 0.96  0.94  0.95  0.94 0.88  0.93
## DECI  0.09  0.80  0.80 0.96 0.98 1.00 0.96  0.94  0.95  0.95 0.87  0.92
## PREP  0.01  0.88  0.86 0.98 0.96 0.96 1.00  0.99  0.98  0.99 0.85  0.95
## FAMI -0.03  0.87  0.84 0.96 0.94 0.94 0.99  1.00  0.98  0.99 0.84  0.94
## ORAL -0.01  0.91  0.91 0.95 0.95 0.95 0.98  0.98  1.00  0.99 0.89  0.98
## WRIT -0.04  0.91  0.89 0.96 0.94 0.95 0.99  0.99  0.99  1.00 0.86  0.97
## PHYS  0.05  0.74  0.79 0.81 0.88 0.87 0.85  0.84  0.89  0.86 1.00  0.91
## RTEN -0.03  0.94  0.94 0.93 0.93 0.92 0.95  0.94  0.98  0.97 0.91  1.00
```

```
corrplot(cor(USJudgeRatings))
```

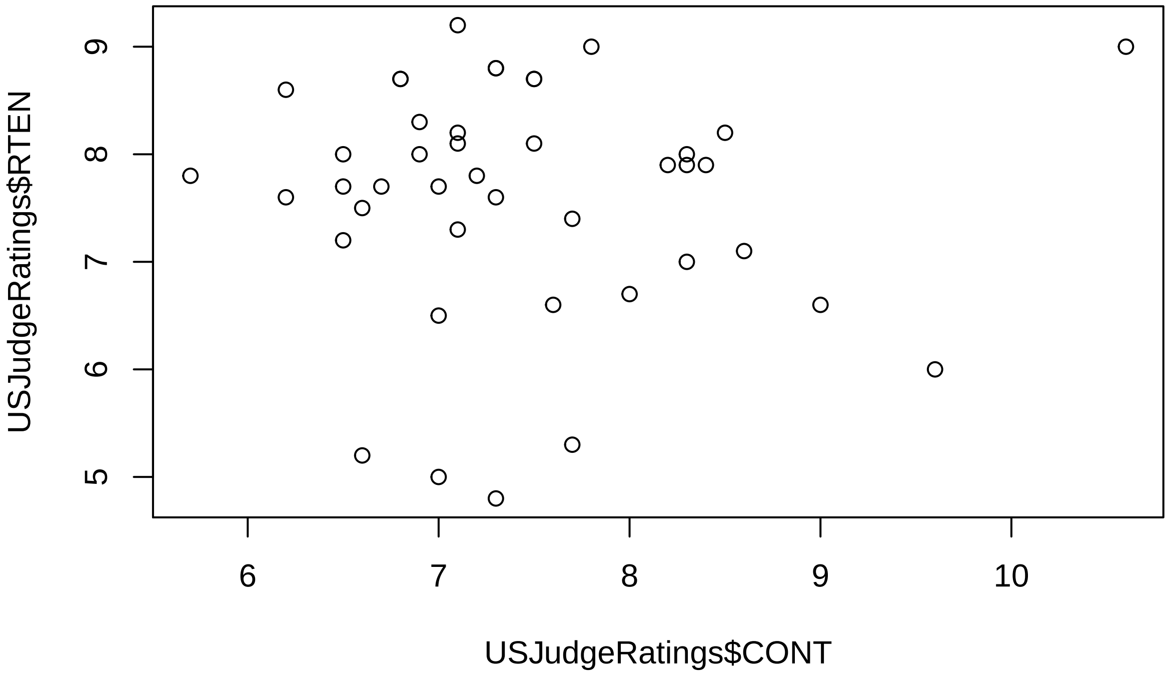


Logically, we can observe strong positive correlations (close to 1) between all variables - except for the CONT ("Number of contacts of lawyer with judge") variable which isn't correlated to the others.

The positive correlations could be related to the fact that lawyers are consistent in their ratings: when they attribute good marks to a judge, they tend to do so across all rating criteria.

As for CONT, its isolation can be explained by the fact that the number of contacts is the only variable which isn't a rating criterion (it's not a judgment from lawyers, just an evaluation of their personal ties with the judge). Interestingly (and reassuringly!), it seems that personal ties aren't correlated to the marks attributed to the judge.

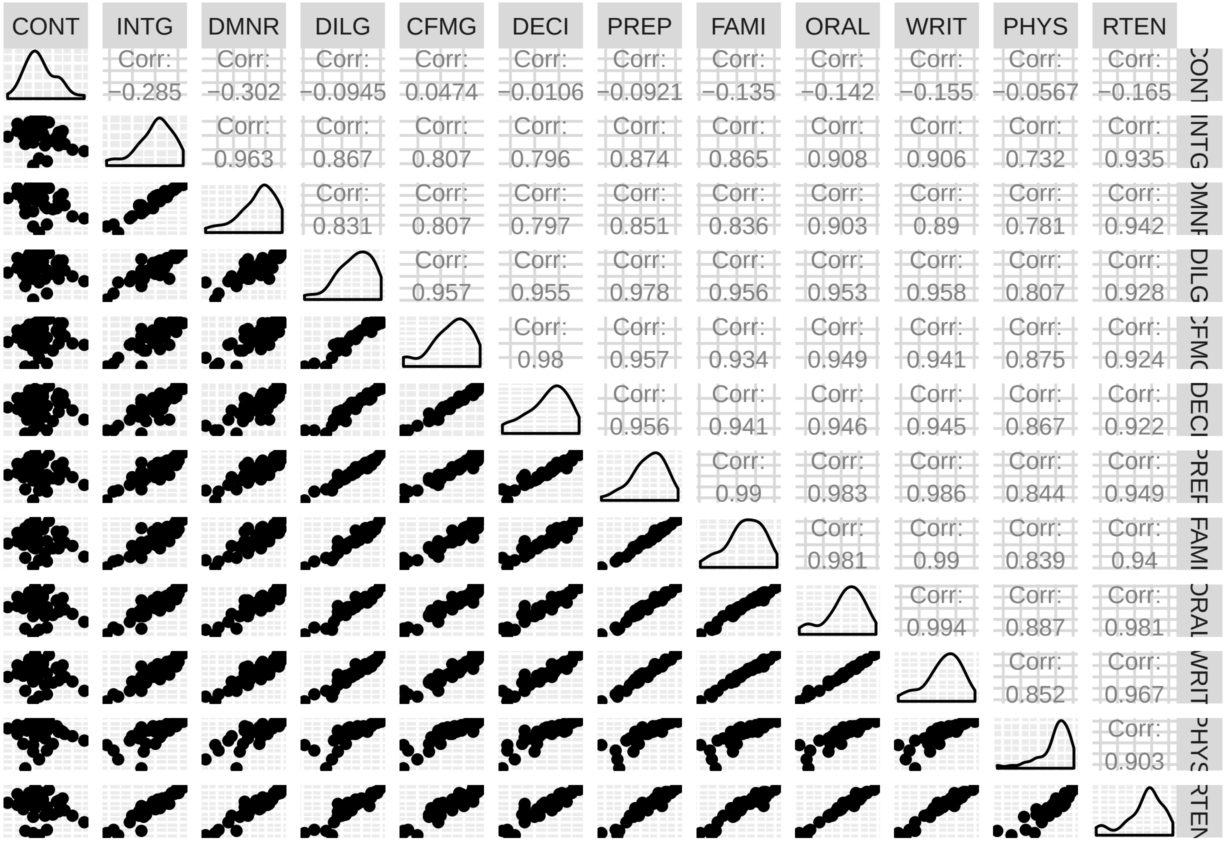
```
plot(USJudgeRatings$CONT, USJudgeRatings$RTEN)
```



As expected following the correlation plot, the scatter plot doesn't show any linear relationship between CONT and (for example) RTEN ("Worthy of retention"). Plotting this also allows us to see that there is no obvious non-linear relationship either.

This section's findings regarding the relationships between variables can be visualized with ggpairs:

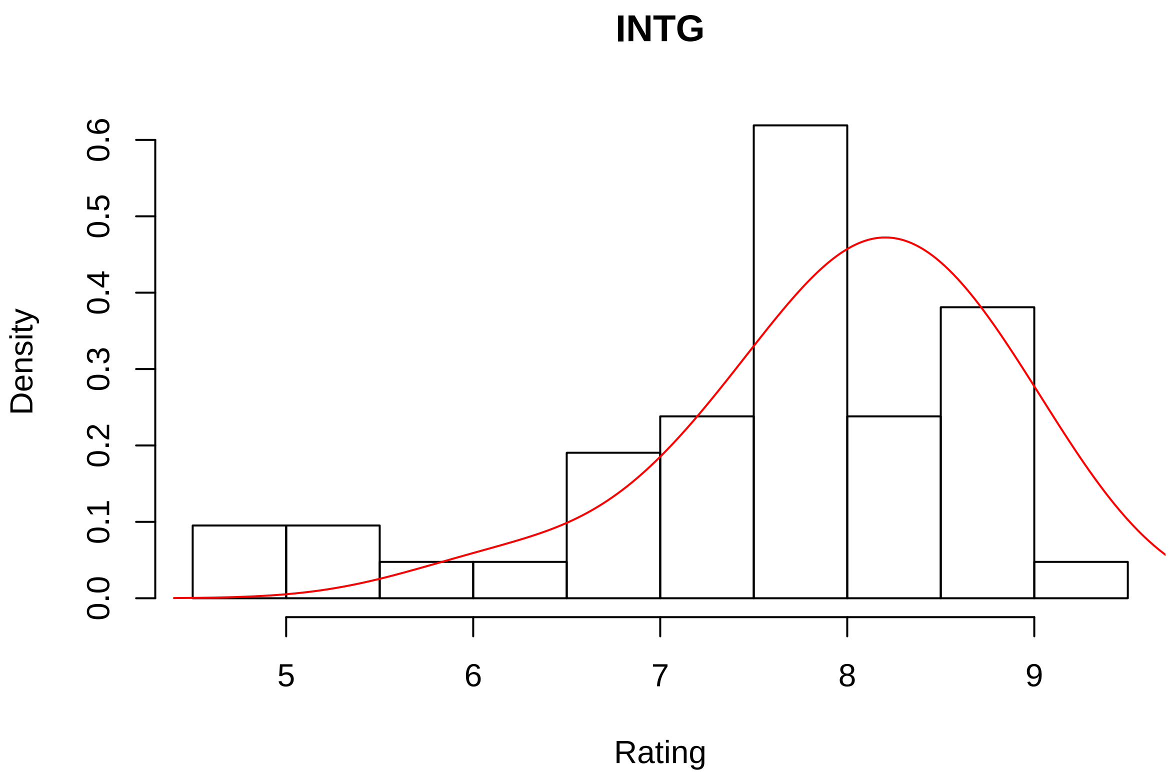
```r
library(GGally)
ggpairs(data, upper = list(continuous = wrap("cor", size = 3)), axisLabels = "none")
```



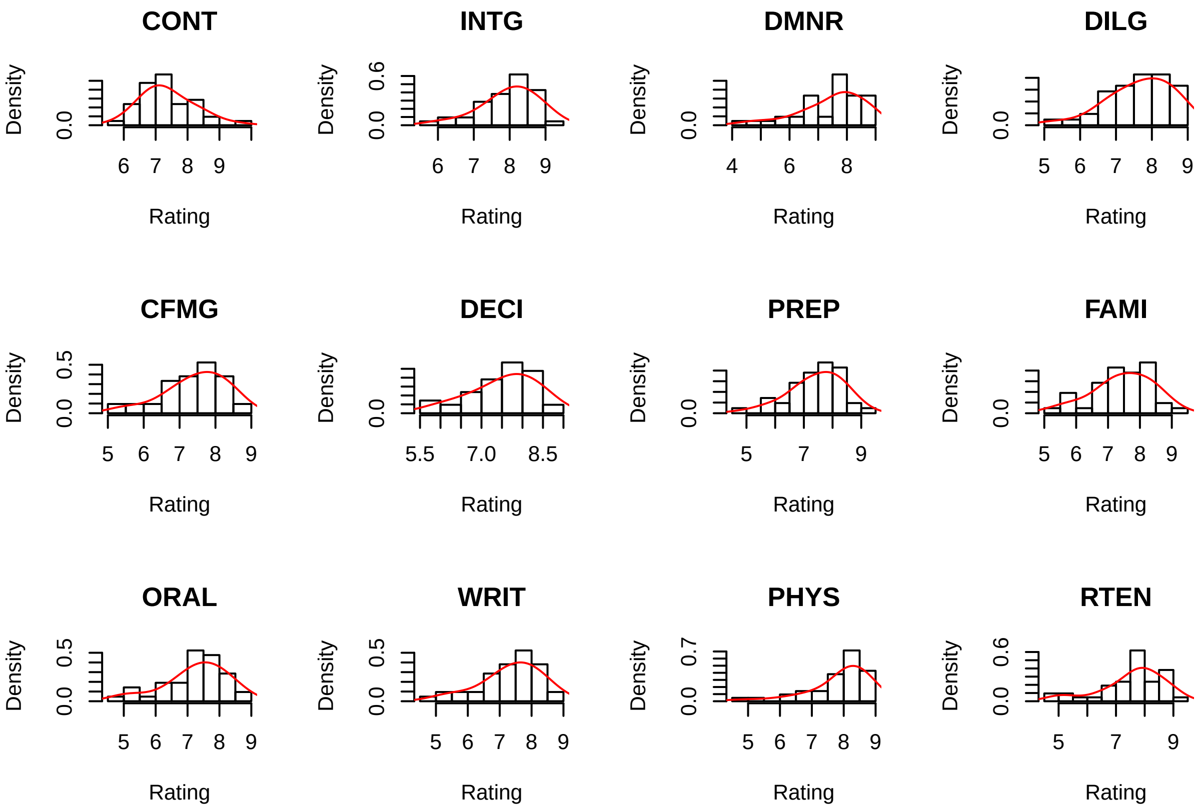
Kernel estimators to estimate the density of the variables

As we could previously notice in the histograms and the pairplot, our variables look like truncated Gaussians, with values taken in $[0, 10]$. Let's take a closer look:

```r
# Example density estimator with the RTEN variable

hist(data$RTEN, main = "INTG", xlab = "Rating", probability = T)
d <- density(data$INTG, bw = 0.5)
lines(d, col = "red")
```



```
# Density estimator for all variables

# With a for loop
par(mfrow = c(3,4))
for (i in colnames(data)) {
  hist(data[[i]], main = colnames(data[i]), xlab = "Rating", probability = T)
  d <- density(data[[i]], bw = 0.5)
  lines(d, col = "red")}

# With an *apply function
invisible(lapply(colnames(data), function(i){
  hist(data[[i]], main = colnames(data[i]), xlab = "Rating", probability = T)
  d <- density(data[[i]], bw = 0.5)
  lines(d, col = "red")}))
```

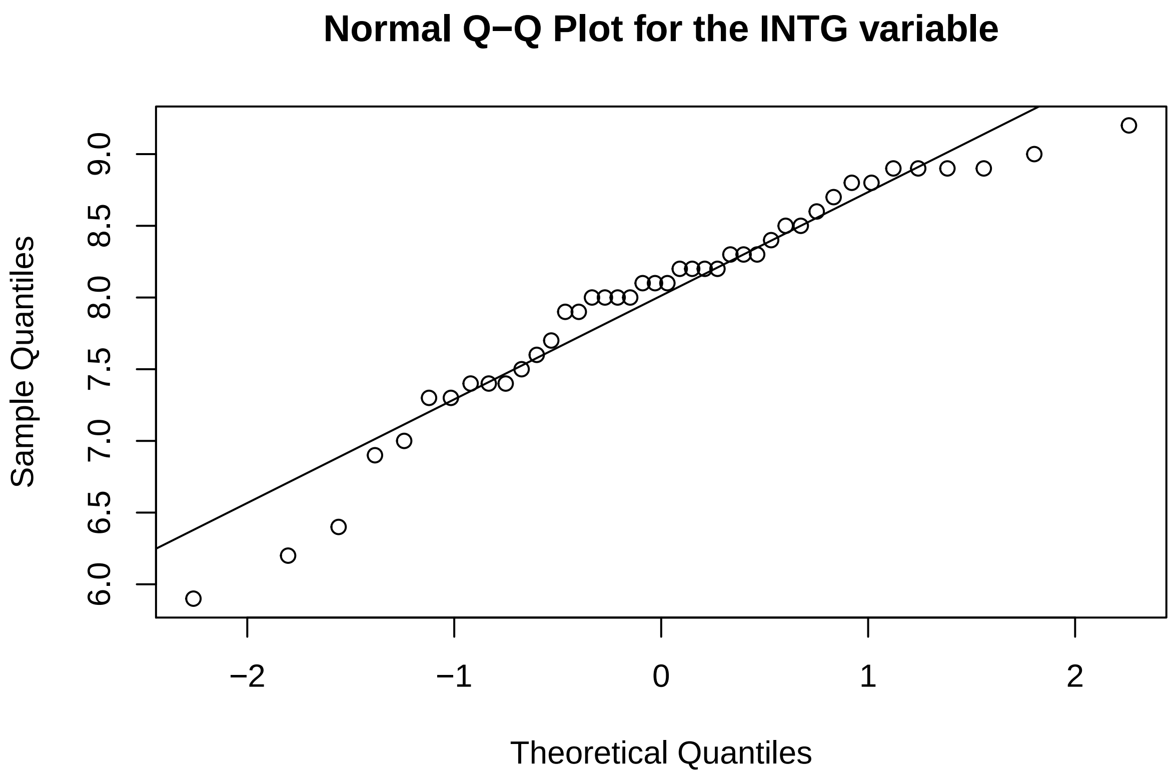


The empirical density curves show that our variables seem to be approximated by Gaussian distributions, capped by 10 along the x-axis.

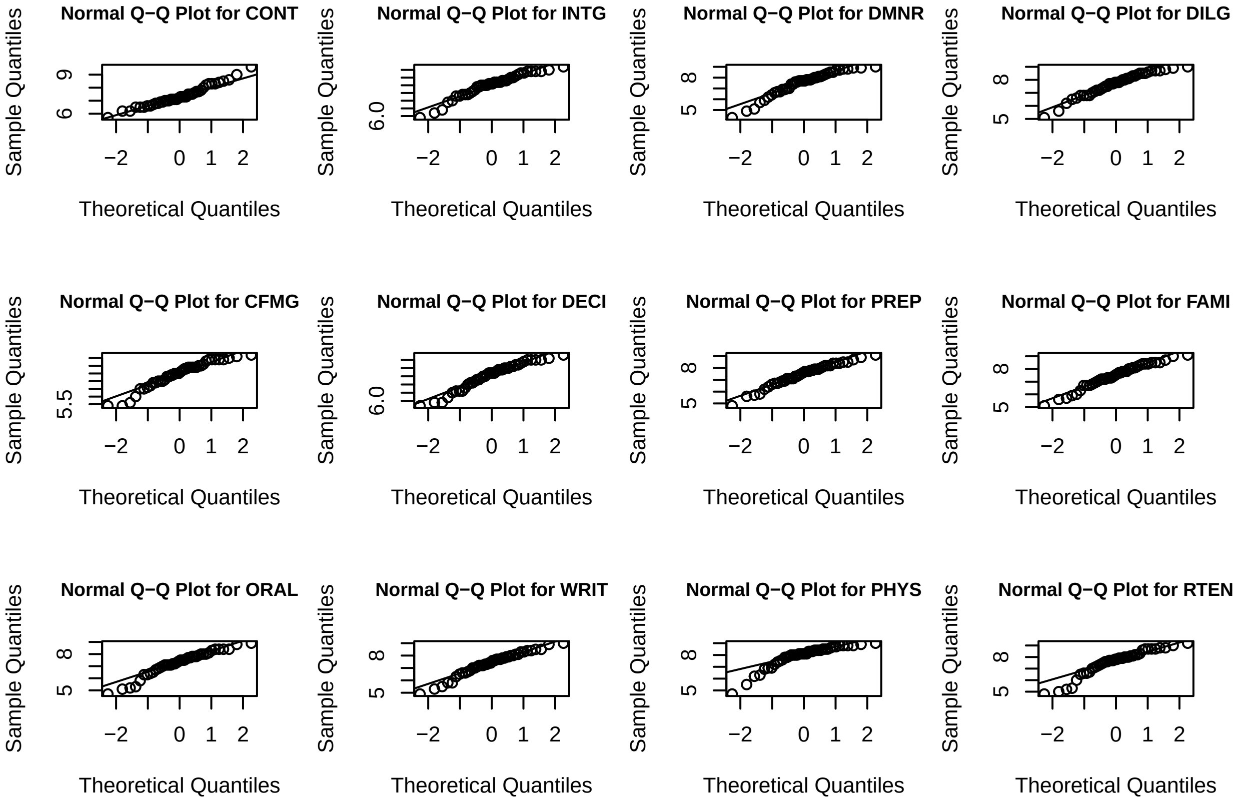
Additionaly, we could also plot Empirical cumulative distribution functions using ecdf(), but this doesn't sound useful for our analysis.

QQplots to visualize if the variables are Gaussian

```
# Example qqplot with the INTG variable

qqnorm(data$INTG, main = "Normal Q-Q Plot for the INTG variable")
qqline(data$INTG)
```



```
# QQplot all variables

# With a for loop
par(mfrow = c(3,4))
for (i in colnames(data)) {
  qqnorm(data[[i]], main = paste("Normal Q-Q Plot for", i), cex.main=0.9)
  qqline(data[[i]])}

# With an *apply function
invisible(lapply(colnames(data), function(i){
  qqnorm(data[[i]], main = paste("Normal Q-Q Plot for", i), cex.main=0.9)
  qqline(data[[i]])}))
```



The QQplot show that the normal distribution seems to be a good approximation of the distribution of each variable on the $[0, 10]$ range.

Final remarks

In addition to this analysis, we have decided not to compute confidence intervals or hypothesis tests for 2 reasons: * If we had data for other judges, confidence intervals could be useful to determine if there is any statistial evidence of one judge performing better than the others. But here we are only analysing ratings of a single judge. * A confidence interval could also help us bound the “true rating” (true mean) of the judge with 5% uncertainty, but in this case a judge cannot have an “intrinsec” or “true” rating, a “true” rating wouldn’t have a meaning *per se* - ratings of a judge only have meaning when compared to ratings of other judges.