MAP 531: Homework

Paul-Antoine GIRARD & Adrien TOULOUSE

Problem 1: Estimating parameters of a Poisson distribution

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 $\mathbb{N}$:

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

Question 1

The poisson distribution is a discrete distribution since it has a countable number of possible values ($\mathbb{N}$).

In statistics, we use this distribution to compute the probability of a given number of (rare) events in a time period.

For example a poisson distribution can model:

- The number of patients arriving in an emergency room between 9 and 10am.
- The number of network failures per day.
- In quality control, the number of manufacturing defect.

Question 2

We assume that $\mathbb{X}$ follows a Poisson distribution with parameter $\theta > 0$.

We will use the fact that $e^\theta = \sum_{i=0}^{\infty}(\frac{\theta^i}{i!}), \forall\theta \in \mathbb{R}$

$$\mathbb{E}[\mathbb{X}] = \sum_{i=0}^{\infty}(i * p(\theta, i)) = \sum_{i=0}^{\infty}(i * e^{-\theta}\frac{\theta^i}{i!}) = \theta * e^{-\theta}\sum_{i=1}^{\infty}(\frac{\theta^{i-1}}{(i-1)!}) = \theta * e^{-\theta}\sum_{i=0}^{\infty}(\frac{\theta^i}{i!}) = \theta * e^{-\theta} * e^\theta = \theta$$

$$\mathbb{E}[\mathbb{X}^2] = \sum_{i=0}^{\infty}(i^2 * p(\theta, i)) = \sum_{i=0}^{\infty}(i^2 * e^{-\theta}\frac{\theta^i}{i!}) = \theta * e^{-\theta}\sum_{i=1}^{\infty}(i\frac{\theta^{i-1}}{(i-1)!}) = \theta * e^{-\theta}\sum_{i=0}^{\infty}((i+1)\frac{\theta^i}{i!})$$

$$= \theta * e^{-\theta}[\sum_{i=0}^{\infty}(i\frac{\theta^i}{i!}) + \sum_{i=0}^{\infty}(\frac{\theta^i}{i!})] = \theta * e^{-\theta}[\theta\sum_{i=0}^{\infty}(\frac{\theta^i}{i!}) + e^\theta] = \theta * e^{-\theta}[\theta * e^\theta + e^\theta] = \theta(\theta + 1)$$

$$\mathbb{V}(\mathbb{X}) = \mathbb{E}[\mathbb{X}^2] - \mathbb{E}[\mathbb{X}]^2 = \theta(\theta + 1) - \theta^2 = \theta$$

Question 3

We are provided with n independent observations of a Poisson random variable of parameter $\theta \in \Theta = \mathbb{R}_+^*$.
Our observations are $X_k \sim Pois(\theta), \forall k \in 1, ..., n$.
The corresponding statistical model is:

$$\mathcal{M}^n = (\mathbb{N}^n, \mathcal{P}(\mathbb{N}^n),\ \{\mathbb{P}_\theta^n,\ \theta \in \Theta\})$$

with $\mathbb{P}_\theta^n = \mathbb{P}_\theta \otimes ... \otimes \mathbb{P}_\theta$ (n times)
We are trying to estimate the parameter θ.

Question 4

The likelihood function is the function on θ that makes our n observations most likely.

Using the independance of the X_k:

$$l(\theta) = \prod_{k=1}^{n} e^{-\theta} \frac{\theta^{X_k}}{X_k!}$$

$$L(\theta) = log(l(\theta)) = \sum_{k=1}^{n}(-\theta + X_k log(\theta) - log(X_k!)) = -n\theta + log(\theta)\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} log(X_k!)$$

By derivating with respect to θ, we have:

$$L'(\theta) = -n + \frac{\sum_{k=1}^{n} X_k}{\theta}$$
$$L''(\theta) = -\frac{\sum_{k=1}^{n} X_k}{\theta^2} < 0$$

Since, the second derivative of the log-likelihood function is negative, the function is concave and admits a global maximum given by:

$$L'(\theta) = 0 \Leftrightarrow -n + \frac{\sum_{k=1}^{n} X_k}{\theta} = 0 \Leftrightarrow \hat{\theta}_{MLE} = \overline{X}$$

So, the maximum likelihood estimator is:

$$\hat{\theta}_{MLE} = \overline{X}$$

Question 5

Since the X_k are iid, we have that:

$$\mathbb{E}[\overline{X}] = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[X_k] = \mathbb{E}[X_1] = \theta$$

$$\mathbb{V}(\overline{X}) = \frac{1}{n^2}\sum_{k=1}^{n} \mathbb{V}(X_k) = \frac{1}{n}\mathbb{V}[X_1] = \frac{\theta}{n}$$

Applying the central limit theorem, we have that $\sqrt{n}(\hat{\theta}_{MLE} - \theta)$ converges towards a Gaussian $\mathcal{N}(0, \theta)$.

Question 6

The weak law of large numbers gives us that:

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta$$

By continuous mapping, we have:

$$\sqrt{\hat{\theta}_{MLE}} \xrightarrow{p} \sqrt{\theta}$$

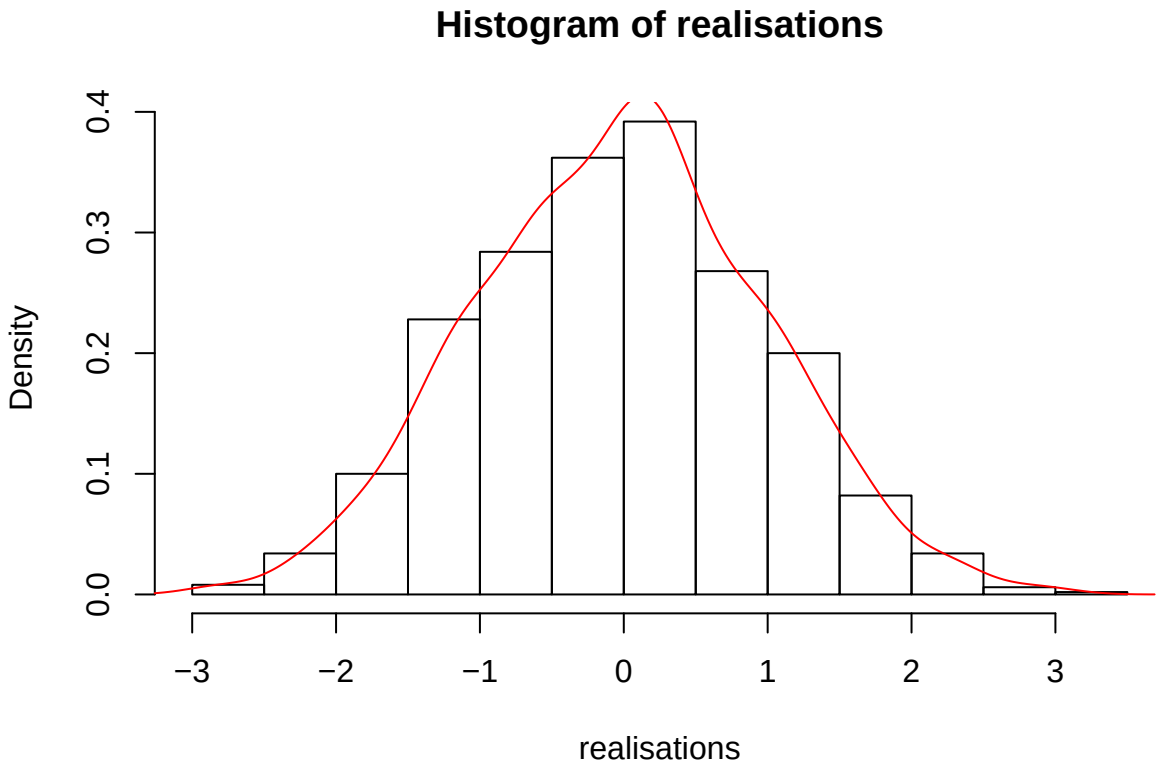
Then, by applying Slutsky's theorem, we have that:

$$\sqrt{n}\frac{(\hat{\theta}_{MLE} - \theta)}{\sqrt{\hat{\theta}_{MLE}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Now, let's check this result in R by simulating 1000 times our random variable $\sqrt{n}\frac{(\hat{\theta}_{MLE} - \theta)}{\sqrt{\hat{\theta}_{MLE}}}$ with a sample size of 100:

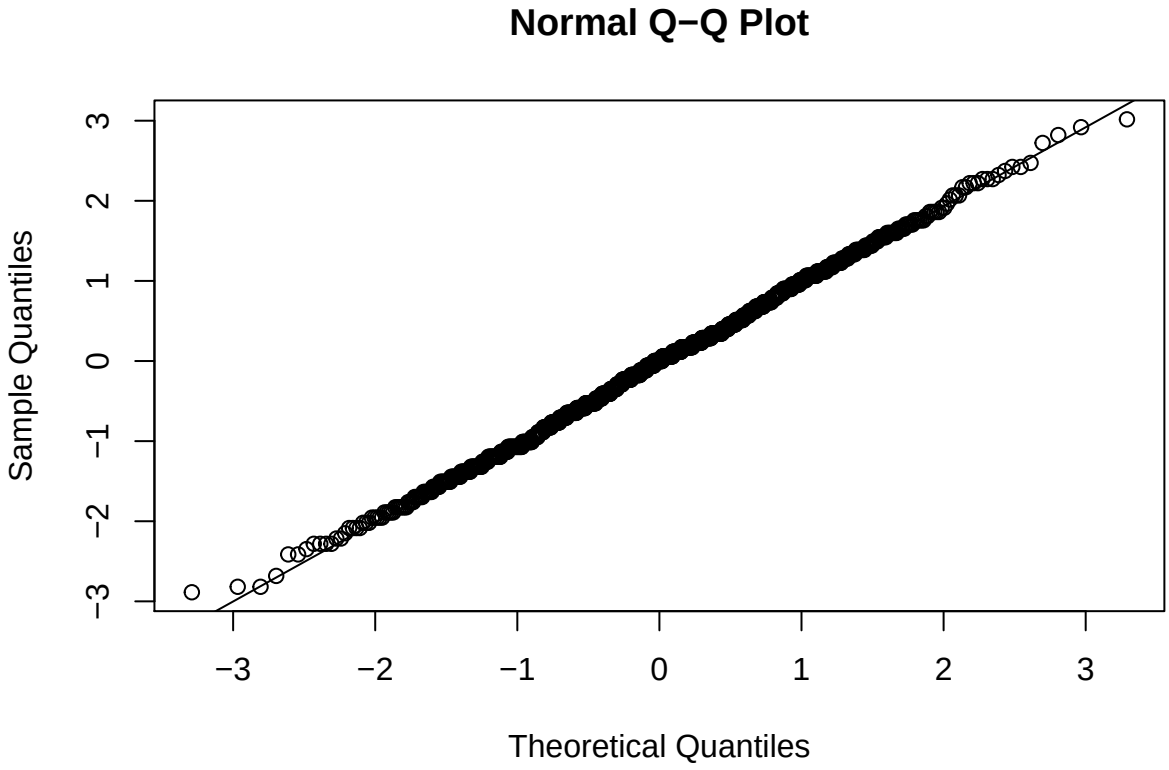
```r
estim <- function(x, theta){
  n <- length(x)
  est <- sqrt(n) * (mean(x) - theta) / sqrt(mean(x))
  return(est)
  }
```

```r
set.seed(23)
Nattempts = 1e3
nsample = 100
theta = 3

samples <- lapply(1:Nattempts, function(i) rpois(nsample, theta))
realisations <- sapply(samples, function(x) estim(x, theta))

hist(realisations, probability = TRUE)
d = density(realisations, kernel='gaussian')
lines(d, col = 'red')
```



The histogram confirms what we found theoretically. In fact, by plotting the density associated to the histogram we can observe a curve that represents a gaussian distribution. It is symetric around its expectation that seems to be zero. So, we can conclude that the random variable $\sqrt{n}\frac{(\hat{\theta}_{MLE}-\theta)}{\sqrt{\hat{\theta}_{MLE}}}$ follows a standard gaussian distribution.

```r
qqnorm(realisations)
qqline(realisations)
```



The Q-Q plot compares the theoritical quantiles of a standard gaussian distribution to the ones of our estimated distribution. We can observe that the points approximately lie on the line y = x, so the distributions compared are similar and this plot also confirms that the random variable $\sqrt{n}\frac{(\hat{\theta}_{MLE}-\theta)}{\sqrt{\hat{\theta}_{MLE}}}$ follows a standard gaussian distribution.

Question 7

Let $Z_n = \sqrt{n}\frac{(\hat{\theta}_{MLE}-\theta)}{\sqrt{\hat{\theta}_{MLE}}}$ be our random variable.

Denote z_α the α-quantile for the standard Normal distribution for $\alpha \in (0,\ 1)$.

$$\lim_{n\to+\infty} \mathbb{P}(-z_{1-\alpha/2} \leq Z_n \leq z_{1-\alpha/2}) \geq 1-\alpha \Leftrightarrow \lim_{n\to+\infty} \mathbb{P}(-z_{1-\alpha/2}\sqrt{\frac{\hat{\theta}_{MLE}}{n}} \leq \hat{\theta}_{MLE}-\theta \leq z_{1-\alpha/2}\sqrt{\frac{\hat{\theta}_{MLE}}{n}}) \geq 1-\alpha$$

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

$$[\hat{\theta}_{MLE} - z_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_{MLE}}}{\sqrt{n}};\ \hat{\theta}_{MLE} + z_{1-\alpha/2}\frac{\sqrt{\hat{\theta}_{MLE}}}{\sqrt{n}}]$$

Question 8

We apply the δ-method with $g(x) = 2\sqrt{x}$
We have: $g'(x) = \frac{1}{\sqrt{x}}$
So,

$$\sqrt{n}(g(\hat{\theta}_{MLE}) - g(\theta)) \xrightarrow{d} \mathcal{N}(0,\ g'(\theta)^2 \times \theta) \Leftrightarrow \sqrt{n}(g(\hat{\theta}_{MLE}) - g(\theta)) \xrightarrow{d} \mathcal{N}(0,1)$$

$$\Leftrightarrow \sqrt{n}(2\sqrt{\hat{\theta}_{MLE}} - 2\sqrt{\theta}) \xrightarrow{d} \mathcal{N}(0,1)$$

Question 9

Let $W_n = \sqrt{n}(2\sqrt{\hat{\theta}_{MLE}} - 2\sqrt{\theta})$ be our random variable.

We know by the last question that $W_n \xrightarrow{d} \mathcal{N}(0,1)$.

$$\lim_{n\to+\infty} \mathbb{P}(-z_{1-\alpha/2} \leq W_n \leq z_{1-\alpha/2}) \geq 1-\alpha \Leftrightarrow \lim_{n\to+\infty} \mathbb{P}(-\frac{z_{1-\alpha/2}}{2\sqrt{n}} \leq \sqrt{\hat{\theta}_{MLE}} - \sqrt{\theta} \leq \frac{z_{1-\alpha/2}}{2\sqrt{n}}) \geq 1-\alpha$$

$$\Leftrightarrow \lim_{n\to+\infty} \mathbb{P}(\sqrt{\hat{\theta}_{MLE}} - \frac{z_{1-\alpha/2}}{2\sqrt{n}} \leq \sqrt{\theta} \leq \sqrt{\hat{\theta}_{MLE}} + \frac{z_{1-\alpha/2}}{2\sqrt{n}}) \geq 1-\alpha$$

When n goes towards infinity, $\frac{z_{1-\alpha/2}}{2\sqrt{n}}$ goes to 0. Since $\sqrt{\hat{\theta}_{MLE}}$ is positive, there exists a n_0 such that $\forall n \geq n_0$, $\sqrt{\hat{\theta}_{MLE}} - \frac{z_{1-\alpha/2}}{2\sqrt{n}}$ is positive and we can take the squares in the inequality without changing the order of the inequalities:

$$\Leftrightarrow \lim_{n\to+\infty} \mathbb{P}((\sqrt{\hat{\theta}_{MLE}} - \frac{z_{1-\alpha/2}}{2\sqrt{n}})^2 \leq \theta \leq (\sqrt{\hat{\theta}_{MLE}} + \frac{z_{1-\alpha/2}}{2\sqrt{n}})^2) \geq 1-\alpha$$

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

$$[(\sqrt{\hat{\theta}_{MLE}} - \frac{z_{1-\alpha/2}}{2\sqrt{n}})^2;\ (\sqrt{\hat{\theta}_{MLE}} + \frac{z_{1-\alpha/2}}{2\sqrt{n}})^2]$$

Question 10

Based on the first moment of a poisson distribution, we easily have that:

$$\hat{\theta}_{MME} = \overline{X}$$

We can remark that $\hat{\theta}_{MME} = \hat{\theta}_{MLE}$

Based on the second moment of a poisson distribution, we have:

$$n^{-1}\sum_{k=1}^{n} X_k^2 = \hat{\theta}_2(\hat{\theta}_2 + 1)$$

Let's define the function $h(x) = x(x+1)$
Its inverse on $\mathbb{R}_+^*$ is $h^{-1}(x) = \frac{1}{2}[-1 + \sqrt{4x+1}]$ and this gives us another estimator of θ:

$$\hat{\theta}_2 = \frac{1}{2}[-1 + \sqrt{(4n^{-1}\sum_{k=1}^{n} X_k^2) + 1}]$$

Question 11

$\mathbb{E}[\hat{\theta}_{MLE}] = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[X_k]$ by linearity of the expectation. So,

$$\mathbb{E}[\hat{\theta}_{MLE}] = \frac{1}{n} * n * \theta = \theta$$

Therefore, $\hat{\theta}_{MLE}$ is an unbiased estimator of θ, ie. ${b_\theta}^*(\hat{\theta}_{MLE}) = 0$

$\mathbb{V}(\hat{\theta}_{MLE}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(X_i)$ by independance of the X_k.

$$\mathbb{V}(\hat{\theta}_{MLE}) = \frac{1}{n^2} * n * \theta = \frac{\theta}{n}$$

The quadratic risk Q is:

$$Q = {b_\theta}^*(\hat{\theta}_{MLE})^2 + \mathbb{V}^*(\hat{\theta}_{MLE}) = 0 + \frac{\theta}{n} = \frac{\theta}{n}$$

Question 12

$\hat{\theta}_{MLE}$ is an unbiased estimator. So the Cramer-Rao bound is given by:

$$\frac{1}{I_n(\theta^*)} = \frac{1}{\mathbb{E}[-L''(\theta^*)]}$$

By derivating the log-likelihood function with respect to θ, we have:

$$L'(\theta^*) = -n + \frac{\sum_{i=1}^{n} X_k}{\theta}$$

$$-L''(\theta^*) = \frac{\sum_{i=1}^{n} X_k}{\theta^2}$$

Therefore,

$$\mathbb{E}[-L''(\theta^*)] = \frac{\sum_{i=1}^{n} \mathbb{E}[X_k]}{\theta^2} = \frac{n}{\theta}$$

Finally,

$$\frac{1}{I_n(\theta^*)} = \frac{\theta}{n} = \mathbb{V}(\hat{\theta}_{MLE})$$

We can conclude that our estimator $\hat{\theta}_{MLE}$ is efficient.

Question 13

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

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2 + (\theta - \overline{X_n})^2 + \frac{2}{n}(\theta - \overline{X_n})\sum_{i=1}^{n}(X_i - \theta) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2 + (\theta - \overline{X_n})^2 + 2(\theta - \overline{X_n})(\overline{X_n} - \theta)$$

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

Question 14

$$\mathbb{E}[(\theta - \overline{X_n})^2] = \mathbb{E}[\theta^2 - 2\theta\overline{X_n} + \overline{X_n}^2] = \theta^2 - 2\theta\mathbb{E}[\overline{X_n}] + \mathbb{E}[\overline{X_n^2}]$$

$$= -\theta^2 + \mathbb{V}(\overline{X_n}) + \mathbb{E}[\overline{X_n}^2] = -\theta^2 + \frac{\theta}{n} + \theta^2 = \frac{\theta}{n}$$

$$\mathbb{E}[\hat{\theta}_2] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2 - (\theta - \overline{X_n})^2] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(X_i - \theta)^2] - \mathbb{E}[(\theta - \overline{X_n})^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{V}(X_i) - \frac{\theta}{n} = \theta(1 - \frac{1}{n})$$

Therefore the bias is:

$$b_{\hat{\theta}_2} = -\frac{\theta}{n}$$

We can get an unbiased estimator $\hat{\theta}_3$ by defining $\hat{\theta}_3 = (1 - \frac{1}{n})^{-1}\hat{\theta}_2$

Question 15

Using the previous questions, we know that:

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

therefore, we have:

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

where:

$$\forall i \in [\![1, n]\!], Yi = (X_i - \theta)^2$$

$$\bar{Y_n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Since:

$$\mathbb{E}[Y_i] = \mathbb{V}(X_i) = \theta$$

$$\mathbb{V}(Y_i) = 2\theta^2 + \theta$$

We can apply the central limit theorem, and we have that:

$$\sqrt{n}(\overline{Y_n} - \theta) \xrightarrow{d} \mathcal{N}(0, 2\theta^2 + \theta)$$

We also have the following equalities:

$$\sqrt{n}(\theta - \overline{X_n})^2 = \sqrt{n}(\overline{X_n} - \theta)^2 = \sqrt{n}(\overline{X_n} - \theta)(\overline{X_n} - \theta)$$

By applying the central limit theorem, we now have:

$$\sqrt{n}(\overline{X_n} - \theta) \xrightarrow{d} \mathcal{N}(0, \theta)$$

On the other hand, applying the law of large numbers gives us:

$$(\theta - \overline{X_n}) \xrightarrow{p} 0$$

Then, by applying Slutsky's theorem, we have that $\sqrt{n}(\theta - \overline{X_n})^2$ converges in distribution towards the constant 0. Therefore, it converges in probability towards 0.

Now, we can apply Slutsky's theorem to $\sqrt{n}(\overline{Y_n} - \theta) - \sqrt{n}(\theta - \overline{X_n})^2$ which gives us finally that:

$$\sqrt{n}(\hat{\theta}_2 - \theta) \xrightarrow{d} \mathcal{N}(0, 2\theta^2 + \theta)$$

We can now compute another asymptotic confidence interval centered in $\hat{\theta}_2$.

We know by the first part of question that $\frac{\sqrt{n}(\hat{\theta}_2 - \theta)}{\sqrt{2\theta^2 + \theta}} \xrightarrow{d} \mathcal{N}(0, 1)$.

Let's use $\hat{\theta}_2$ as an estimator of θ for the denominator.
First, in order to apply slutsky's theorem, we need to prove that $\hat{\theta}_2 \xrightarrow{p} \theta$.

Let $\epsilon > 0$. By Chebyshev's inequality,

$$P(\mid \hat{\theta}_2 - \theta \mid > \epsilon) \leq \frac{Var(\hat{\theta}_2 - \theta)}{\epsilon^2} = \frac{\theta^2 + \theta}{n\epsilon^2} \underset{n \to +\infty}{\to} 0$$

So, we have that $\hat{\theta}_2 \xrightarrow{p} \theta$ and by continuous mapping,

$$\sqrt{2\hat{\theta}_2^2 + \hat{\theta}_2} \xrightarrow{p} \sqrt{2\theta^2 + \theta}$$

Then, by Slutsky's theorem, we have that $V_n = \sqrt{n}\frac{(\hat{\theta}_2 - \theta)}{\sqrt{2\hat{\theta}_2^2 + \hat{\theta}_2}}$ converges in law towards a gaussian $\mathcal{N}(0,1)$, and we can use it as pivotal quantity to find a confidence interval for θ.

$$\lim_{n \to +\infty} \mathbb{P}(-z_{1-\alpha/2} \leq V_n \leq z_{1-\alpha/2}) \geq 1-\alpha \Leftrightarrow \lim_{n \to +\infty} \mathbb{P}(-z_{1-\alpha/2}\frac{\sqrt{2\hat{\theta}_2^2 + \hat{\theta}_2}}{\sqrt{n}} \leq \hat{\theta}_2 - \theta \leq z_{1-\alpha/2}\frac{\sqrt{2\hat{\theta}_2^2 + \hat{\theta}_2}}{\sqrt{n}}) \geq 1-\alpha$$

$$\Leftrightarrow \lim_{n \to +\infty} \mathbb{P}(\hat{\theta}_2 - z_{1-\alpha/2}\frac{\sqrt{2\hat{\theta}_2^2 + \hat{\theta}_2}}{\sqrt{n}} \leq \theta \leq \hat{\theta}_2 + z_{1-\alpha/2}\frac{\sqrt{2\hat{\theta}_2^2 + \hat{\theta}_2}}{\sqrt{n}}) \geq 1-\alpha$$

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

$$[\hat{\theta}_2 - z_{1-\alpha/2}\frac{\sqrt{2\hat{\theta}_2^2 + \hat{\theta}_2}}{\sqrt{n}};\ \hat{\theta}_2 + z_{1-\alpha/2}\frac{\sqrt{2\hat{\theta}_2^2 + \hat{\theta}_2}}{\sqrt{n}}]$$

This asymptotic interval has a bigger range compared to the first one

Furthermore, we have that:

$$var(\hat{\theta}_2) = \frac{2\theta^2 + \theta}{n} = \frac{\theta(2\theta+1)}{n} = (2\theta+1)Var(\hat{\theta}_{MLE})$$

Question 16

Let $s \in \mathbb{R}$. The probability generating function of the Poisson distribution is given by:

$$G_{\mathbb{X}}(s) = \mathbb{E}[exp(s\mathbb{X})] = \sum_{k=0}^{\infty} e^{ks}e^{-\theta}\frac{\theta^k}{k!} = e^{-\theta}\sum_{k=0}^{\infty}\frac{(\theta e^s)^k}{k!} = e^{-\theta}e^{\theta e^s} = e^{\theta(e^s-1)}$$

In order to compute the first and second moment of the Poisson distribution, we can now use the moment generating function. Let's compute its first and second order derivatives.

$$G'_{\mathbb{X}}(s) = \theta e^s e^{\theta(e^s-1)}$$
$$G''_{\mathbb{X}}(s) = \theta[e^s e^{\theta(e^s-1)} + \theta e^{2s}e^{\theta(e^s-1)}] = \theta e^s[e^{\theta(e^s-1)} + \theta e^s e^{\theta(e^s-1)}]$$

Then, we have:

$$\mathbb{E}[\mathbb{X}] = G'_{\mathbb{X}}(0) = \theta$$
$$\mathbb{E}[\mathbb{X}^2] = G''_{\mathbb{X}}(0) = \theta(1+\theta)$$
$$\mathbb{V}(\mathbb{X}) = \mathbb{E}[\mathbb{X}^2] - \mathbb{E}[\mathbb{X}]^2 = \theta(1+\theta) - \theta^2 = \theta$$

We will now show that: $\mathbb{V}[(\mathbb{X}_i - \theta)^2] = 2\theta^2 + \theta$

$$G_{\mathbb{X}}^{(3)}(s) = (1 + 3\theta e^s + \theta^2 e^{2s})\theta e^{s+\theta(e^s-1)}$$

$$G_{\mathbb{X}}^{(4)}(s) = (1 + \theta^3 e^{3s} + 6\theta^2 e^{2s} + 7\theta e^s)\theta e^{s+\theta(e^s-1)}$$

$$\mathbb{E}[\mathbb{X}^3] = G_{\mathbb{X}}^{(3)}(0) = \theta + 3\theta^2 + \theta^3$$

$$\mathbb{E}[\mathbb{X}^4] = G_{\mathbb{X}}^{(4)}(0) = \theta^4 + 6\theta^3 + 7\theta^2 + \theta$$

$$\mathbb{V}[(\mathbb{X}_i - \theta)^2] = \mathbb{E}[(\mathbb{X} - \theta)^4] - \mathbb{E}[(\mathbb{X} - \theta)^2]^2 = \mathbb{E}[\mathbb{X}^4] - 4\theta\mathbb{E}[\mathbb{X}^3] + 6\theta^2\mathbb{E}[\mathbb{X}^2] - 4\theta^3\mathbb{E}[\mathbb{X}] + \theta^4 - Var(\mathbb{X})^2$$

$$= \theta^4 + 6\theta^3 + 7\theta^2 + \theta - 4\theta(\theta + 3\theta^2 + \theta^3) + 6\theta^2(\theta + \theta^2) - 4\theta^4 + \theta^4 - \theta^2 = 2\theta^2 + \theta$$

Problem 2: Analysis of the USJudgeRatings dataset

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.

Introduction

We are given to analyse a dataset, named USJudgeratings, containing various ratings of state judges in the US Superior Court made by lawyers. The different variables given help us to determine if a judge is worthy staying in the US Superior Court or not. We will start by doing a general description of the data and applying descriptive statistics to better apprehend the data.

General description

We start by uploading our data.

```
data(USJudgeRatings)
```

First, let's see how the dataset is organized.

```
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 ...
```

The data is stored in a dataframe nad we can observe that all the variables are numeric.

```
dim(USJudgeRatings)
```

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

We are provided with n = 43 observations and p = 12 quantitative variables.

We can have a full view of the dataset by using the kable function:

```
library(knitr)
library(kableExtra)
kable(USJudgeRatings, 'latex', caption = "Ratings of US judges", booktabs = T) %>%
  kable_styling(latex_options = "striped", font_size = 6)
```

Table 1: Ratings of US judges

	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
CALLAHAN,R.J.	10.6	9.0	8.9	8.7	8.5	8.5	8.5	8.5	8.6	8.4	9.1	9.0
COHEN,S.S.	7.0	5.9	4.9	5.1	5.4	5.9	4.8	5.1	4.7	4.9	6.8	5.0
DALY,J.J.	7.3	8.9	8.9	8.7	8.6	8.5	8.4	8.4	8.4	8.5	8.8	8.8
DANNEHY,J.F.	8.2	7.9	6.7	8.1	7.9	8.0	7.9	8.1	7.7	7.8	8.5	7.9
DEAN,H.H.	7.0	8.0	7.6	7.4	7.3	7.5	7.1	7.2	7.1	7.2	8.4	7.7
DEVITA,H.J.	6.5	8.0	7.6	7.2	7.0	7.1	6.9	7.0	7.0	7.1	6.9	7.2
DRISCOLL,P.J.	6.7	8.6	8.2	6.8	6.9	6.6	7.1	7.3	7.2	7.2	8.1	7.7
GRILLO,A.E.	7.0	7.5	6.4	6.8	6.5	7.0	6.6	6.8	6.3	6.6	6.2	6.5
HADDEN,W.L.JR.	6.5	8.1	8.0	8.0	7.9	8.0	7.9	7.8	7.8	7.8	8.4	8.0
HAMILL,E.C.	7.3	8.0	7.4	7.7	7.3	7.3	7.3	7.2	7.1	7.2	8.0	7.6
HEALEY.A.H.	8.0	7.6	6.6	7.2	6.5	6.5	6.8	6.7	6.4	6.5	6.9	6.7
HULL,T.C.	7.7	7.7	6.7	7.5	7.4	7.5	7.1	7.3	7.1	7.3	8.1	7.4
LEVINE,I.	8.3	8.2	7.4	7.8	7.7	7.7	7.7	7.8	7.5	7.6	8.0	8.0
LEVISTER,R.L.	9.6	6.9	5.7	6.6	6.9	6.6	6.2	6.0	5.8	5.8	7.2	6.0
MARTIN,L.F.	7.1	8.2	7.7	7.1	6.6	6.6	6.7	6.7	6.8	6.8	7.5	7.3
MCGRATH,J.F.	7.6	7.3	6.9	6.8	6.7	6.8	6.4	6.3	6.3	6.3	7.4	6.6
MIGNONE,A.F.	6.6	7.4	6.2	6.2	5.4	5.7	5.8	5.9	5.2	5.8	4.7	5.2
MISSAL,H.M.	6.2	8.3	8.1	7.7	7.4	7.3	7.3	7.3	7.2	7.3	7.8	7.6
MULVEY,H.M.	7.5	8.7	8.5	8.6	8.5	8.4	8.5	8.5	8.4	8.4	8.7	8.7
NARUK,H.J.	7.8	8.9	8.7	8.9	8.7	8.8	8.9	9.0	8.8	8.9	9.0	9.0
O'BRIEN,F.J.	7.1	8.5	8.3	8.0	7.9	7.9	7.8	7.8	7.8	7.7	8.3	8.2
O'SULLIVAN,T.J.	7.5	9.0	8.9	8.7	8.4	8.5	8.4	8.3	8.3	8.3	8.8	8.7
PASKEY,L.	7.5	8.1	7.7	8.2	8.0	8.1	8.2	8.4	8.0	8.1	8.4	8.1
RUBINOW,J.E.	7.1	9.2	9.0	9.0	8.4	8.6	9.1	9.1	8.9	9.0	8.9	9.2
SADEN.G.A.	6.6	7.4	6.9	8.4	8.0	7.9	8.2	8.4	7.7	7.9	8.4	7.5
SATANIELLO,A.G.	8.4	8.0	7.9	7.9	7.8	7.8	7.6	7.4	7.4	7.4	8.1	7.9
SHEA,D.M.	6.9	8.5	7.8	8.5	8.1	8.2	8.4	8.5	8.1	8.3	8.7	8.3
SHEA,J.F.JR.	7.3	8.9	8.8	8.7	8.4	8.5	8.5	8.5	8.4	8.4	8.8	8.8
SIDOR,W.J.	7.7	6.2	5.1	5.6	5.6	5.9	5.6	5.6	5.3	5.5	6.3	5.3
SPEZIALE,J.A.	8.5	8.3	8.1	8.3	8.4	8.2	8.2	8.1	7.9	8.0	8.0	8.2
SPONZO,M.J.	6.9	8.3	8.0	8.1	7.9	7.9	7.9	7.7	7.6	7.7	8.1	8.0
STAPLETON,J.F.	6.5	8.2	7.7	7.8	7.6	7.7	7.7	7.7	7.5	7.6	8.5	7.7
TESTO,R.J.	8.3	7.3	7.0	6.8	7.0	7.1	6.7	6.7	6.7	6.7	8.0	7.0
TIERNEY,W.L.JR.	8.3	8.2	7.8	8.3	8.4	8.3	7.7	7.6	7.5	7.7	8.1	7.9
WALL,R.A.	9.0	7.0	5.9	7.0	7.0	7.2	6.9	6.9	6.5	6.6	7.6	6.6
WRIGHT,D.B.	7.1	8.4	8.4	7.7	7.5	7.7	7.8	8.2	8.0	8.1	8.3	8.1
ZARRILLI,K.J.	8.6	7.4	7.0	7.5	7.5	7.7	7.4	7.2	6.9	7.0	7.8	7.1

An observation in this dataset represents the different ratings received by a judge (given by his name) in the US Superior Court. In order to study this dataset, we will first define properly what each variable means.

```
colnames(USJudgeRatings)
```

```
##  [1] "CONT" "INTG" "DMNR" "DILG" "CFMG" "DECI" "PREP" "FAMI" "ORAL" "WRIT"
## [11] "PHYS" "RTEN"
```

The variables are:

- *CONT* : The number of contacts of the lawyer with judge.
- *INTG* : The judicial integrity of the judge.
- *DMNR* : Demeanor of the judge.
- *DILG* : Diligence of the judge.
- *CFMG* : Case flow managed by the judge.
- *DECI* : Prompt decisions taken by the judge.
- *PREP* : How the judge is prepared trial.
- *FAMI* : The judge's familiarity with law.
- *ORAL* : The judge's sound oral rulings.
- *WRIT* : The judge's sound written rulings.
- *PHYS* : The judge's physical ability.
- *RTEN* : Scaling if the judge is worthy to retain in the US Superior court.

Descriptive dataset analysis

Let's inspect the dataframe for missing values, outliers and errors:

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

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

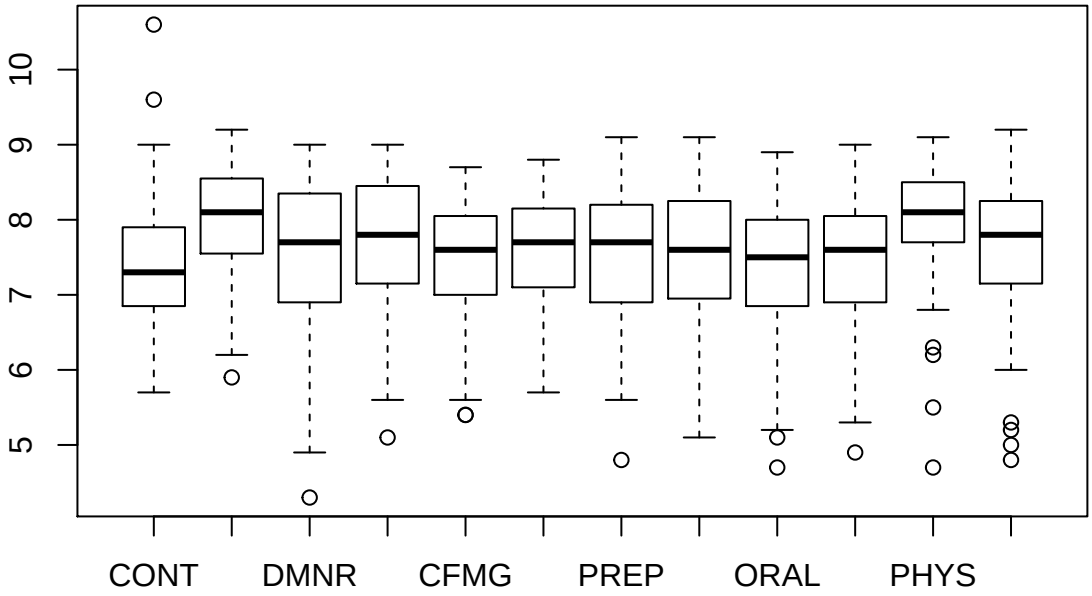
There are no missing values in the dataframe.

```
summary(USJudgeRatings)
```

```
##       CONT             INTG            DMNR            DILG      
##  Min.   : 5.700   Min.   :5.900   Min.   :4.300   Min.   :5.100  
##  1st Qu.: 6.850   1st Qu.:7.550   1st Qu.:6.900   1st Qu.:7.150  
##  Median : 7.300   Median :8.100   Median :7.700   Median :7.800  
##  Mean   : 7.437   Mean   :8.021   Mean   :7.516   Mean   :7.693  
##  3rd Qu.: 7.900   3rd Qu.:8.550   3rd Qu.:8.350   3rd Qu.:8.450  
##  Max.   :10.600   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.:6.950  
##  Median :7.600   Median :7.700   Median :7.700   Median :7.600  
##  Mean   :7.479   Mean   :7.565   Mean   :7.467   Mean   :7.488  
##  3rd Qu.:8.050   3rd Qu.:8.150   3rd Qu.:8.200   3rd Qu.:8.250  
##  Max.   :8.700   Max.   :8.800   Max.   :9.100   Max.   :9.100  
##       ORAL            WRIT            PHYS            RTEN      
##  Min.   :4.700   Min.   :4.900   Min.   :4.700   Min.   :4.800  
##  1st Qu.:6.850   1st Qu.:6.900   1st Qu.:7.700   1st Qu.:7.150  
##  Median :7.500   Median :7.600   Median :8.100   Median :7.800  
##  Mean   :7.293   Mean   :7.384   Mean   :7.935   Mean   :7.602  
##  3rd Qu.:8.000   3rd Qu.:8.050   3rd Qu.:8.500   3rd Qu.:8.250  
##  Max.   :8.900   Max.   :9.000   Max.   :9.100   Max.   :9.200
```

All the variables (except the variable CONT) are ranged between 0 and 10. Furthermore, we can observe that each variable admits a median that is close to the mean. For example, the median for the INTG variable is 8.021 and its median is equal to 8.100. So the variables seem to follow symetric distributions.
We will now take a look on the boxplots of the variables to see if there is outliers and errors, and to compare the interquartile ranges.

```
Outvals = boxplot(USJudgeRatings)
```



We observe the presence of outliers for 10 of the 12 variables (with larger values for CONT and with lower values for the other variables). Most of the variables have only one or two outliers, except the variables PHYS and RTEN, which have four outliers each. We are not provided with extra information and nothing indicated that these outliers correspond to mistakes. Thus, we will assume that they aren't mistakes and keep them in our analysis.
Note that the PHYS variable has a small interquartile range compare to the other variables. That means that all the judges seem to have a good physical ability (superior to 7) except the four outliers.
Also, the four outliers for the RTEN variable mean that the laywers think that these four judges should not stay in the US Superior court. Therefore, we need to see if there a link between the other variable and the variable RTEN.

Let's take a closer look at these outliers.

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

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

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

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

The judge with the biggest number of contacts of lawyer is judge Callahan with a a number of 10.6 contacts.

```
max(USJudgeRatings$RTEN)
```

```
## [1] 9.2
```

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

```
## [1] "RUBINOW,J.E."
```

The judge with the highest rating for worthiness of retention is judge Rubinow with a rating of 9.2.

We can take a look at his other ratings.

```
USJudgeRatings[which.max(USJudgeRatings$RTEN),]
```

```
##              CONT INTG DMNR DILG CFMG DECI PREP FAMI ORAL WRIT PHYS RTEN
## RUBINOW,J.E.  7.1  9.2    9    9  8.4  8.6  9.1  9.1  8.9    9  8.9  9.2
```

```
min(USJudgeRatings$RTEN)
```

```
## [1] 4.8
```

```
rownames(USJudgeRatings)[which.min(USJudgeRatings$RTEN)]
```

```
## [1] "BRACKEN,J.J."
```

The judge with the lowest rating for worthiness of retention is judge Bracken with a rating of 4.8. We can also take a look at his other ratings.

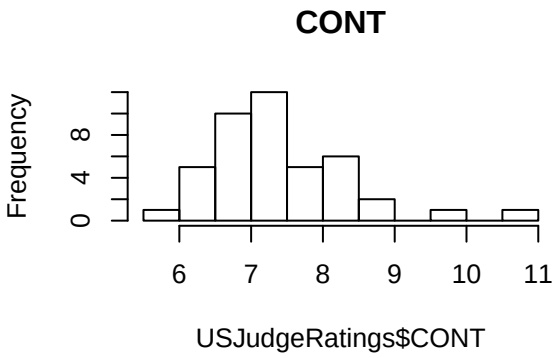
```
USJudgeRatings[which.min(USJudgeRatings$RTEN),]
```

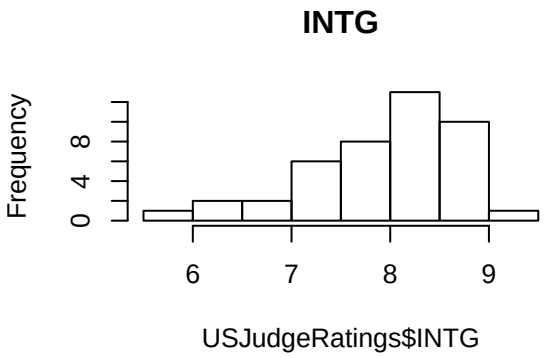
```
##              CONT INTG DMNR DILG CFMG DECI PREP FAMI ORAL WRIT PHYS RTEN
## BRACKEN,J.J.  7.3  6.4  4.3  6.5    6  6.2  5.7  5.7  5.1  5.3  5.5  4.8
```

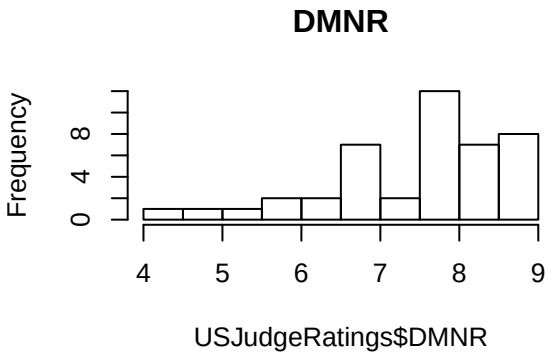
By comparing the ratings of these two judges, we can observe that the judge with the highest RTEN rate has better ratings in the other variables except for the variable CONT. Thus, there seems to be a correlation between the RTEN variable and the others (except CONT). We will look deeper into that correlation later in this report.

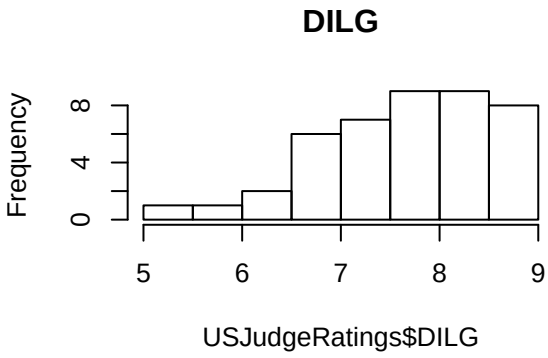
Let's see the distribution of the variable

```
par(mfrow=c(2,2))
hist(USJudgeRatings$CONT, main="CONT")
hist(USJudgeRatings$INTG, main="INTG")
hist(USJudgeRatings$DMNR, main="DMNR")
hist(USJudgeRatings$DILG, main="DILG")
```

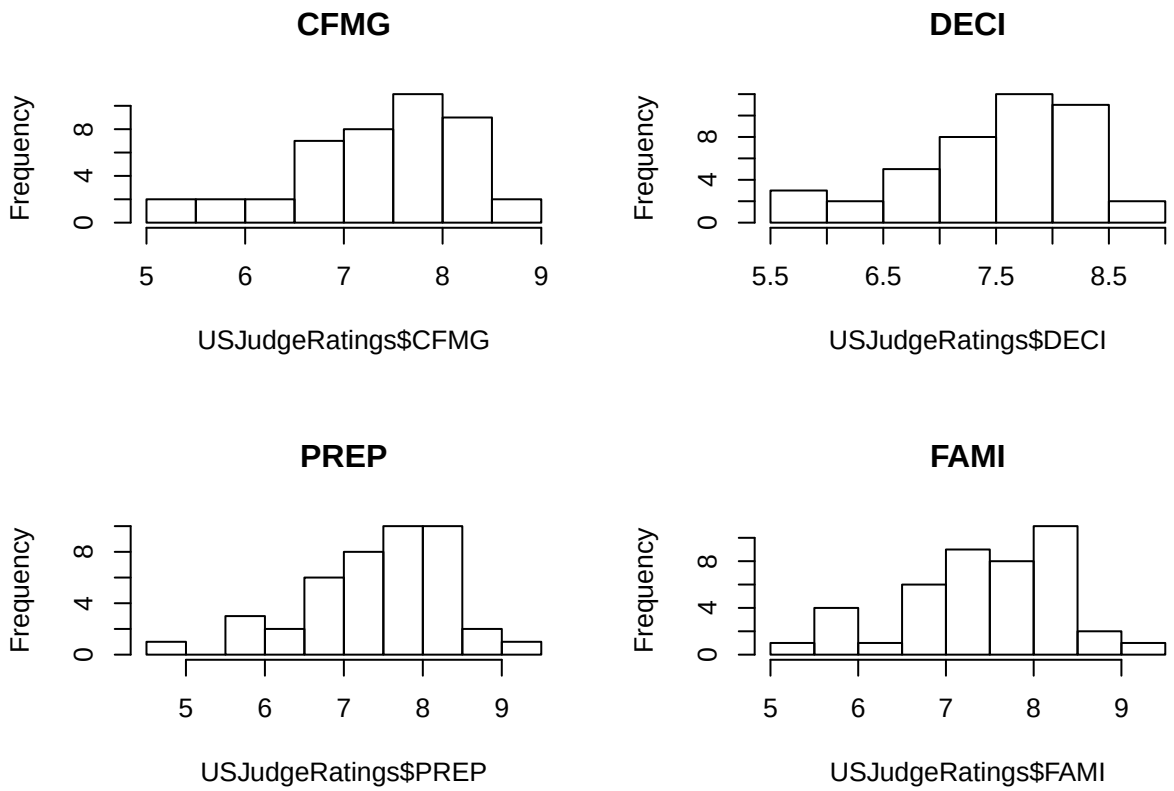




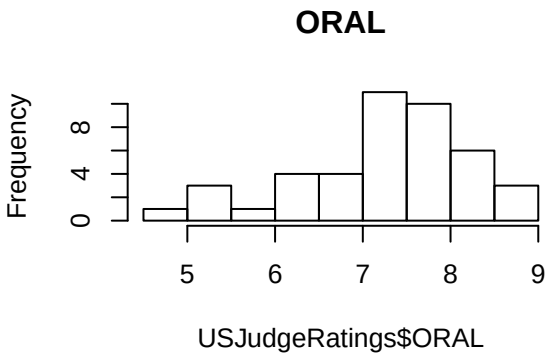


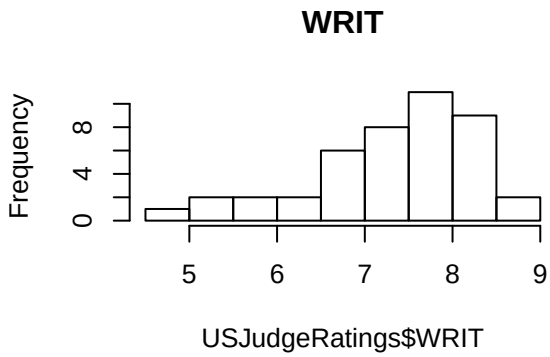


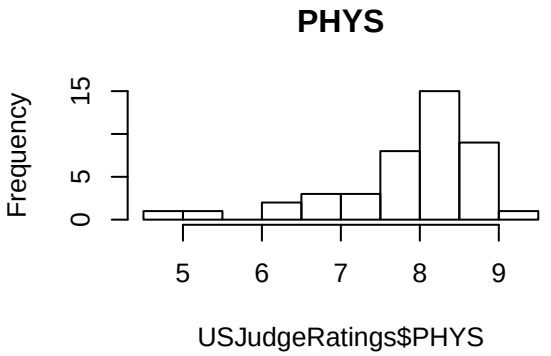

```
par(mfrow=c(2,2))
hist(USJudgeRatings$CFMG, main="CFMG")
hist(USJudgeRatings$DECI, main="DECI")
hist(USJudgeRatings$PREP, main="PREP")
hist(USJudgeRatings$FAMI, main="FAMI")
```

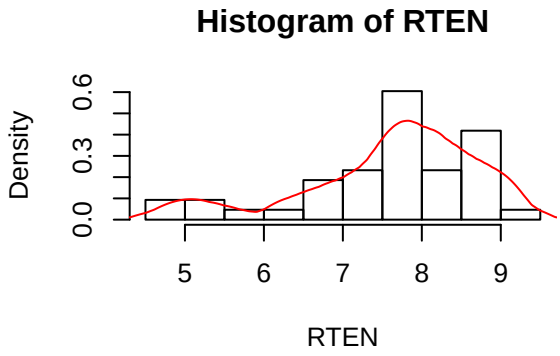



```r
par(mfrow=c(2,2))
hist(USJudgeRatings$ORAL, main="ORAL")
hist(USJudgeRatings$WRIT, main="WRIT")
hist(USJudgeRatings$PHYS, main="PHYS")
hist(USJudgeRatings$RTEN, probability= TRUE, main="Histogram of RTEN" , xlab="RTEN")
d = density(USJudgeRatings$RTEN, kernel = 'o', bw = 0.3)
lines(d, col="red")
```









To comment...

Let's analyze the correlation between the variables

```
round(var(USJudgeRatings), 2)
```

```
##       CONT  INTG  DMNR DILG CFMG DECI PREP  FAMI  ORAL  WRIT PHYS  RTEN
## CONT  0.89 -0.10 -0.17 0.01 0.11 0.07 0.01 -0.02 -0.01 -0.04 0.05 -0.03
## INTG -0.10  0.59  0.85 0.60 0.54 0.50 0.64  0.64  0.71  0.67 0.54  0.79
## DMNR -0.17  0.85  1.31 0.86 0.80 0.74 0.93  0.91  1.05  0.98 0.85  1.19
## DILG  0.01  0.60  0.86 0.81 0.74 0.69 0.84  0.82  0.87  0.83 0.69  0.92
## CFMG  0.11  0.54  0.80 0.74 0.74 0.68 0.79  0.76  0.83  0.78 0.71  0.88
## DECI  0.07  0.50  0.74 0.69 0.68 0.64 0.73  0.72  0.77  0.73 0.66  0.82
## PREP  0.01  0.64  0.93 0.84 0.79 0.73 0.91  0.90  0.95  0.90 0.76  1.00
## FAMI -0.02  0.64  0.91 0.82 0.76 0.72 0.90  0.90  0.94  0.90 0.75  0.98
## ORAL -0.01  0.71  1.05 0.87 0.83 0.77 0.95  0.94  1.02  0.96 0.85  1.09
## WRIT -0.04  0.67  0.98 0.83 0.78 0.73 0.90  0.90  0.96  0.92 0.77  1.02
## PHYS  0.05  0.54  0.85 0.69 0.71 0.66 0.76  0.75  0.85  0.77 0.88  0.94
## RTEN -0.03  0.79  1.19 0.92 0.88 0.82 1.00  0.98  1.09  1.02 0.94  1.21
```

```
round(sqrt(diag(var(USJudgeRatings))),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
```

```
print('The smallest standard deviation is: ')
```

```
## [1] "The smallest standard deviation is: "
```

```
min(round(sqrt(diag(var(USJudgeRatings))),2))
```

```
## [1] 0.77
```

```r
print('The largest standard deviation is: ')
```

```
## [1] "The largest standard deviation is: "
```

```r
max(round(sqrt(diag(var(USJudgeRatings))),2))
```

```
## [1] 1.14
```

Regarding the dispersion, we look at the interquartile range (given by the boxplots) and the empirical standard deviation. Overall, the dispersions are not very high (around 1). We find that the variables DMNR and RTEN have the largest standard deviation, while the DECI variable has the smallest.

Let's measure the correlations between the 11 first variables and the variable RTEN. For this we use the correlations function and the pairs function to visualize the scatter plots of the variables two by two.

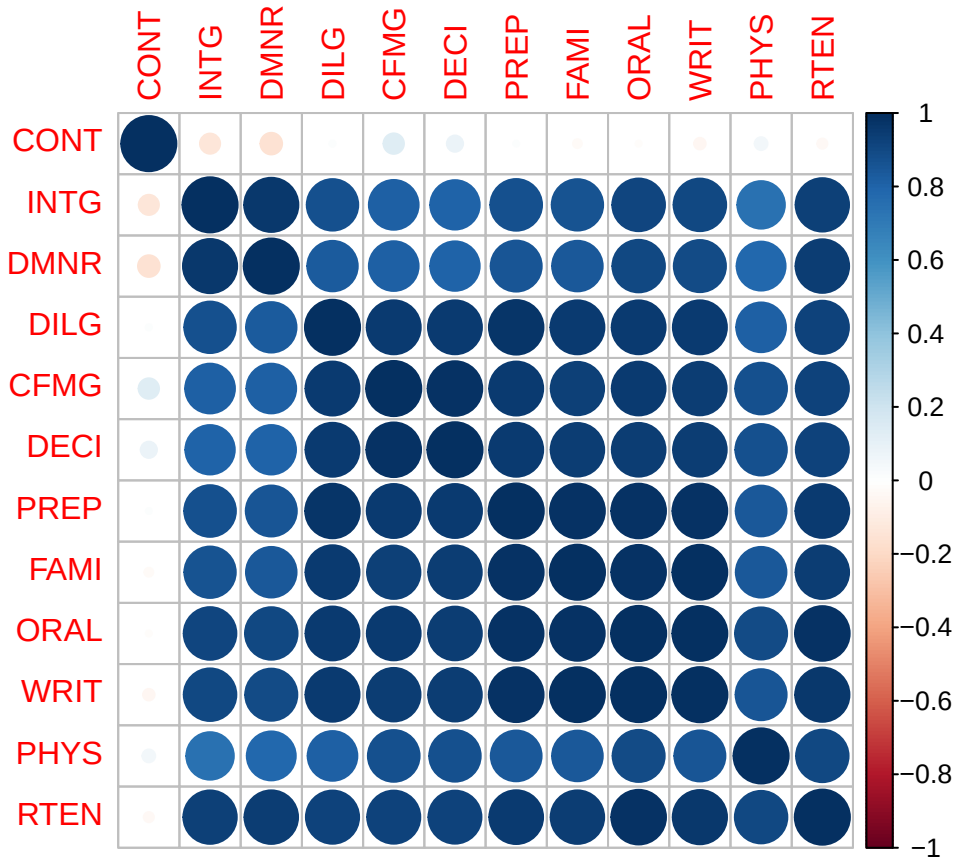
```r
round(cor(USJudgeRatings),2)
```

```
##       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
```

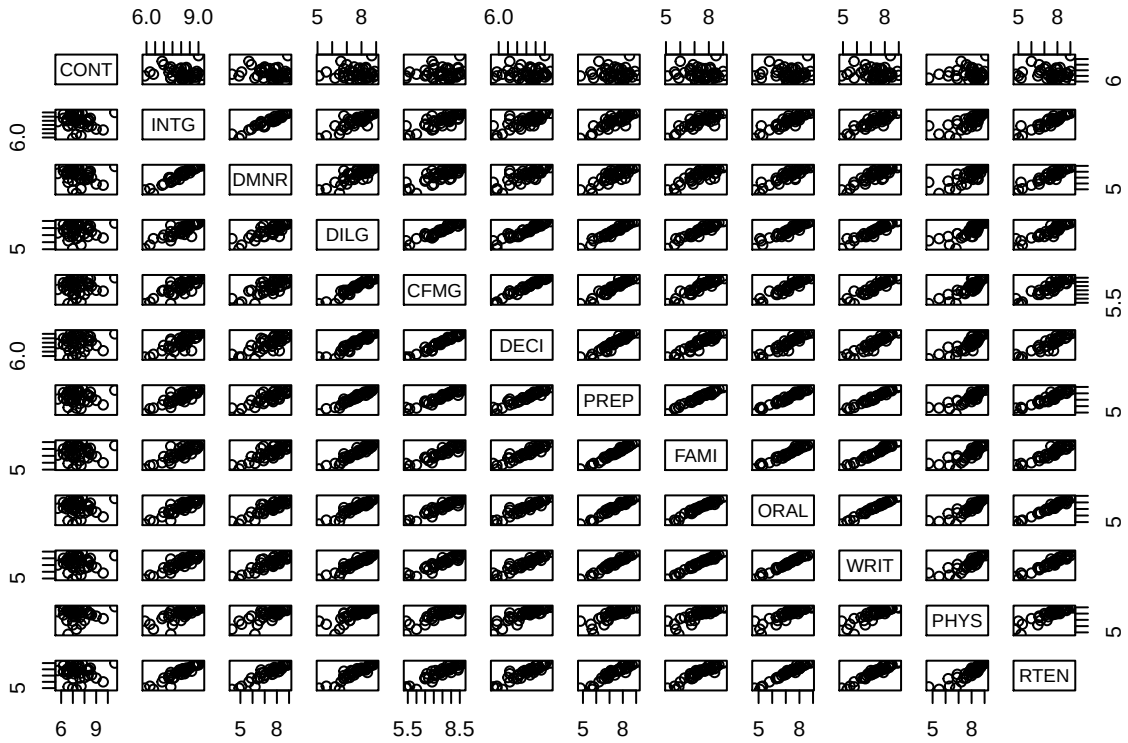
```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

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



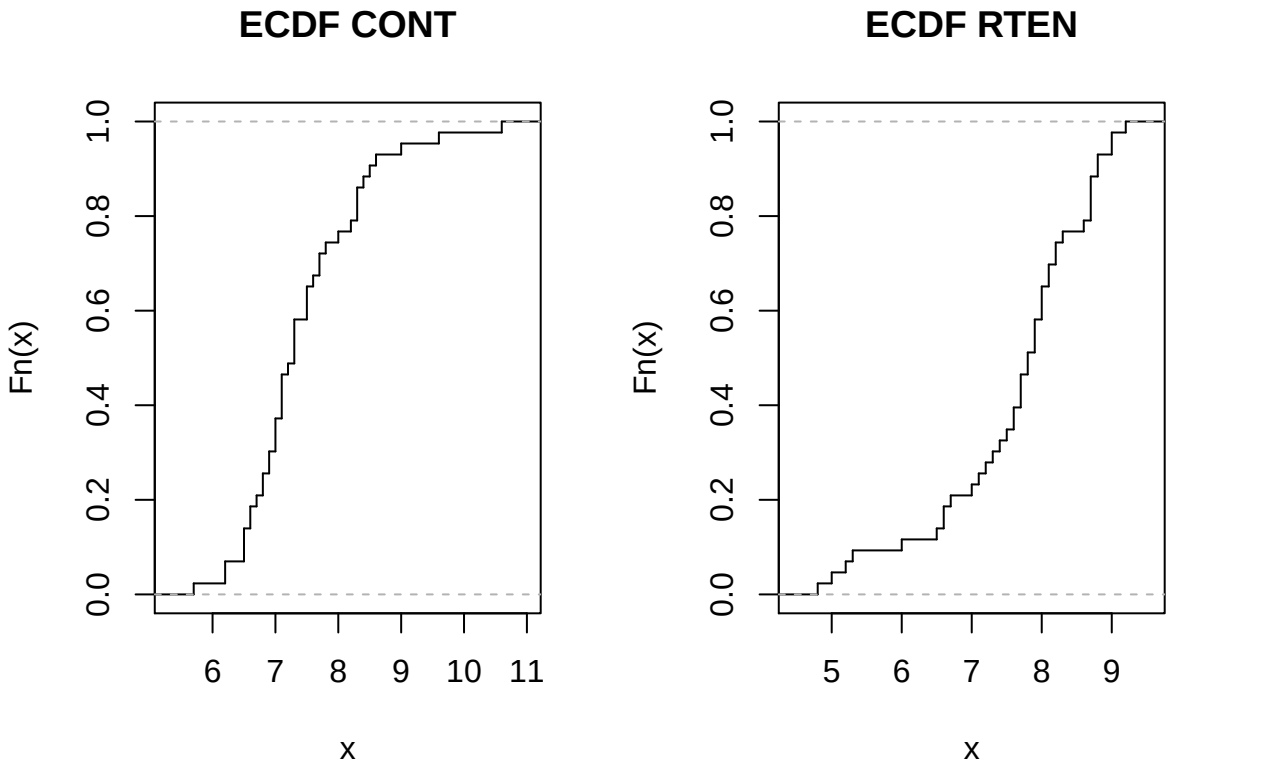

```
pairs(USJudgeRatings)
```



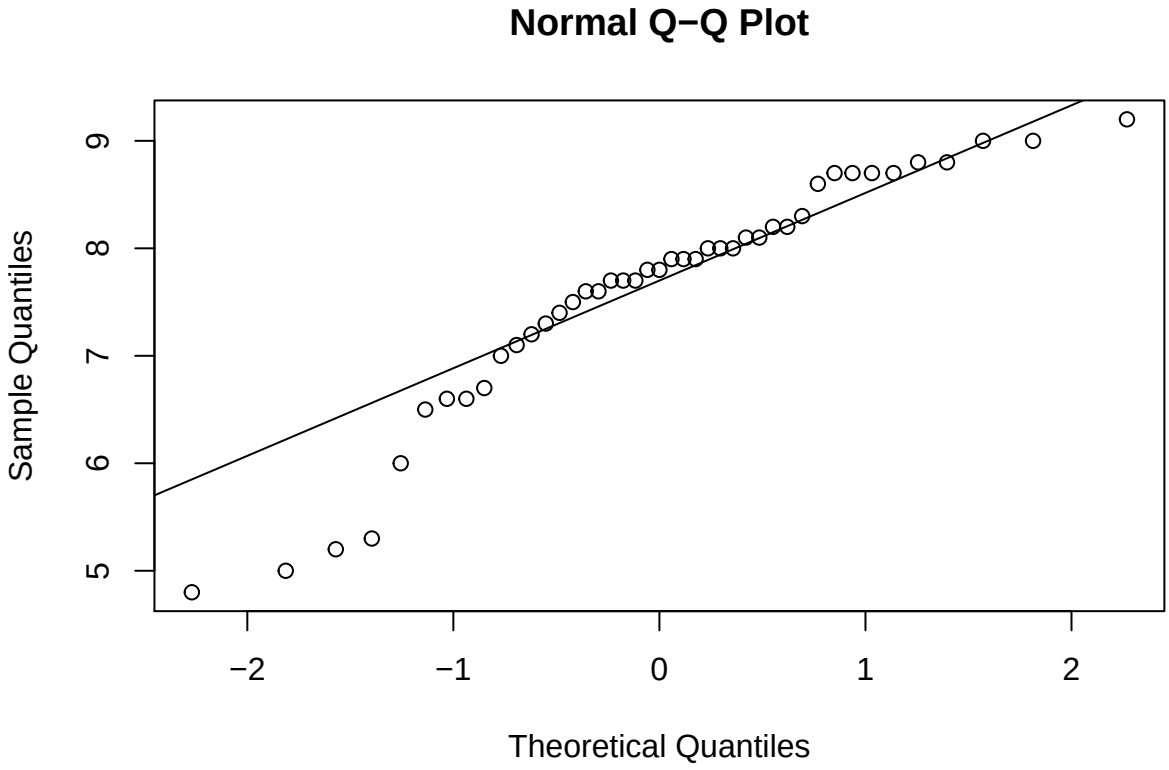
All the variables have a strong positive correlation two by two except the variable CONT which is not correlated to all the other variables. The number of contacts of a lawyer with the judge doesn't seem to explain the

ratings received by the judge.

```
par(mfrow=c(1,2))
plot(ecdf(USJudgeRatings$CONT), verticals = TRUE, do.points = FALSE, main = "ECDF CONT")
plot(ecdf(USJudgeRatings$RTEN), verticals = TRUE, do.points = FALSE, main = "ECDF RTEN")
```




```
qqnorm(USJudgeRatings$RTEN)
qqline(USJudgeRatings$RTEN)
```



The QQ plots suggests that the RTEN variable seems to follow a Gaussian distribution except for lower values.

```
library(e1071)
kurtosis(USJudgeRatings$RTEN)
```

```
## [1] 0.2557421
```

```
skewness(USJudgeRatings$RTEN)
```

```
## [1] -0.9373609
```

Skewness and kurtosis figures for the retention variable are both between -1 and 1 which indicates no substantial skewness or kurtosis.

Conclusion

Our analysis has