— title: “Statistics homework 1” author: “Sarah Jallot and Victoire de Termont” date: “31/10/2019” output: “pdf_document” df_print: paged — ## # Statistics Homework, Assignment 1. # Author: Sarah Jallot and Victoire de Termont.
## Problem 1 : Estimating parameters of a Poisson distribution to model the number of goals scored in football.
### Question 1.
It is a discrete distribution as its support is N^*. The Poisson law is used to describe rare events in a large population. Examples of experiences appropriately modelled by a Poisson distribution are:
i. The number of ice-creams sold on a beach during a hot summer day.
ii. The number of power failures occurring in Oslo in the winter over a period of 100 years.
iii. A restaurant staying empty for a whole night although it normally serves around 1000 people each night.

Question 2.

i. Computing the mean of the Poisson distribution:
$E(X) = \sum_{k=0}^{\infty} k * \frac{e^{-\theta}\theta^k}{k!}$ # summing directly to infinity because we know that the sum converges, but a more proper way would have been to sum to n then do the limit
$\iff E(X) = \theta * \sum_{k=1}^{\infty} \frac{e^{-\theta}\theta^{k-1}}{(k-1)!}$ # simplifying by k as in 0 the sum term is null, and then factorise by lambda
$\iff E(X) = \theta * \sum_{k=0}^{\infty} \frac{e^{-\theta}\theta^k}{(k)!}$ # now changing the indice to make the terms of the Poisson law appear
$E(X) = \theta * 1$ # their sum is by definition equal to one
$\iff E(X) = \theta$

ii. Computing the variance of the Poisson distribution:
We have : $Var(X) = E(X^2) - E(X)^2$.
Let’s compute $E(X^2)$:
$E(X^2) = \sum_{k=0}^{\infty} k^2 * \frac{e^{-\theta}\theta^k}{k!}$ $E(X^2) = \sum_{k=0}^{\infty} k(k-1+1) * \frac{e^{-\theta}\theta^k}{k!}$ # again we already know that the sum converges so we slightly abuse notations.
$\iff E(X^2) = \sum_{k=0}^{\infty} k(k-1) * \frac{e^{-\theta}\theta^k}{k!} + \sum_{k=0}^{\infty} k * \frac{e^{-\theta}\theta^k}{k!}$
$\iff E(X^2) = \theta^2 * \sum_{k=2}^{\infty} \frac{e^{-\theta}\theta^{k-2}}{k-2!} + E(X)$ # in the same way as for the expectation, we simplify by $k*(k-1)$ and then change the indices to make the Poisson sum of probability terms appear. We then note that this sum is equal to one.
$\iff E(X^2) = \theta^2 * 1 + \theta$
Finally, we get : $Var(X) = E(X^2) - E(X)^2 = \theta^2 + \theta - \theta^2 = \theta$

Question 3.

i. Observations:
Assume there are n games. Then, for each game, our observation is the number of goals scored and we have n observations.

ii. Model:
It is the set of all possible Poisson laws whose parameter is in $R_+^\star$.
We have $M = \left\{p(.|\theta), \theta \in R_+^\star\right\}$ where $p(.|\theta) = exp(-\theta) * \frac{\theta^k}{k!}$

iii. Parameter:
We are trying to estimate θ, which is the mean but also the variance of our Poisson law.

Question 4.

The likelihood of a Poisson distribution for a given x in R^n is: $l(x_1, x_2 ... x_n) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta_i^x}{x_i!}$ Thus the log-likelihood is:
$L(x_1, x_2 ... x_n) = \sum_{i=1}^{n} \log(\frac{e^{-\theta}\theta_i^x}{x_i!})$
$L(x_1, x_2 ... x_n) = -n * \theta + \log\theta * \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \sum_{k=1}^{x_i} k$

The derivative is therefore:
$d\theta L(x_1, x_2 ... x_n) = -n + \frac{1}{\theta} * \sum_{i=1}^{n} x_i$
So, the likelihood is maximised for $\theta = \frac{1}{n} * \sum_{i=1}^{n} x_i$, when the derivative of the log-likelihood is null. Thus the mean and the MLE are the same for the Poisson distribution.

Question 5.

We have that the MLE is in fact the mean of the random variables. As our Poisson laws are all independent and identically distributed of mean θ, by the central limit theorem we get that $\sqrt{n} * (\hat{\theta}_{MLE} - \theta)$ converges in distribution to $N(0, \theta)$ given that the variance is θ.

Question 6.

We know that $\hat{\theta}_{MLE}$ is an estimator of θ (the mean as well as the variance of our Poisson law).
$\hat{\theta}_{MLE}$ converges in probability to θ.
So, given that $g : x \longrightarrow \frac{1}{\sqrt{x}}$ is continuous on $R_{+}^{\star}$, by the continuous mapping theorem we get that $\frac{1}{\sqrt{\hat{\theta}_{MLE}}}$ converges in probability to $\frac{1}{\sqrt{\theta}}$.
We showed in question 5 that $\sqrt{n} * (\hat{\theta}_{MLE} - \theta)$ converged in distribution to $\mathcal{N}(0, \theta)$.
Thus, by Slutsky, $\sqrt{n} \frac{(\hat{\theta}_{MLE} - \theta)}{\sqrt{\hat{\theta}_{MLE}}}$ converges in law to $N(0, 1)$.

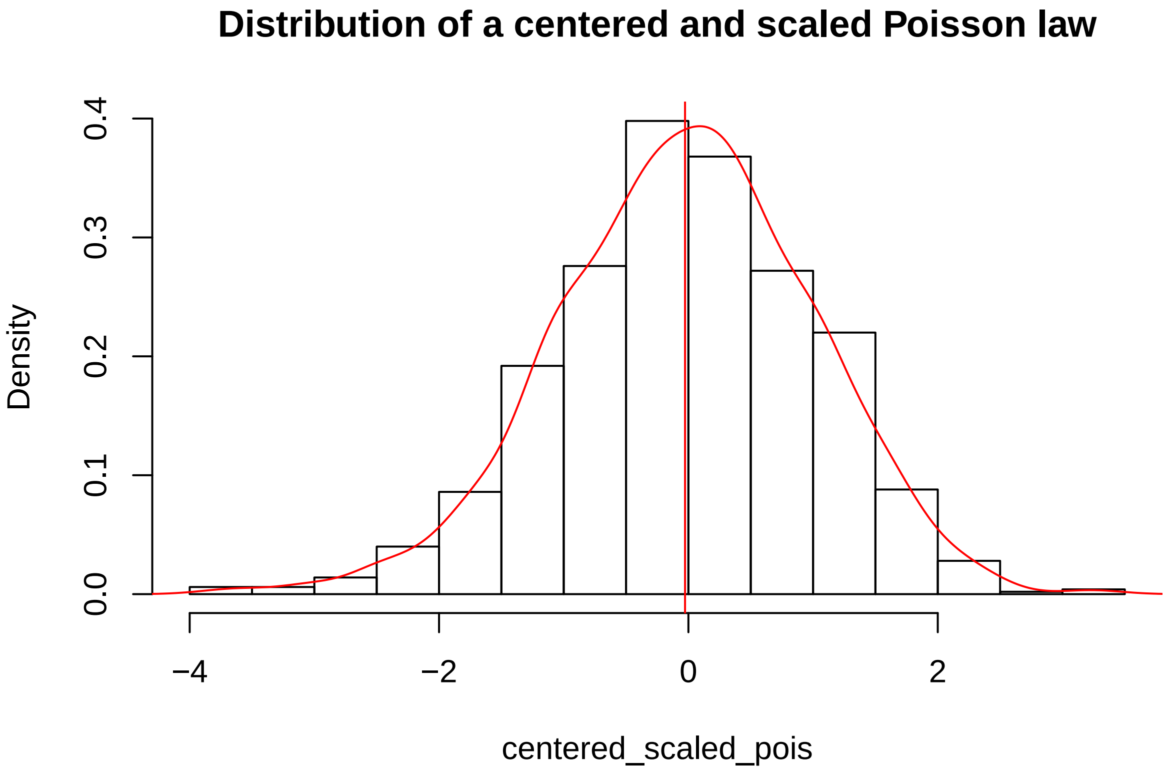
Empirical verification using R:

i. First defining the hypotheses and creating the realisations we want to plot.

```
N_attempts = 1000
n_sample = 100
theta= 3
centered_scaled_pois = c()
for (i in 1:N_attempts) {
x_pois = rpois(n_sample,theta)
theta_mle = mean(x_pois)
centered_scaled_pois[i] =sqrt(n_sample)*(theta_mle - theta)/sqrt(theta_mle)}
```

ii. Now plotting the histogram.

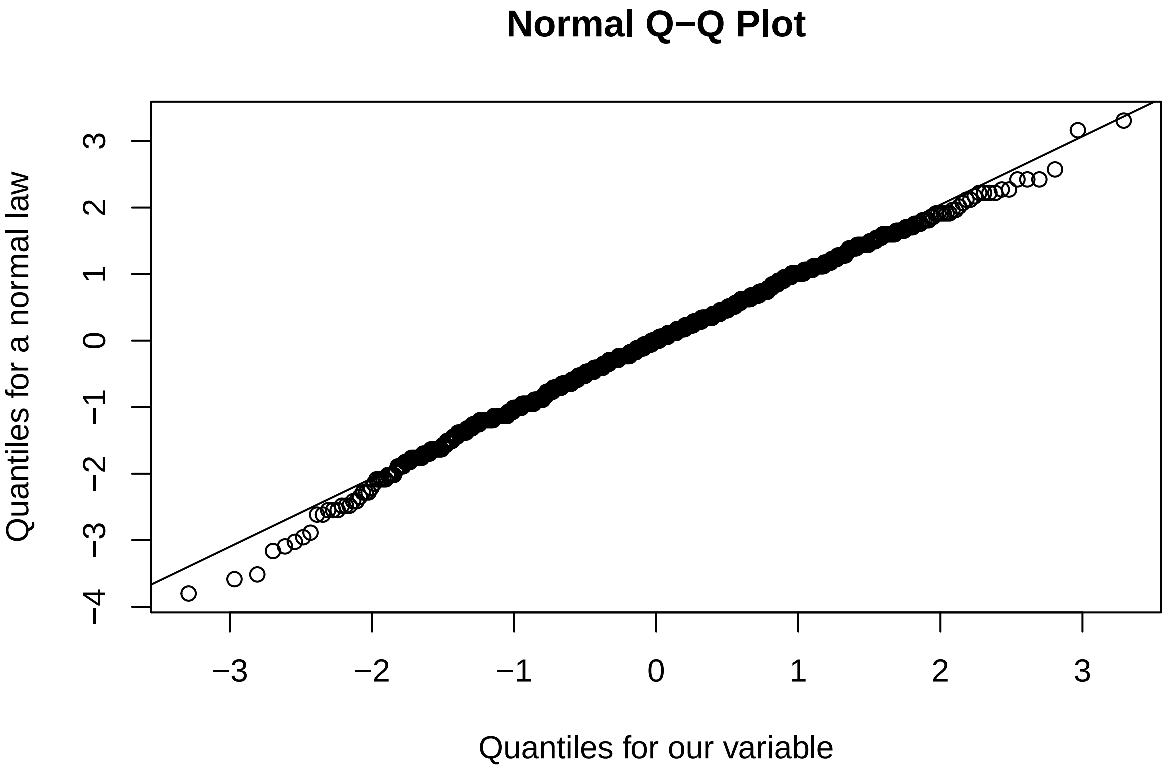
```
hist(centered_scaled_pois, main ="Distribution of a centered and scaled Poisson law", prob = TRUE)
d = density(centered_scaled_pois)
lines(d, col='red')
abline(v=mean(centered_scaled_pois), col='red')
```



We notice that all values are centered around zero and symmetrically distributed around the mean. The tails are very narrow. This resembles normal distribution of mean 0 and standard deviation one.

iii. Now plotting the qqplot.

```r
qqnorm(centered_scaled_pois, xlab = "Quantiles for our variable", ylab = "Quantiles for a normal law")
qqline(centered_scaled_pois)
```



This qqplot confirms that our centered scaled Poisson follows a normal law.

Question 7.

We know that $\sqrt{n}\frac{(\hat{\theta}_{MLE}-\theta)}{\sqrt{\hat{\theta}_{MLE}}}$ converges in law to $N(0,1)$. Let $Z \sim \mathcal{N}(0,1)$ and z_α its α quartile. We get the following interval:
$\lim_{n\to\infty} P(-z_{1-\alpha/2} \leq \sqrt{n}\frac{(\hat{\theta}_{MLE}-\theta)}{\sqrt{\hat{\theta}_{MLE}}} \leq z_{1-\alpha/2}) = 1-\alpha$
We have:
$-z_{1-\alpha/2} \leq \sqrt{n}\frac{(\hat{\theta}_{MLE}-\theta)}{\sqrt{\hat{\theta}_{MLE}}} \leq z_{1-\alpha/2} \iff -\frac{\sqrt{\hat{\theta}_{MLE}}}{\sqrt{n}} z_{1-\alpha/2} \leq (\hat{\theta}_{MLE}-\theta) \leq \frac{\sqrt{\hat{\theta}_{MLE}}}{\sqrt{n}} z_{1-\alpha/2}$
$\iff \hat{\theta}_{MLE} - \frac{\sqrt{\hat{\theta}_{MLE}}}{\sqrt{n}} z_{1-\alpha/2} \leq \theta \leq \hat{\theta}_{MLE} + \frac{\sqrt{\hat{\theta}_{MLE}}}{\sqrt{n}} z_{1-\alpha/2}$
$\forall \epsilon > 0$, if we define: i. $\alpha_n = \hat{\theta}_{MLE} - \frac{\sqrt{\hat{\theta}_{MLE}}}{\sqrt{n}} z_{1-\alpha/2} - \epsilon$
ii. $\beta_n = \alpha_n = \hat{\theta}_{MLE} + \frac{\sqrt{\hat{\theta}_{MLE}}}{\sqrt{n}} z_{1-\alpha/2} + \epsilon$ Then, we get $\lim_{n\to\infty} P(\theta \in [\alpha_n, \beta_n]) \geq 1-\alpha$.

Question 8.

We know from q5) that $\sqrt{n} * (\hat{\theta}_{MLE} - \theta) \sim \mathcal{N}(0,\theta)$.
We define $g : x \longrightarrow 2\sqrt{x}$ a continuously derivable function on R_+^*.
$\forall x \in R_+^*$, we have $g'(x) = \frac{1}{\sqrt{x}}$.
So, the delta-method gives us that: $\sqrt{n}(2\sqrt{\hat{\theta}_{MLE}} - 2\sqrt{\theta}) \sim \mathcal{N}(0, \theta * g'(\theta)^2)$ $\sqrt{n}(2\sqrt{\hat{\theta}_{MLE}} - 2\sqrt{\theta}) \sim \mathcal{N}(0, \theta * \frac{1}{\theta})$
$\sqrt{n}(2\sqrt{\hat{\theta}_{MLE}} - 2\sqrt{\theta}) \sim \mathcal{N}(0,1)$

Question 9.

We know that $\sqrt{n}(2\sqrt{\hat{\theta}_{MLE}} - 2\sqrt{\theta}) \sim \mathcal{N}(0,1)$.
So, we get $lim_{n\to\infty} P(-z_{1-\alpha/2} \leq \sqrt{n}(2\sqrt{\hat{\theta}_{MLE}} - 2\sqrt{\theta}) \leq z_{1-\alpha/2}) = 1-\alpha$

We have: $-z_{1-\alpha} \leq \sqrt{n}(2\sqrt{\hat{\theta}_{MLE}} - 2\sqrt{\theta}) \leq z_{1-\alpha}$
$\iff -\frac{1}{2\sqrt{n}}z_{1-\alpha} \leq (\sqrt{\hat{\theta}_{MLE}} - \sqrt{\theta}) \leq \frac{1}{2\sqrt{n}}z_{1-\alpha}$
$\iff \sqrt{\hat{\theta}_{MLE}} - \frac{1}{2\sqrt{n}}z_{1-\alpha} \leq \sqrt{\theta} \leq \sqrt{\hat{\theta}_{MLE}} + \frac{1}{2\sqrt{n}}z_{1-\alpha}$ (1) When n goes to infinity, this inequality is greater than zero almost surely given that the left hand side of the inequality converges in probability to $\sqrt{\theta}$ almost surely.
Thus, given that $g : x \longrightarrow x^2$ is continuous and strictly increasing on $R_+^\star$, we get:
(1) $\iff \theta \in [(\sqrt{\hat{\theta}_{MLE}} - \frac{1}{2\sqrt{n}}z_{1-\alpha})^2, (\sqrt{\hat{\theta}_{MLE}} + \frac{1}{2\sqrt{n}}z_{1-\alpha})^2]$. Thus, $\forall \epsilon > 0$, if we define:
$c_n = (\sqrt{\hat{\theta}_{MLE}} - \frac{1}{2\sqrt{n}}z_{1-\alpha})^2 - \epsilon$
$d_n = (\sqrt{\hat{\theta}_{MLE}} + \frac{1}{2\sqrt{n}}z_{1-\alpha})^2 + \epsilon$
Then, we have $lim_{n \to \infty} P(\theta \in [c_n, d_n] \geq (1-\alpha)$

Question 10.

We can denote $\hat{\theta}_1 = \frac{1}{n}\sum_{k=1}^{n} X_i$ the sample mean, an unbiased estimator of θ.
And $\hat{\theta}_2 = \frac{1}{n}\sum_{k=1}^{n}(X_i - \overline{X})^2$ the sample variance, a biased estimator of θ given that the mean and the variance of a Poisson law are the same. We note that $\hat{\theta}_1$ is the MLE estimator of θ.
N.B. It is also the one with minimal variance among all unbiased estimators.
Indeed, we have: $Var(\hat{\theta}_1) = \frac{1}{n^2} * n * \theta$ # it is the variance of the MLE estimator $\iff Var(\hat{\theta}_1) = \frac{\theta}{n}$

Question 11.

i. Computation of the bias:
$E(\hat{\theta}_{MLE}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i)$
$E(\hat{\theta}_{MLE}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$ by linearity of the expectation
$E(\hat{\theta}_{MLE}) = \frac{1}{n} * n\theta \; E(\hat{\theta}_{MLE}) = \theta$
So, $\hat{\theta}_{MLE}$ is an unbiased estimator of θ and $b^\star(\hat{\theta}_{MLE}) = 0$.

ii) Computation of the variance:
$Var^*(\hat{\theta}_{MLE}) = Var^*(\frac{1}{n}\sum_{i=1}^{n} X_i)$
$Var^*(\hat{\theta}_{MLE}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$ because the X_is are i.i.d.
$Var^*(\hat{\theta}_{MLE}) = \frac{\theta}{n}$

iii) Computation of the quadratic risk of $\hat{\theta}_{MLE}$.
$risk = b^*(\hat{\theta})^2 + Var^*(\hat{\theta})$
$risk = 0 + \frac{\theta}{n}$

Question 12.

Cramer-Rao bound for $\hat{\theta}_{MLE}$:
i. Let's compute $L''(x_1, x_2, ...x_n)$.
$L'(x_1, x_2, ...x_n) = -n + \frac{1}{\theta} * \sum_{i=1}^{n} x_i$
$\iff L''(x_1, x_2, ...x_n) = -\frac{1}{\theta^2} * \sum_{i=1}^{n} x_i$
ii. So, we get:
$I(\theta) = E(-L''(X_1, X_2, ..., X_n)$
$\iff E(-L''(X_1, X_2, ..., X_n)) = E(\frac{1}{\theta^2} * \sum_{i=1}^{n} X_i))$
$\iff E(-L''(X_1, X_2, ..., X_n)) = \frac{n*\theta}{\theta^2}$
$\iff E(-L''(X_1, X_2, ..., X_n)) = \frac{n}{\theta}$
So, the variance of any unbiased estimator of θ will always be superior to $\frac{\theta}{n}$ using Fisher's theorem.
So, by Fisher, $\hat{\theta}_1$ is the estimator of θ with minimal variance over all unbiased estimators as its variance is equal to the Cramer Rao bound.

Question 13.

We have:
$\hat{\theta}_2 = \frac{1}{n} * \sum_{k=1}^{n}(X_i - \overline{X_n})^2$ $\hat{\theta}_2 = \frac{1}{n} * \sum_{k=1}^{n}(X_i - \theta + \theta - \overline{X_n})^2$
$\hat{\theta}_2 = \frac{1}{n} * \sum_{k=1}^{n}((X_i - \theta)^2 + (\theta - \overline{X_n})^2 + 2 * (X_i - \theta) * (\theta - \overline{X_n}))$
$\hat{\theta}_2 = \frac{1}{n} * \sum_{k=1}^{n}(X_i - \theta)^2 + (\theta - \overline{X_n})^2 - 2 * (\theta - \overline{X_n}) * \frac{1}{n} * \sum_{k=1}^{n}(\theta - X_i)$
$\hat{\theta}_2 = \frac{1}{n} * \sum_{k=1}^{n}(X_i - \theta)^2 + (\theta - \overline{X_n})^2 - 2 * (\theta - \overline{X_n})^2$
$\hat{\theta}_2 = \frac{1}{n} * \sum_{k=1}^{n}(X_i - \theta)^2 - (\theta - \overline{X_n})^2$

Question 14.

i. $E((\theta - \overline{X_n})^2) = E(\theta^2 + \overline{X_n}^2 - 2 * \theta * \overline{X_n})$
$E((\theta - \overline{X_n})^2) = \theta^2 + Var(\overline{X_n}) + E(\overline{X_n})^2 - 2 * \theta * E(\overline{X_n})$
$E((\theta - \overline{X_n})^2) = \theta^2 + \frac{\theta}{n} + \theta^2 - 2 * \theta^2$ $E((\theta - \overline{X_n})^2) = \frac{\theta}{n}$

ii. We also have:
$E(\frac{1}{n}\sum_{i=1}^{n}(X_i - \theta^2)) = \frac{1}{n}\sum_{i=1}^{n}(E(X_i^2) - 2\theta * E(X_i) + \theta^2)$ $E(\frac{1}{n}\sum_{i=1}^{n}(X_i - \theta^2)) = \frac{1}{n}\sum_{i=1}^{n}(\theta + \theta^2 - 2\theta^2 + \theta^2)$
# by using $E(X^2) = Var(X) + E(X)^2$
$E(\frac{1}{n}\sum_{i=1}^{n}(X_i - \theta^2)) = \theta$
So, we have: $E(\hat{\theta_2}) = \frac{1}{n} * \sum_{k=1}^{n}(X_i - \theta)^2 - (\theta - \overline{X_n})^2$ $E(\hat{\theta_2}) = \frac{n-1}{n}\theta$ # by summing the two equations we obtained earlier
So, to get an unbiased estimator, we can define $\hat{\theta}_3 = \frac{n}{n-1}\hat{\theta_2} = \frac{1}{n-1} * \sum_{k=1}^{n}(X_i - \overline{X_n})^2$

Question 15.

We know thanks to the decomposition in q)13 that:
$\hat{\theta}_2 = \frac{1}{n} * \sum_{k=1}^{n}(X_i - \theta)^2 - (\theta - \overline{X_n})^2$
i. We first note that $\overline{X_n} \to \theta$ in probability.
So, as $g : x \longrightarrow (x - \theta)$ is continuous on R, we get that $(\overline{X_n} - \theta) \to 0$ in probability.
ii. We know that the X_i are i.i.d. So, the $(X_i - \theta)^2$ are also i.i.d.
$E((X_i - \theta)^2) = E(X_i^2) - 2\theta E(X_i) + \theta^2 = Var(X_i) + E(X_i)^2 - 2\theta^2 + \theta^2 = \theta.$
So, by using the central limit theorem, we get that $\sqrt{n}(\frac{1}{n} * \sum_{k=1}^{n}(X_i - \theta)^2 - \theta)$ converges in distribution to $\sim \mathcal{N}(0, 2\theta^2 + \theta)$.
iii. As the X_i are i.i.d and $\overline{X}$ converges in probability to θ, we apply the CLT again to get $\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} X_i - \theta) \sim \mathcal{N}(0, \theta)$ when n goes to infinity. So, as $\overline{X} - \theta$ converges in probability to 0, we get that $\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} X_i - \theta) * (\overline{X} - \theta)$ converges in law to zero.
However we know from the course that convergence in law to a constant implies convergence in probability to that constant. iv. Finally, by Slutsky, we obtain that: $A = \sqrt{n}(\frac{1}{n}\sum_{k=1}^{n}(X_i - \hat{\theta})^2 - \theta) - \sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} X_i - \theta)(\overline{X} - \theta) \sim \mathcal{N}(0, 2\theta^2 + \theta)$ when n goes to infinity. We now just have to notice that $A = \sqrt{n}(\hat{\theta}_2 - \theta)$. So, we get that $\sqrt{n}(\hat{\theta}_2 - \theta) \sim \mathcal{N}(0, 2\theta^2 + \theta)$.

Question 16.

i. Computing the moment generating function. We have:
$E(e^{sX}) = \exp(-\theta)\sum_{k=0}^{\infty} e^{sk}\frac{\theta^k}{k!}$ $E(e^{sX}) = \exp(-\theta) * [exp(\theta e^s)\exp(-\theta e^s)] * \sum_{k=0}^{\infty}\frac{(\theta e^s)^k}{k!}$
$E(e^{sX}) = \exp(-\theta) * exp(\theta e^s)[\exp(-\theta e^s) * \sum_{k=0}^{\infty}\frac{(\theta e^s)^k}{k!}]$ # we are making the sum on N of the probability terms of a Poisson law of parameter θe^s appear.
Finally, we obtain $E(e^{sX}) = \exp(\theta(e^s - 1))$.

ii. Computing the expectation:
$G_X(s) = E(e^{sX}) = \sum_{k=0}^{\infty}(e^s)^k P(X = k)$
$G'_X(s) = \sum_{k=0}^{\infty} k(e^s)^k P(X = k)$.
So, $G'_X(0) = E(X)$.

But $\forall s > 0, G'_X(s) = (\theta e^s)\exp(\theta(e^s - 1))$ # by using the expression found in i.
Thus, $E(X) = \theta$.

iii. Computing the variance:
$G'_X(s) = \sum_{k=0}^{\infty} k(e^s)^k P(X = k)$
$G''_X(s) = \sum_{k=0}^{\infty} k^2(e^s)^k P(X = k)\$$
So, $G''_X(0) = E(X^2)$
$G''_X(0) = Var(X) + E(X)^2$
$G''_X(0) = Var(X) + \theta^2$
But $\forall s > 0, G''_X(s) = G'_X(s) + (\theta e^s)^2 \exp(\theta(e^s - 1))$ # by using the expressions found in i. and ii. for G_X and its first derivative Thus, $G''_X(0) = \theta + \theta^2$ Thus, $Var(X) + E(X)^2 = \theta + \theta^2$ and $E(X)^2 = \theta^2$
We finally obtain $Var(X) = \theta$.

iv. Recovering that a sum of independent Poisson laws of respective parameters λ_1 and λ_2 is a Poisson law of parameter $\lambda_1 + \lambda 2$:
$G_{X_1+X_2}(s) = E(e^{s(X_1+X_2)}) = E(e^{sX_1}e^{X_2}) = G_{X_1}(s) * G_{X_2}(s)$ by independence of X_1 and X_2
$\iff G_{X_1+X_2}(s) = \exp(\lambda_1(e^s - 1)) * \exp(\lambda_2(e^s - 1)) = \exp((\lambda_1 + \lambda_2)(e^s - 1))$.
So, the result we set out to prove is true given the unicity of the moment generating function.

v. We have:

a) $G_{X-\theta}(s) = E(\exp(sX - s\theta)) = \exp(-s\theta) * E(\exp(sX)) = \exp(-s\theta) * G_X(s)$

b) $G_{X-\theta}(s) = \exp(-\theta)\sum_{k=0}^{\infty} \exp((k-\theta)s)) * \frac{\theta^k}{k!}$
$G'_{X-\theta}(s) = \exp(-\theta)\sum_{k=0}^{\infty}(k-\theta)\exp((k-\theta)s)) * \frac{\theta^k}{k!}$
$G'_{X-\theta}(0) = \exp(-\theta)\sum_{k=0}^{\infty}(k-\theta) * \frac{\theta^k}{k!} = E(X-\theta)$
$G''_{X-\theta}(s) = \exp(-\theta)\sum_{k=0}^{\infty}(k-\theta)^2\exp((k-\theta)s)) * \frac{\theta^k}{k!} = E((X-\theta)^2)$ # by an evident recurrence due to the derivative of the exponential, we get that $\forall n \in N^\star, G^{(n)}_{X-\theta}(0) = E((X-\theta)^n)$

c) $G_{X-\theta}(s) = \exp(\theta(e^s - 1 - s))$ so $G_{X-\theta}(0) = 1$
$\Rightarrow G'_{X-\theta}(s) = \theta(e^s - 1) * G_{X-\theta}(s)$ so $G'_{X-\theta}(0) = 0$
$\Rightarrow G''_{X-\theta}(s) = \theta e^s(G_{X-\theta}(s) + G'_{X-\theta}) - \theta G'_{X-\theta}(s)$ so $G''_{X-\theta}(0) = \theta$
$\Rightarrow G^{(3)}_{X-\theta}(s) = \theta e^s(G_{X-\theta}(s) + 2 * G'_{X-\theta}(s) + G''_{X-\theta}(s)) - \theta G''_{X-\theta}(s)$ so $G^{(3)}_{X-\theta}(0) = \theta(1 + 0 + \theta) - \theta^2 = \theta$
$\Rightarrow G^{(4)}_{X-\theta}(s) = (\theta e^s)(G_{X-\theta}(s) + 3G'_{X-\theta}(s) + 3G''_{X-\theta}(s) + G^{(3)}_{X-\theta}(s)) - \theta G^{(3)}_{X-\theta}(s)$ so $G^{(4)}_{X-\theta}(0) = \theta(1 + 0 + 3\theta + \theta) - \theta^2 = 3\theta^2 + \theta$.
By using the result we found in b, we finally obtain that $E((X_i - \theta)^4) = 3\theta^2 + \theta$.
So using $E((X_i - \theta)^2) = \theta$ as demonstrated in q.15) ii., we finally obtain:
$Var((X_i - \theta)^2)) = E((X_i - \theta)^4) - E((X_i - \theta)^2)^2) = 3\theta^2 + \theta - \theta^2 = 2\theta^2 + \theta$.

Problem 2 - Analysis of the USJudgeRatings dataset.

I. First contact and preliminary observations.

1) Introduction to the dataset.

```
# First loading the dataset
data(USJudgeRatings)
```

The dataset name suggests that we are looking into a panel of US judges's ratings on a set of criteria. We do not know the outcome of the experiment.

```
# Printing the first ten rows of the dataset
USJudgeRatings[1:10,]
```

```
##                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
```

Each observation corresponds to one judge of the US Superior Court. The variables, all numeric, are the columns. We suppose them to be the judge's ratings on a set of criteria. We do not know whether one of the columns is in fact an output.

```
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 ...
```

We have 12 variables in this dataframe, all of which are numeric. A priori, they seem pretty close in terms of extreme values and dispersion, an insight to be confirmed through boxplots.

We are now diving deeper in the context surrounding the experiment and our variables.

```
# We check that our dataframe isn't missing any values
print(paste("There are", sum(is.na(USJudgeRatings)), "missing values in the dataframe"))
```

```
## [1] "There are 0 missing values in the dataframe"
```

2) Contextual information on the data and the experiment methodology.

```
?USJudgeRatings
```

1. For a given observation, each of our numeric variables measures the performance of the judge along a given criterion as rated by lawyers. For instance, "DECI" measures the ability of a judge to make quick decisions.

2. We note that the data is old: it dates back to 1977, which means caution when handling it to draw conclusions today.

3. There are little indications on the period of time over which these ratings were given. It would be interesting to gather more data on the frequency at which they were reported, who required them and with what in mind.
4. We have no information on the grading methodology. Were the lawyers briefed in advance? Did they have a list of elements to observe, or did they go with their gut feeling?

5. We also have little information on the examinators themselves, the lawyers.
   For a given criterion, numbers are relatively round. We can assume that one lawyer was reponsible for grading a given judge on all criteria, and not a set of lawyers of which we averaged the results: this leaves higher possibilities of bias. Was it always the same lawyer, or did it change? How objective were they relatively to the trial, or trials, they observed?
6. We have no indication on the range of values the ratings could take. The help function gives us more information on the variables themselves:
   [,1] CONT Number of contacts of lawyer with judge.
   # This could be an indicator of the integrity of the lawyer when rating the judge. This variable should not be correlated to others if the lawyers are objective in rating the judge's performance.
   [,2] INTG Judicial integrity.
   # This criterion is difficult to evaluate relying on objective facts. It relies mostly on the lawyer's impression.
   [,3] DMNR Demeanor.
   # Idem. Judicial integrity can be a part of demeanor, so these two variages might be linked.
   [,4] DILG Diligence.
   # Idem
   [,5] CFMG Case flow managing.
   # If we know of a set of best practices, then this criterion might be more reliable [,6] DECI Prompt decisions.
   # We expect DECI and diligence to be correlated. Moreover, is it more important to make prompt decisions or to give sound rulings? Further analysis of the dataset will give us more information on the
   [,7] PREP Preparation for trial.
   # Nothing to report [,8] FAMI Familiarity with law.
   # This should be one of the most objective rating criteria given that we are referring to a pre-defined set of norms. We expect it to be correlated to preparation for trial.
   [,9] ORAL Sound oral rulings.
   # This variable is already a judgement on the performance of the judge depending on other factors (preparation, for instance) [,10] WRIT Sound written rulings.
   # Same remark as previously. We would expect ORAL and WRIT to be strongly correlated given that it is only the form of the judgement that changes [,11] PHYS Physical ability.
   # Nothing to report
   [,12] RTEN Worthy of retention.
   # This seems to be a final rating answering the following question: should the judge be retained at the US Superior Court?. However, we do not know how we decide who stays or not given this grade. Indeed: will we keep judges above a certain grade threshold? Will we keep a certain number of judges, and then decide on the grade threshold that will help us decide who to keep? This is an important question which should guide the next steps of the experiment.

The eleven first columns seem to be grades that lead to a final evaluation on whether or not the judge should be retained in the US Superior Court. As noted in the comments, some criteria rely on more objective observations than others.

```r
summary(USJudgeRatings)
```

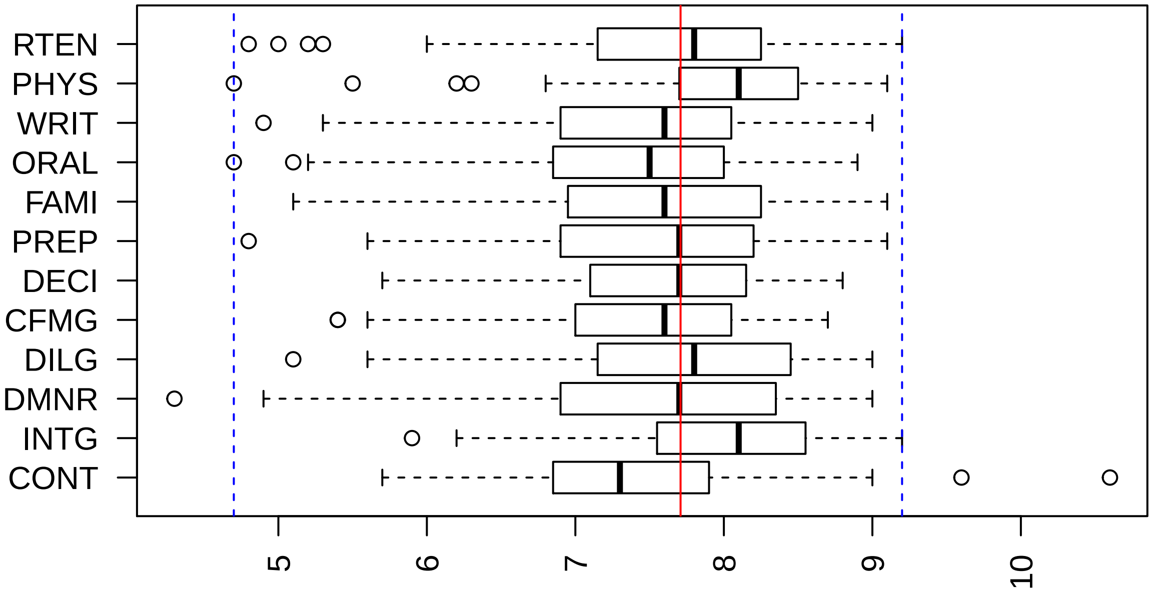
```
##       CONT             INTG            DMNR            DILG      
##  Min.   : 5.700   Min.   :5.900   Min.   :4.300   Min.   :5.100  
##  1st Qu.: 6.850   1st Qu.:7.550   1st Qu.:6.900   1st Qu.:7.150  
##  Median : 7.300   Median :8.100   Median :7.700   Median :7.800  
##  Mean   : 7.437   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  
```

The variables means range in $[7.3, 8.0]$ which means they are pretty close to one another.
Their dispersion seems to be pretty similar also: we have one variable with a minimal value under 4.7, which is DMNR, and one variable whose maximum above 9.2, which is CONT. So except DMNR and CONT, all variables are contained within the $[4.7, 9.2]$ interval.
We note that means are close to the medians, but lower. So most variables are slightly skewed towards the left with some grades significantly lower than others, as we will confirm by studying the boxplots and skewness.

II. Multivariate and univariate analysis - Descriptive statistics.

1) Position and dispersion comparison via boxplots.

```
# Plotting the boxplots horizontally because it makes it easier to compare their medians and dispersion
boxplot(USJudgeRatings, las = 2, horizontal = TRUE)
# Plotting the mean of the median values to make any distance to it more visible
a = mean(c(7.3,8.1,7.7,7.8,7.6,7.7,7.7,7.6,7.5,7.6,8.1,7.8))
abline(v=a, col = 'red')
# Plotting the boundaries we noticed when analysing the summary
abline( v = 4.7, col = 'blue', lty = 2)
abline(v = 9.2, col = 'blue', lty = 2)
```



1. We confirm that our variables are approximately distributed within the same ranges, as visible by plotting in blue the boundaries identified previously, and by observing the shapes of the boxes and their whiskers.
2. In terms of **position**, all medians are close to 7.8, except PHYS and INTG which are significantly

higher, both at (8.1).
We can assume that the physical ability and integrity of judges is generally good.
CONT is significantly lower at (7.3), which could mean that the lawyers didn't have too many contacts with the judges (or they reported it that way).

3. In terms of **dispersion**, the boxplots are slightly skewed towards the lower boundaries versus the median, confirming that some low ratings deflate the mean ratings that judges get.
   Physical ability is narrow enough compared to the other variables, confirming that this physical ability is generally good for everyone.
   DMNR has more spread and a smaller lower whisker than the others.

4. There are a few **outliers** we should look out for. We first observe that outliers are most of the time to the lower boundaries, which coincides with our earlier observation that the means were slightly lower than the medians.
   The three variables for which outliers are the most visible are CONT, PHYS and DMNR, because they are far from the boxplots and distant from the blue boundaries identified in the summary.

a) CONT has an outlier above 10, which seems weird for two reasons:

i. The other data points suggest that 10 should be the upper boundary for all the ratings.
ii. The median value for CONT is significantly lower than for others. However, we can consider that CONT is not a grade but simply the measure of the number of contacts the lawyer had with the judge. However, it then seems weird that the values in CONT aren't integers…

b) The outliers for PHYS are far from the whiskers, even though they are within our blue boundaries. min(PHYS) = 4.7 << 7.7 which is the first quartile. It makes sense for PHYS to have some exceptional outliers, even though they are far-left from the box whiskers.

c) DMNR has an outlier at 4.3, which doesn't seem to disturbing given that:

i. It seems like an acceptable grade

ii. As we will see later, INTG and DMNR are strongly correlated, and the judge whose INTGR grade is of 4.3 has a low DMNR grade of 6.4 (< first quartile) which makes sense.

```r
USJudgeRatings$INTG[USJudgeRatings$DMNR == min(USJudgeRatings$DMNR)]
```

```
## [1] 6.4
```

d) We will consider the other outliers are acceptable for several reasons:

i. They are not too far from the lower-box whisker (WRIT, ORAL in particular). PREP might be more intriguing with an outlier at 4.8, well under the first quartile (6.9).

ii. We expect RTEN to present some outliers if the variables it depends on present some too.

Skewness and kurtosis of the data

```r
#install.packages("e1071")
library(e1071)
```

Given our observations of the boxplot, we expect the skewness of the dataset to be negative for each variable:

```r
for (name in colnames(USJudgeRatings)){
s = skewness(USJudgeRatings[,name])
print(c(name,s))}
```

```
## [1] "CONT"             "1.04831082852977"
## [1] "INTG"               "-0.813541922448701"
## [1] "DMNR"               "-0.914698975831847"
## [1] "DILG"              "-0.75619701024528"
## [1] "CFMG"              "-0.76275292434365"
## [1] "DECI"               "-0.621558748765366"
## [1] "PREP"               "-0.657253624881199"
## [1] "FAMI"               "-0.537792326172604"
## [1] "ORAL"               "-0.753038885790875"
## [1] "WRIT"               "-0.672682520721674"
## [1] "PHYS"              "-1.50417637184898"
## [1] "RTEN"               "-0.937360930294454"
```

As expected, we notice that practically all variables are slightly skewed towards the left except PHYS which is highly skewed towards the left and CONT which is highly skewed towards the right (absolute value greater than one).

```r
for (name in colnames(USJudgeRatings)){
k = kurtosis(USJudgeRatings[,name])
print(c(name,k))}
```

```
## [1] "CONT"             "1.51221182877214"
## [1] "INTG"              "0.257028146294933"
## [1] "DMNR"              "0.274266854918179"
## [1] "DILG"             "0.14697661975641"
## [1] "CFMG"                "-0.0964006409476323"
## [1] "DECI"               "-0.531858166900887"
## [1] "PREP"                "-0.0036222033312967"
## [1] "FAMI"              "-0.36534841231648"
## [1] "ORAL"               "0.0126668778846812"
## [1] "WRIT"               "-0.108868716116502"
## [1] "PHYS"             "2.15947201560131"
## [1] "RTEN"              "0.255742065912552"
```

Kurtosis analysis confirms that outliers PHYS (kurt = 2.2 >>1) and CONT (kurt = 1.5 > 1) outliers are significantly far from the mean, as observed on the boxplot.

```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

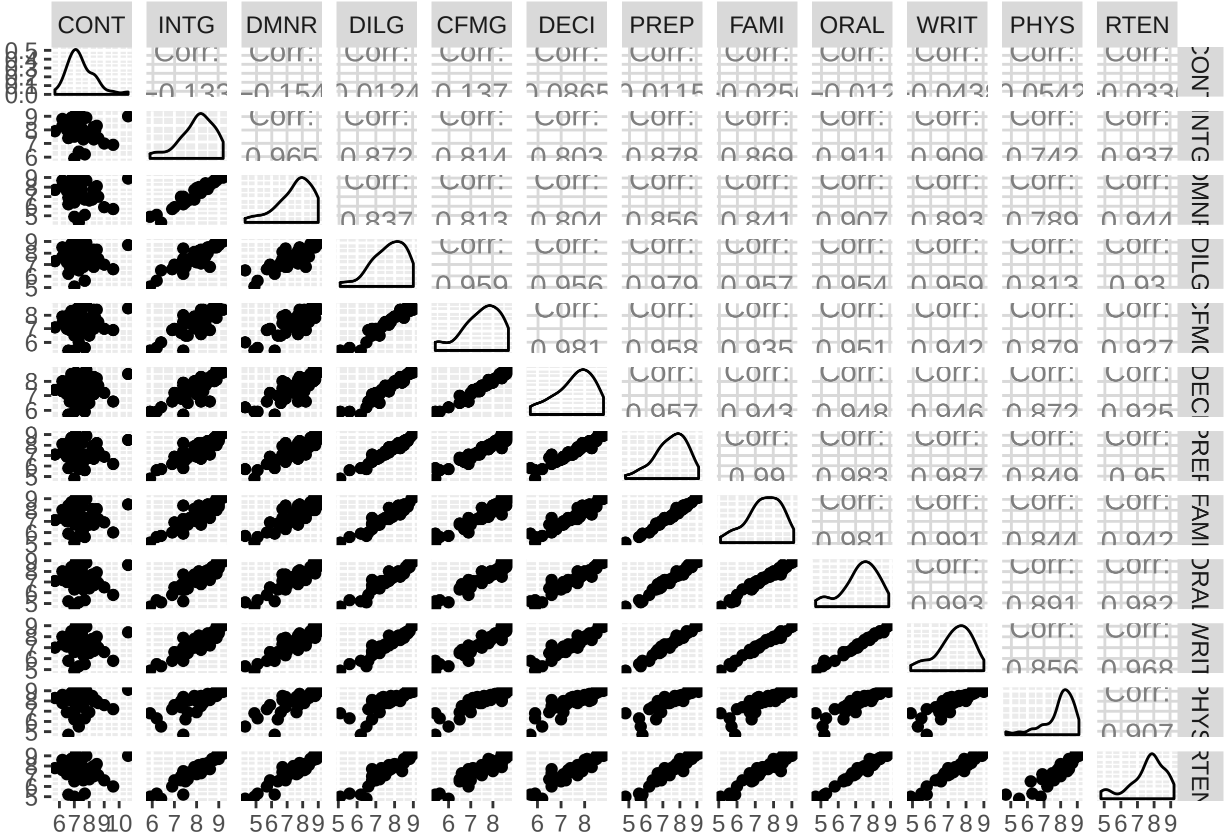
```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

2) General multivariate analysis: scatterplots and correlation matrix.

a) Scatterplots.

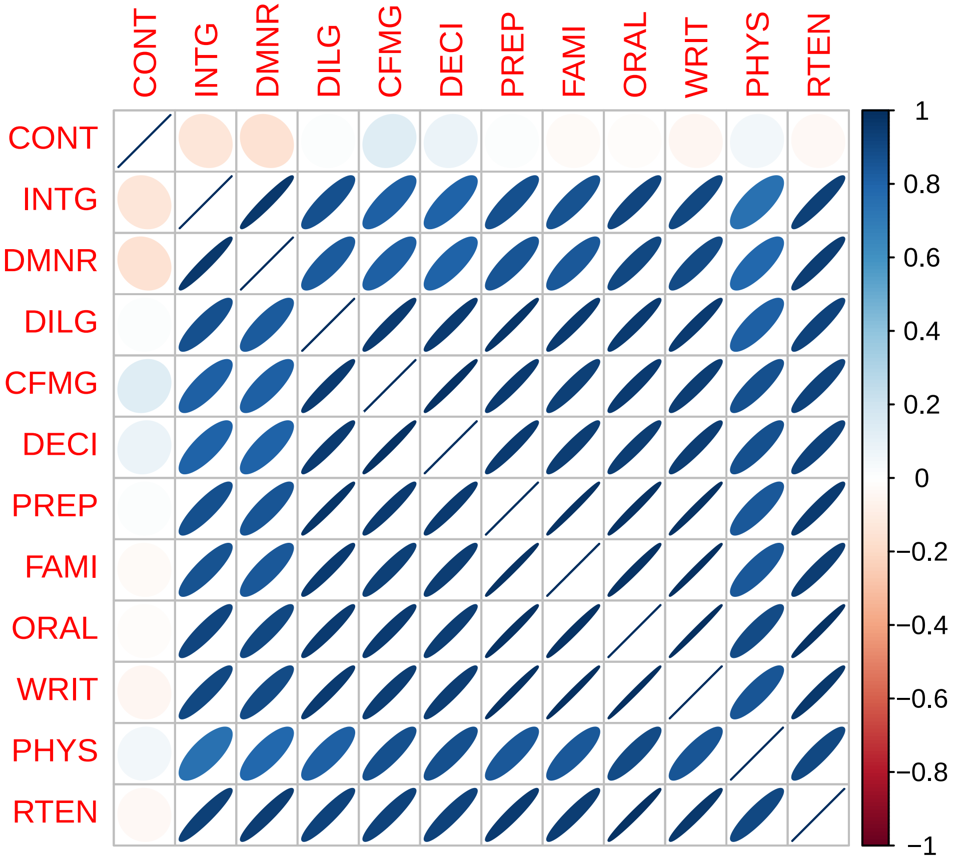
```
# We are first plotting the scatterplots of each variable relatively to the others to see if any clear
ggpairs(USJudgeRatings)
```



1. We observe that our variables are in fact strongly correlated, given that some scatterplots display near linear relations. This entails that if we want to predict RTEN using other variables, we will have to discriminate between the most useful features in explaining RTEN (maybe using fast forward selection). The relation between PREP and FAM in particular is quasi linear. This shouldn't surprise us, because it seems logical that the more one prepares a case, the better one knows the law surrounding it.
2. We remark that CONT is the only variable that has very little correlation with the others, with a highest correlation of ~0.2 in absolute value with DMNR.
3. It seems that the best linear relation between RTEN, our output, and another variable is for ORAL, meaning that ORAL would be the best single feature explaining the outcome.

b) Correlation matrix.

```
Corr = cor(USJudgeRatings)
corrplot(Corr, method = "ellipse")
```

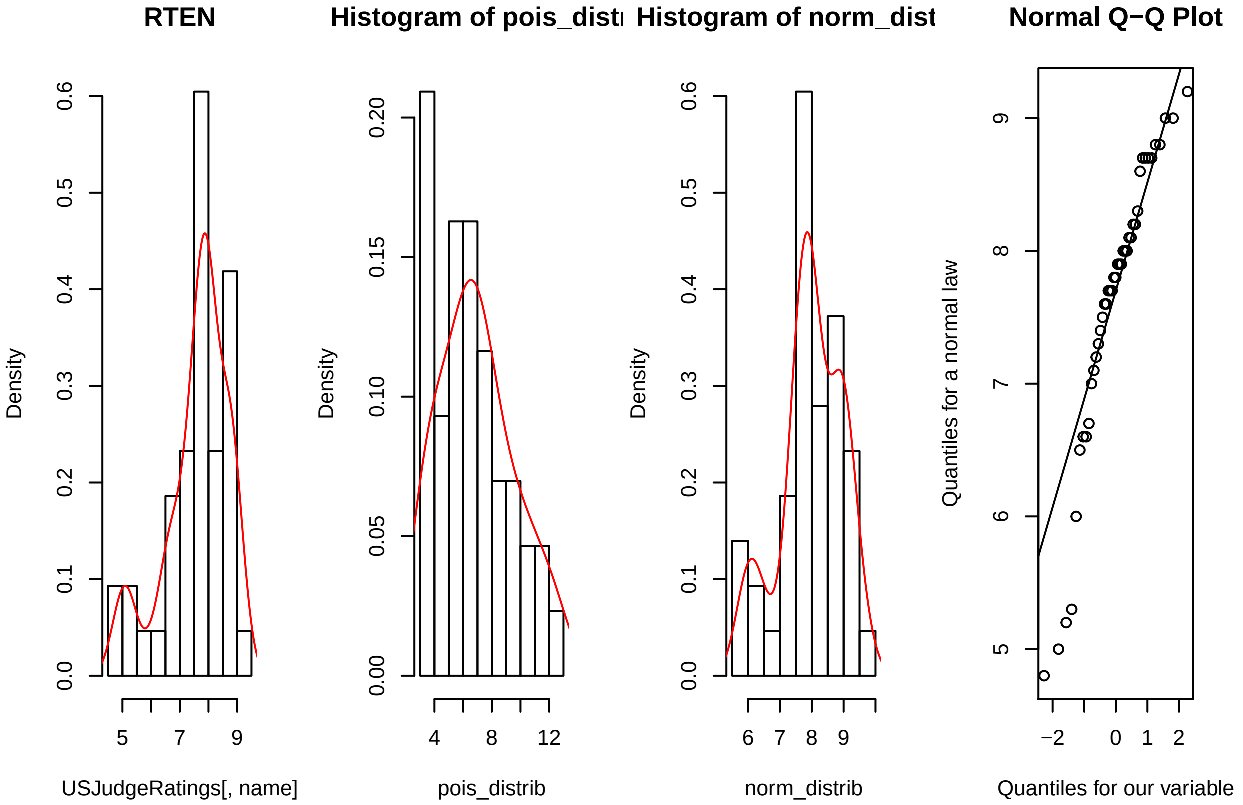


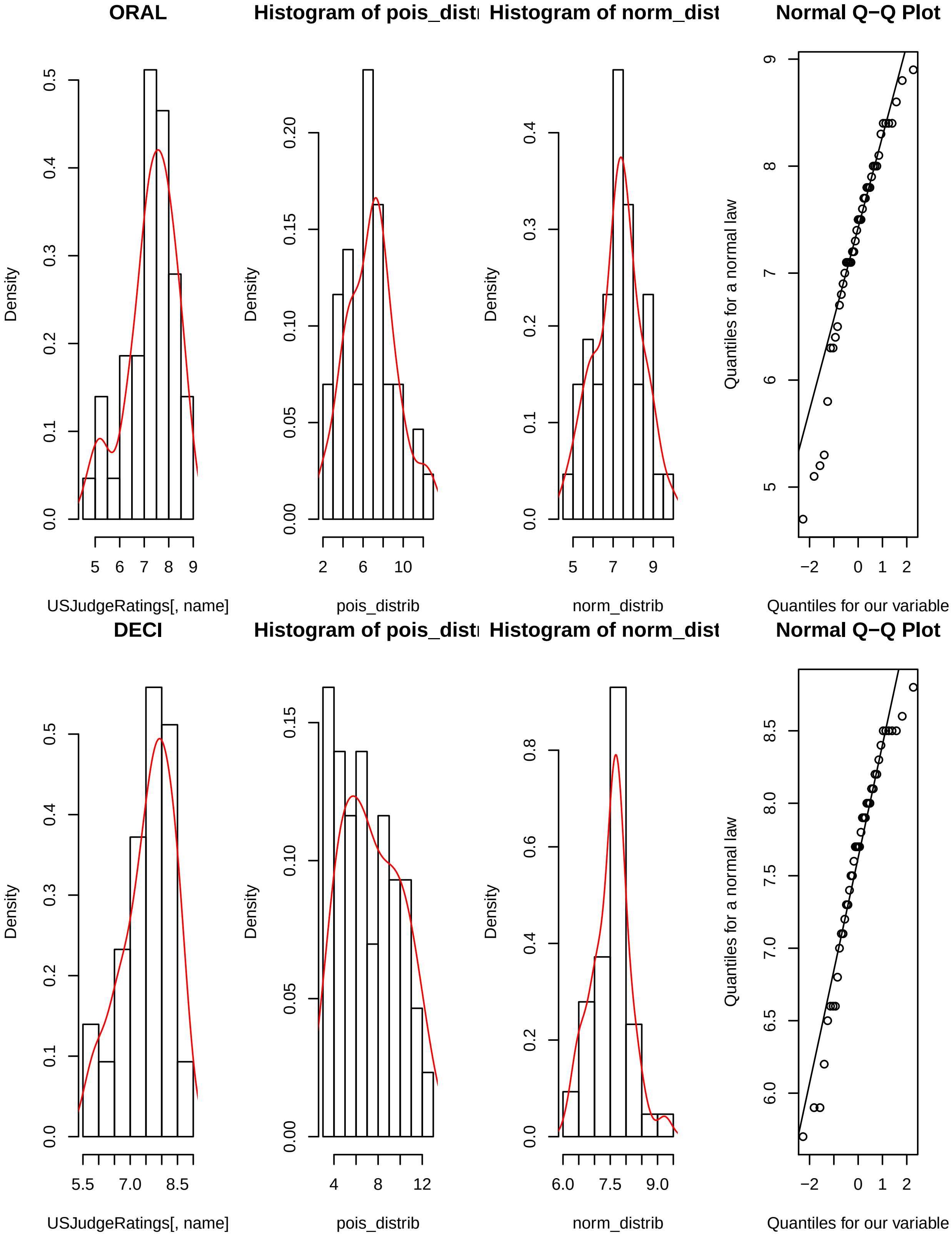
We confirm that our variables are very strongly correlated to one another, with most correlations between 0.8 and 1. The only variable which is rather uncorrelated to others is indeed CONT.
We have the following relations (analysing RTEN in a separate point as it is the output rating):
1. Variables that are highly correlated to multiple other variables include DILG, CFMG, DECI, PREP, FAMI, ORAL, and WRIT.
a. WRIT and ORAL are very strongly correlated to one another (corr = 0.99), as well as to FAMI and PREP (corr > 0.98). b. CFMG and DECI are strongly correlated (corr= 0.98) : promptitude plays a role in case flow management. To a lesser extent, they are also very strongly correlated to DILG, PREP and ORAL (corr > 0.95)
2. Variables that are highly correlated to less than two variables include INTG, DMNR, and PHYS.
a. INTG and DMNR are strongly correlated (corr = 0.96), but are less correlated to the other variables. They are both very slightly negatively correlated to CONT.
b. Among strongly correlated variables, PHYS is the one that least depends on others. However, it is strongly correlated to CFMG, DECI and ORAL (correlation >0.87). It would be an interesting variable to include in a model if it explained RTEN well: however, it does have one of the lowest correlation coeffs to the variable (0.91).
3. The only uncorrelated variable is CONT. CONT isn't related to other variables, and in particular it doesn't correlate to RTEN which is our output, so it probably isn't good towards explaining the data. However, it does underline the correlations between other codependent variables.
4. RTEN is highly correlated to all variables except CONT, with corr > 0.93 for each one except CONT (-0.03) and PHYS (0.91).
As noticed earlier, the highest correlation between RTEN and another variable occurs for ORAL (corr =

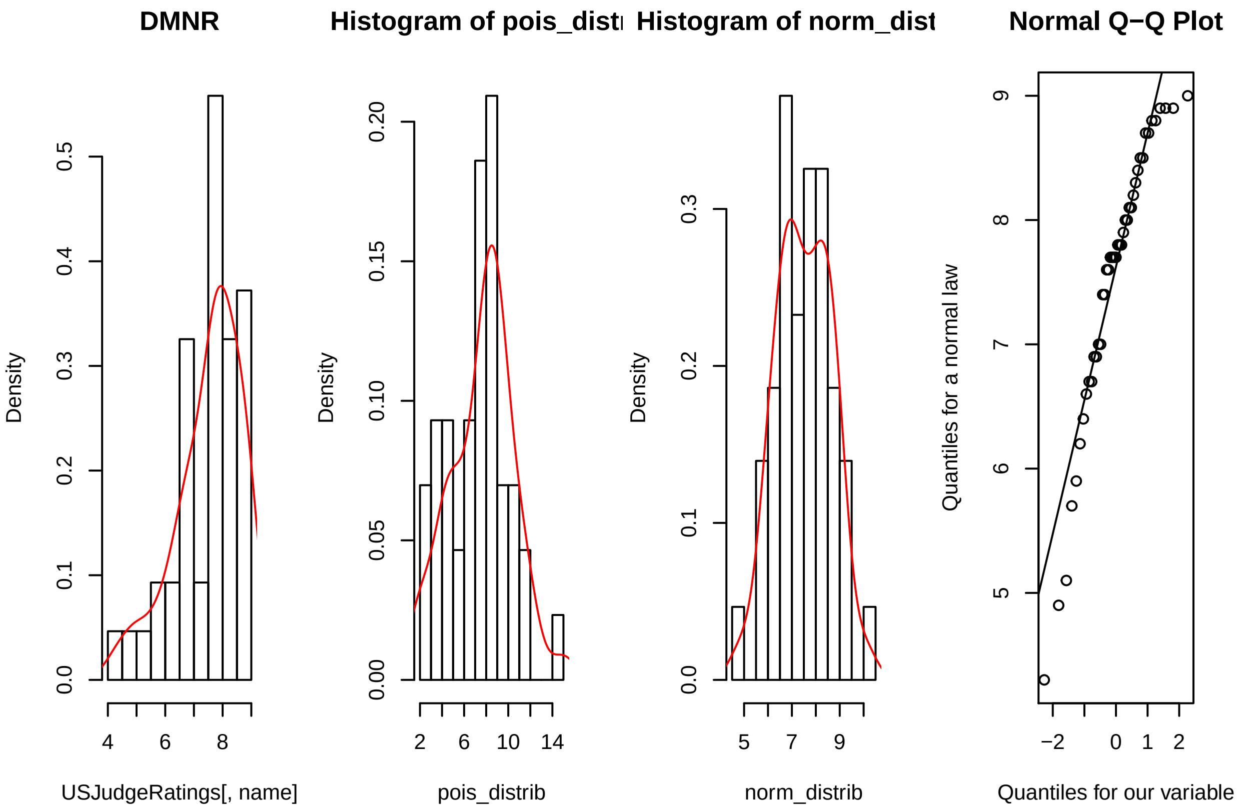
0.98).

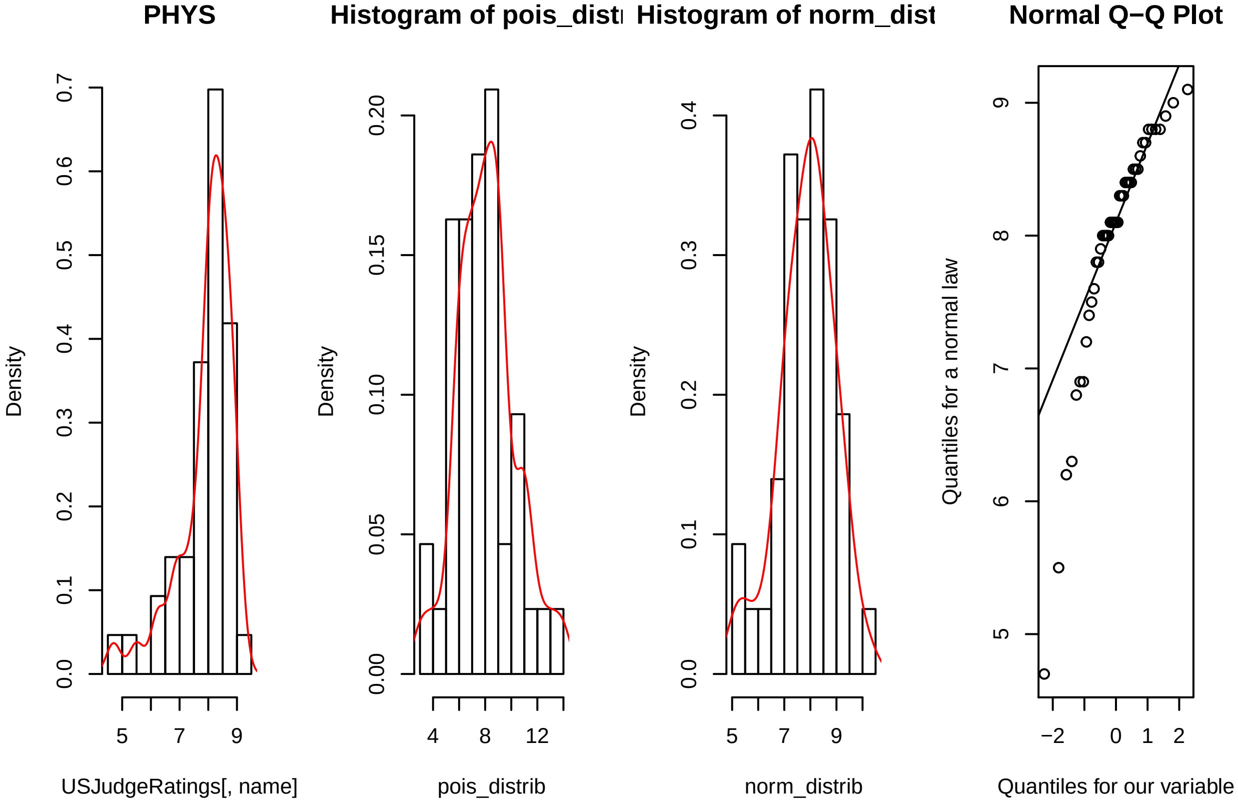
3) Histograms: getting the laws of the essential variables.

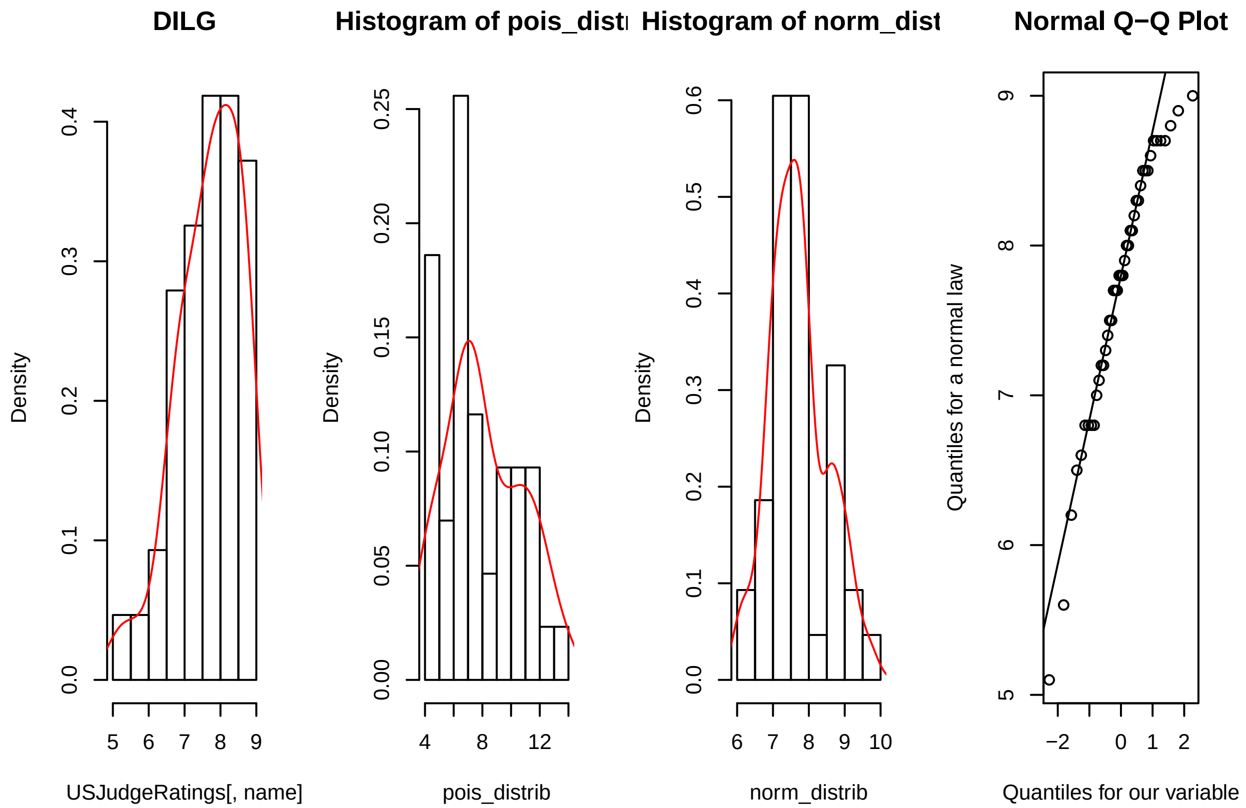
```r
par(mfrow=c(1,4))
names = c("RTEN", "ORAL", "DECI", "DMNR", "PHYS", "DILG", "CONT") # selected RTEN as it is the output;

for (name in names) {

hist(USJudgeRatings[,name], main = name, proba = TRUE, breaks = 10)
d= density(USJudgeRatings[,name])
lines(d, col = 'red')

pois_distrib=rpois(43,mean(USJudgeRatings[,name]))
hist(pois_distrib, prob=TRUE, breaks = 10)
d = density(pois_distrib)
lines(d, col='red')

norm_distrib=rnorm(43,mean(USJudgeRatings[,name]),sqrt(var(USJudgeRatings[,name])))
hist(norm_distrib, prob=TRUE, breaks = 10)
d = density(norm_distrib)
lines(d, col='red')

qqnorm(USJudgeRatings[,name], xlab = "Quantiles for our variable", ylab = "Quantiles for a normal law")
qqline(USJudgeRatings[,name])}
```

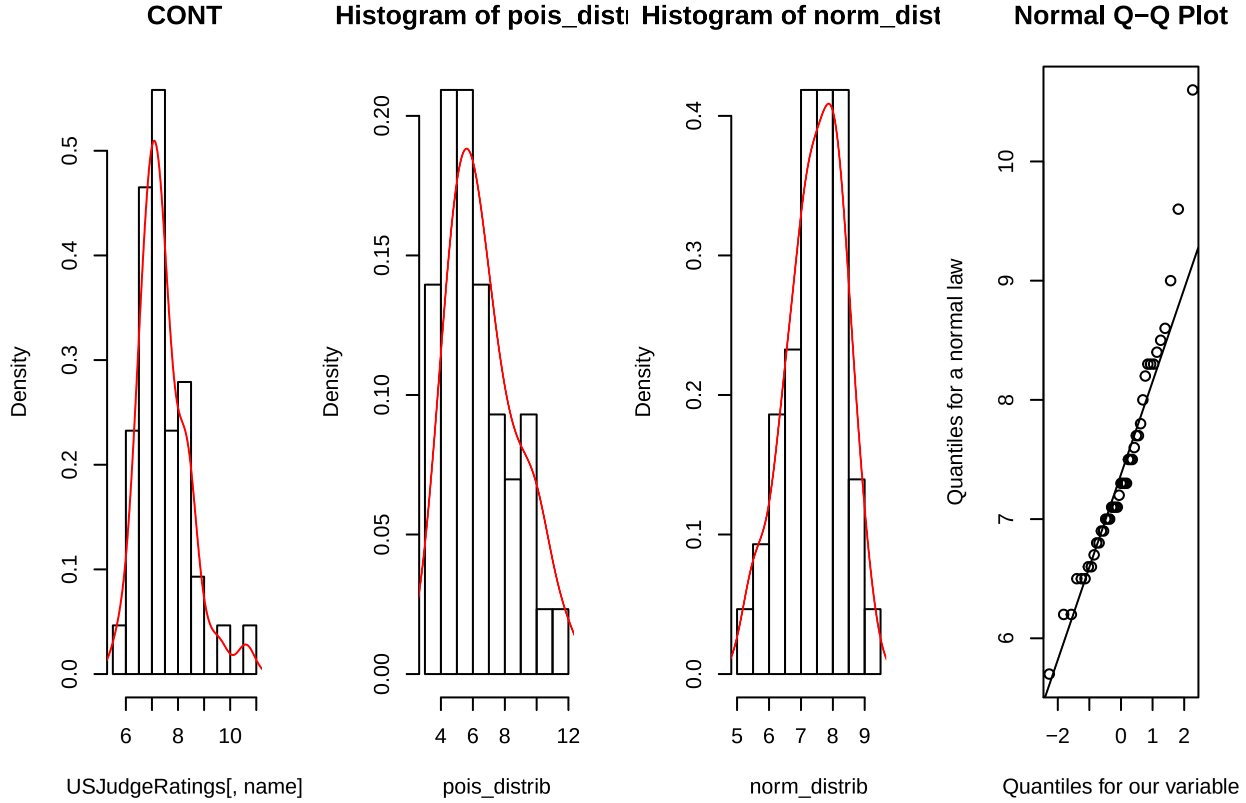



ORAL
Histogram of pois_distı
Histogram of norm_dist
Normal Q−Q Plot
Density
0.0
0.1
0.2
0.3
0.4
0.5
5
6
7
8
9
USJudgeRatings[, name]
0.00
0.05
0.10
0.15
0.20
2
6
10
pois_distrib
0.0
0.1
0.2
0.3
0.4
5
7
9
norm_distrib
Quantiles for a normal law
5
6
7
8
9
−2
0
1
2
Quantiles for our variable
DECI
Histogram of pois_distı
Histogram of norm_dist
Normal Q−Q Plot
0.0
0.1
0.2
0.3
0.4
0.5
5.5
7.0
8.5
USJudgeRatings[, name]
0.00
0.05
0.10
0.15
4
8
12
pois_distrib
0.0
0.2
0.4
0.6
0.8
6.0
7.5
9.0
norm_distrib
6.0
6.5
7.0
7.5
8.0
8.5
−2
0
1
2
Quantiles for our variable


DMNR
Density
0.0
0.1
0.2
0.3
0.4
0.5
4
6
8
USJudgeRatings[, name]
Histogram of pois_distr
Density
0.00
0.05
0.10
0.15
0.20
2
6
10
14
pois_distrib
Histogram of norm_dist
Density
0.0
0.1
0.2
0.3
5
7
9
norm_distrib
Normal Q−Q Plot
Quantiles for a normal law
5
6
7
8
9
−2
0
1
2
Quantiles for our variable


PHYS
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
5
7
9
USJudgeRatings[, name]
Histogram of pois_distr
Density
0.00
0.05
0.10
0.15
0.20
4
8
12
pois_distrib
Histogram of norm_dist
Density
0.0
0.1
0.2
0.3
0.4
5
7
9
norm_distrib
Normal Q−Q Plot
Quantiles for a normal law
5
6
7
8
9
−2
0
1
2
Quantiles for our variable





TO BE COMPLETED

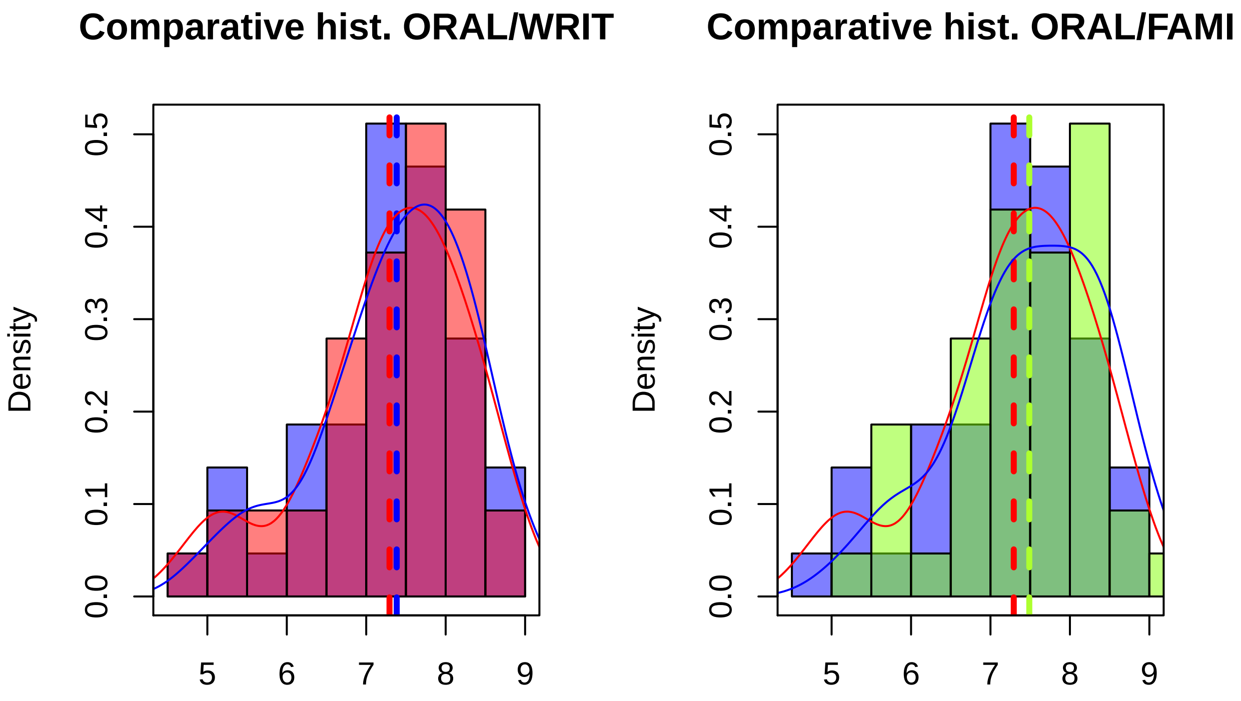
We would rather expect the US Judge Ratings to follow normal laws, given that we can consider the judges

to be iid variables and then apply the CLT to it.
1. First, let us analyse the distribution of RTEN (the output) and ORAL together.
2. DECI 3. DMNR
4. PHYS
5. DILG
6. CONT # normal law

4) Focused multivariate analysis: some interesting combinations to study correlations.

1. ORAL vs WRIT, ORAL vs FAMI

```
par(mfrow=c(1,2))
hist(USJudgeRatings$ORAL, col = rgb(0,0,1,0.5), proba = TRUE, main= "Comparative hist. ORAL/WRIT", breal
hist (USJudgeRatings$WRIT, col = rgb(1,0,0,0.5), proba = TRUE, braks = 15, add = TRUE) # blue
abline(v=mean(USJudgeRatings$ORAL), col="red", lwd=3, lty=2)
abline(v=mean(USJudgeRatings$WRIT), col="blue", lwd=3, lty=2)
d1 = density(USJudgeRatings$ORAL)
lines(d1, col = 'red')
d2 = density(USJudgeRatings$WRIT)
lines(d2, col = 'blue')
box()

hist(USJudgeRatings$ORAL, col = rgb(0,0,1,0.5), proba = TRUE, main= "Comparative hist. ORAL/FAMI", breal
hist (USJudgeRatings$FAMI, col = rgb(0.5,1,0,0.5), proba = TRUE, braks = 15, add = TRUE) #
abline(v=mean(USJudgeRatings$ORAL), col="red", lwd=3, lty=2)
abline(v=mean(USJudgeRatings$FAMI), col="greenyellow", lwd=3, lty=2)
d1 = density(USJudgeRatings$ORAL)
lines(d1, col = 'red')
d2 = density(USJudgeRatings$FAMI)
lines(d2, col = 'blue')
box()
```

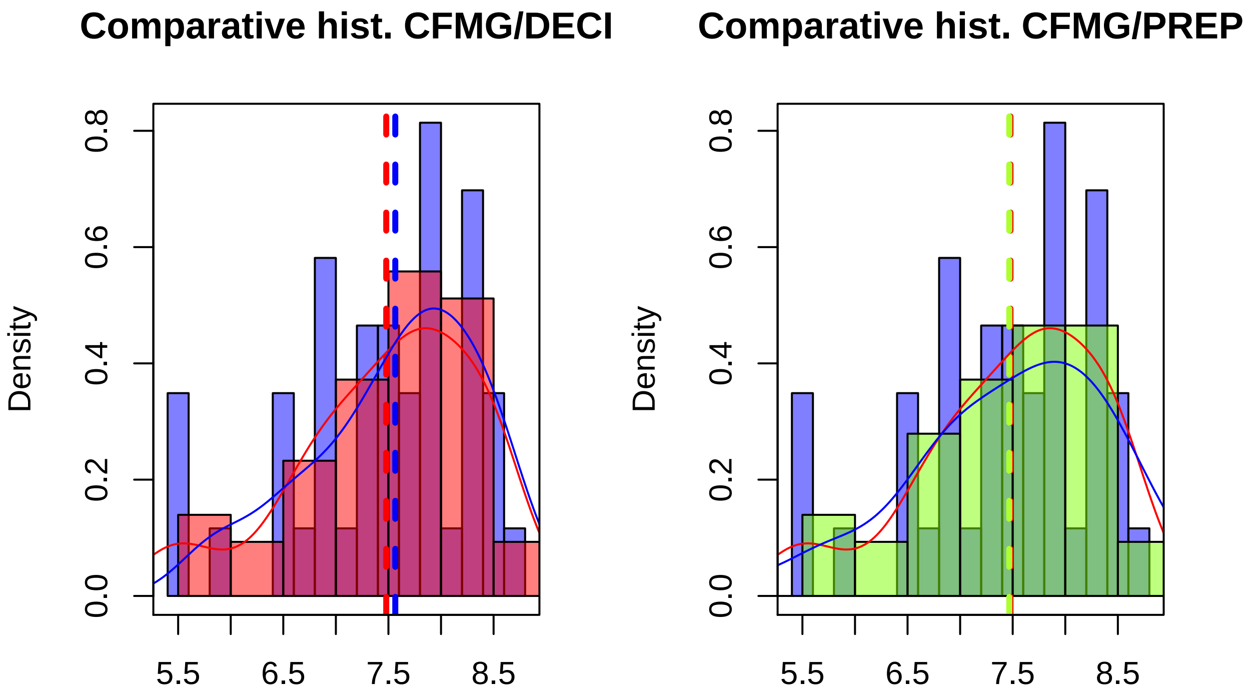


As expected, ORAL and WRIT fit quasi perfectly.

ORAL and FAMI are also very close albeit a little less.

2. CFMG vs DECI, CFMG vs PREP

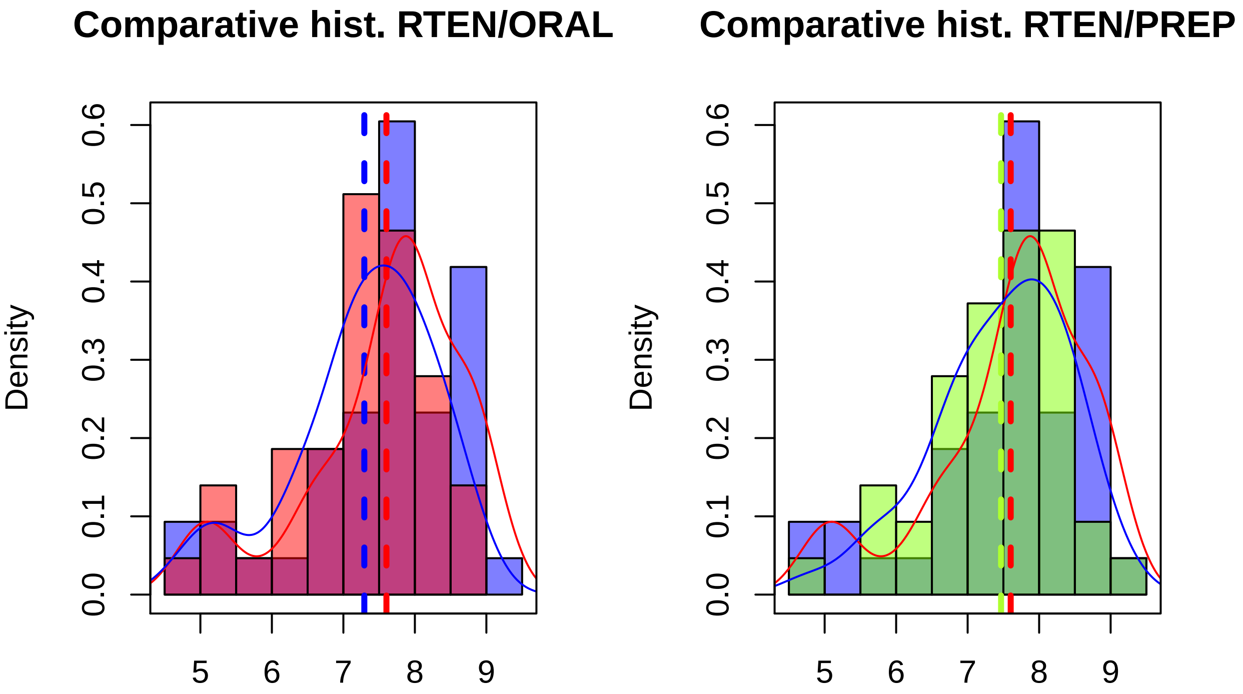
```r
par(mfrow=c(1,2))
hist(USJudgeRatings$CFMG, col = rgb(0,0,1,0.5), proba = TRUE, main= "Comparative hist. CFMG/DECI", breal
hist (USJudgeRatings$DECI, col = rgb(1,0,0,0.5), proba = TRUE, braks = 15, add = TRUE) # blue
abline(v=mean(USJudgeRatings$CFMG), col="red", lwd=3, lty=2)
abline(v=mean(USJudgeRatings$DECI), col="blue", lwd=3, lty=2)
d1 = density(USJudgeRatings$CFMG)
lines(d1, col = 'red')
d2 = density(USJudgeRatings$DECI)
lines(d2, col = 'blue')
box()

hist(USJudgeRatings$CFMG, col = rgb(0,0,1,0.5), proba = TRUE, main= "Comparative hist. CFMG/PREP", breal
hist (USJudgeRatings$PREP, col = rgb(0.5,1,0,0.5), proba = TRUE, braks = 15, add = TRUE) #
abline(v=mean(USJudgeRatings$CFMG), col="red", lwd=3, lty=2)
abline(v=mean(USJudgeRatings$PREP), col="greenyellow", lwd=3, lty=2)
d1 = density(USJudgeRatings$CFMG)
lines(d1, col = 'red')
d2 = density(USJudgeRatings$PREP)
lines(d2, col = 'blue')
box()
```



We confirm closeness of CFMG and DECI, and secondarily CFMG and PREP in terms of distribution.

3. RTEN vs ORAL, RTEN vs PREP.

```r
# ORAL as best explicative variable for RTEN
par(mfrow=c(1,2))
hist(USJudgeRatings$RTEN, col = rgb(0,0,1,0.5), proba = TRUE, main= "Comparative hist. RTEN/ORAL", breal
hist (USJudgeRatings$ORAL, col = rgb(1,0,0,0.5), proba = TRUE, braks = 15, add = TRUE) # blue
abline(v=mean(USJudgeRatings$RTEN), col="red", lwd=3, lty=2)
abline(v=mean(USJudgeRatings$ORAL), col="blue", lwd=3, lty=2)
d1 = density(USJudgeRatings$RTEN)
lines(d1, col = 'red')
d2 = density(USJudgeRatings$ORAL)
lines(d2, col = 'blue')
box()

# PREP as second best explicative variable for RTEN aside WRIT (highly correlated to ORAL)
hist(USJudgeRatings$RTEN, col = rgb(0,0,1,0.5), proba = TRUE, main= "Comparative hist. RTEN/PREP", brea
hist (USJudgeRatings$PREP, col = rgb(0.5,1,0,0.5), proba = TRUE, braks = 15, add = TRUE) #
abline(v=mean(USJudgeRatings$RTEN), col="red", lwd=3, lty=2)
abline(v=mean(USJudgeRatings$PREP), col="greenyellow", lwd=3, lty=2)
d1 = density(USJudgeRatings$RTEN)
lines(d1, col = 'red')
d2 = density(USJudgeRatings$PREP)
lines(d2, col = 'blue')
box()
```



Albeit imperfect, we know that ORAL is approximately the best unique variable explaining RTEN. Indeed, its distribution is close to that of RTEN, and closer than the second best unique variable (except WRIT), PREP.

Conclusion

1. We examined a dataset of grades given by lawyers to US Superior Court judges of which we will assume that the objective is to determine whether these judges were fit to retain, or not (it could've been which ones need training, for instance).
2. Contextual information is lacking, mostly on the objectivity of the measures, and the data is quite old.
3. Most importantly, we do not know how the main question of the experiment will be answered. Feature engineering might be helpful here by deriving a classifier from RTEN: 1 if the judge is retained, 0 if they are not. As discussed in the introduction, the methodology choice is still to be defined. Although

there are more refined options, the two main possibilities are: keeping a number of judges, and then determining the threshold grade to make the cut; or, defining the grade threshold to be met in RTEN, then selecting the judges to keep. We would choose option 1 given that there are probably constraints on the number of judges we want to keep (budgetary or others).

4. We noticed that the variables are generally skewed towards the left-hand-side, as showed by studying their means versus their median as well as skewness.
5. Most variables are highly correlated, which means that if we want to run a model on the dataset, we will have to discriminate closely between those that are relevant towards explaining the data. In particular, WRIT and ORAL are highly correlated. Fast forward selection seems to be the best methodology we could implement towards modelling.
6. The best single variable expaining RTEN, the final grade on which our decision will be based, is ORAL. The second best variable is PREP, but it doesn't mean that it will be included in our model if we use forward selection (forward selection looks at the efficiency of a combination of features).
7.

```r
library(DataExplorer)
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(readr)
library(ggfortify)

#install.packages("tidyverse")
#install.packages("caret")
#install.packages("leaps")
#install.packages("MASS")
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------- tidyverse

## v tibble  2.1.3     v purrr   0.3.2
## v tidyr   1.0.0     v stringr 1.4.0
## v tibble  2.1.3     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------------ tidyverse_confli
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(leaps)
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

Bonus: first steps in choosing a model to fit the data.

```
# Fit the full model
full.model <- lm(RTEN ~., data = USJudgeRatings)
summary(full.model)
```

```
##
## Call:
## lm(formula = RTEN ~ ., data = USJudgeRatings)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22123 -0.06155 -0.01055  0.05045  0.26079
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.11943    0.51904  -4.083 0.000290 ***
## CONT         0.01280    0.02586   0.495 0.624272
## INTG         0.36484    0.12936   2.820 0.008291 **
## DMNR         0.12540    0.08971   1.398 0.172102
## DILG         0.06669    0.14303   0.466 0.644293
## CFMG        -0.19453    0.14779  -1.316 0.197735
## DECI         0.27829    0.13826   2.013 0.052883 .
## PREP        -0.00196    0.24001  -0.008 0.993536
## FAMI        -0.13579    0.26725  -0.508 0.614972
## ORAL         0.54782    0.27725   1.976 0.057121 .
## WRIT        -0.06806    0.31485  -0.216 0.830269
## PHYS         0.26881    0.06213   4.326 0.000146 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1174 on 31 degrees of freedom
## Multiple R-squared:  0.9916, Adjusted R-squared:  0.9886
## F-statistic: 332.9 on 11 and 31 DF,  p-value: < 2.2e-16
```

```
# Fast forward selection with 5 features
models <- regsubsets(RTEN~., data = USJudgeRatings, nvmax = 5,
                     method = "forward")
summary(models)
```

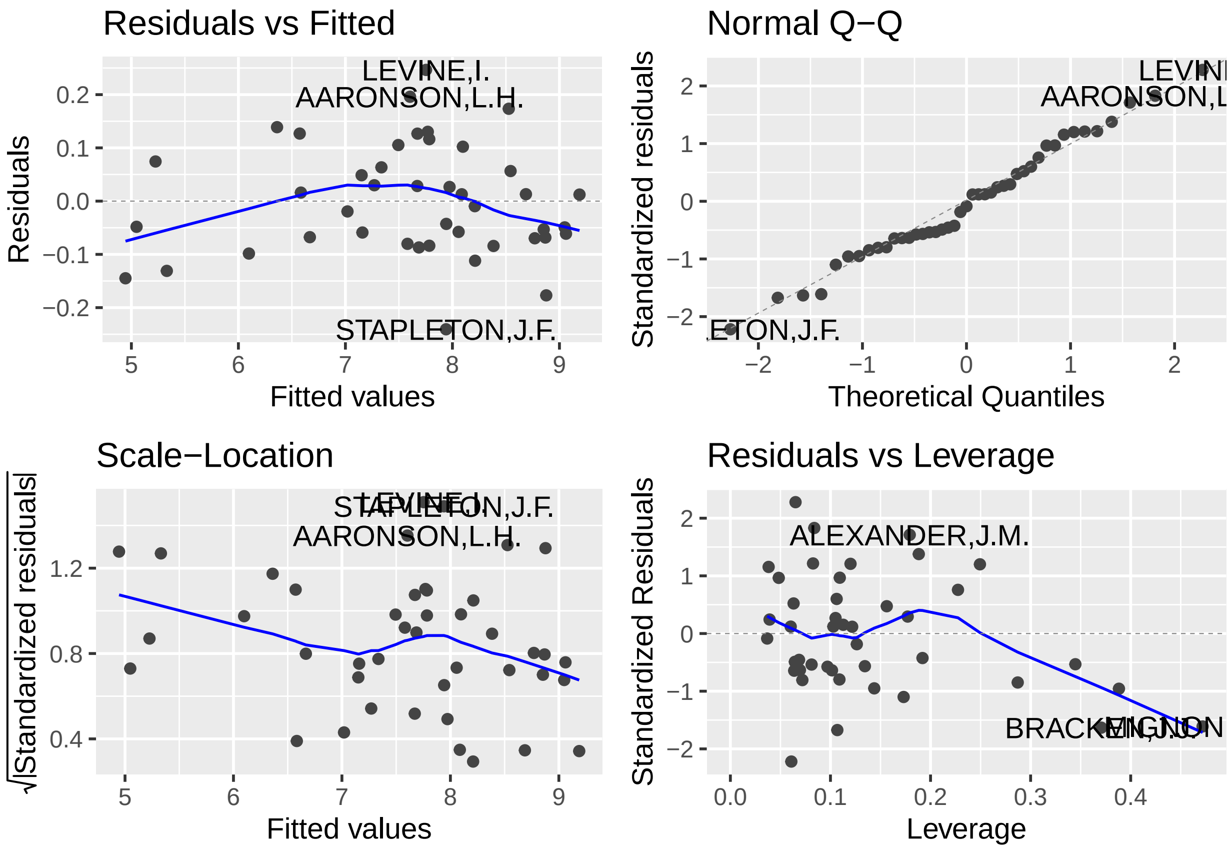
```
## Subset selection object
## Call: regsubsets.formula(RTEN ~ ., data = USJudgeRatings, nvmax = 5,
##     method = "forward")
## 11 Variables  (and intercept)
##      Forced in Forced out
## CONT     FALSE      FALSE
## INTG     FALSE      FALSE
## DMNR     FALSE      FALSE
## DILG     FALSE      FALSE
## CFMG     FALSE      FALSE
## DECI     FALSE      FALSE
## PREP     FALSE      FALSE
## FAMI     FALSE      FALSE
## ORAL     FALSE      FALSE
## WRIT     FALSE      FALSE
## PHYS     FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: forward
##          CONT INTG DMNR DILG CFMG DECI PREP FAMI ORAL WRIT PHYS
## 1  ( 1 ) " "  " "  " "  " "  " "  " "  " "  " "  "*"  " "  " "
## 2  ( 1 ) " "  " "  "*"  " "  " "  " "  " "  " "  "*"  " "  " "
## 3  ( 1 ) " "  " "  "*"  " "  " "  " "  " "  " "  "*"  " "  "*"
## 4  ( 1 ) " "  "*"  "*"  " "  " "  " "  " "  " "  "*"  " "  "*"
## 5  ( 1 ) " "  "*"  "*"  " "  " "  "*"  " "  " "  "*"  " "  "*"
```

```
# Creating the model with the 5 features selected by forward regression.
forward.model = lm(RTEN ~ ORAL + PHYS + INTG + DECI + DMNR, data = USJudgeRatings)
summary(forward.model)
```

```
##
## Call:
## lm(formula = RTEN ~ ORAL + PHYS + INTG + DECI + DMNR, data = USJudgeRatings)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.240656 -0.069026 -0.009474  0.068961  0.246402
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.20433    0.43611  -5.055 1.19e-05 ***
## ORAL         0.29169    0.10191   2.862 0.006887 **
## PHYS         0.28292    0.04678   6.048 5.40e-07 ***
## INTG         0.37785    0.10559   3.579 0.000986 ***
## DECI         0.16672    0.07702   2.165 0.036928 *
## DMNR         0.15199    0.06354   2.392 0.021957 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1119 on 37 degrees of freedom
## Multiple R-squared:  0.9909, Adjusted R-squared:  0.9897
## F-statistic: 806.1 on 5 and 37 DF,  p-value: < 2.2e-16
```

In fact the important KPI to watch here is the adjusted R-squared. 0.9886 for the full model, and 0.9897 for our forward selection model.
So, both models have a very good fit, but the forward selection model with 5 variables is in fact better at explaining the data without using too many variables!

```
autoplot(forward.model)
```



Our standardised residuals are in between 0.6 and 1.0 approximately (well within the [-2,2] range), which means that homsedasticity is a priori verified.

We could choose the log of the sqrt of RTEN to smooth the outliers. The next step would be to determine a classifier derived from RTEN, but that is not the objective of our exercise.