Test di Normalità di Shapiro-Wilks

- La verifica della normalità avviene confrontando due stimatori alternativi della varianza σ^2: uno stimatore non parametrico basato sulla combinazione lineare dei percentili di una variabile aleatoria normale al numeratore, e il consueto stimatore parametrico, ossia la varianza campionaria, al denominatore.
$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
dove $x_{(i)}$ è l'i-esimo valore più piccolo (rango i) del campione, $\overline{x} = (x_1 + \cdots + x_n)/n$ è la media aritmetica del campione e le costanti a_i sono ottenute tenendo conto di una stima della variabilità casuale
- La statistica W può assumere valori da 0 a 1. Qualora il valore della statistica W sia "molto" piccolo, il test rifiuta l'ipotesi nulla che i valori campionari siano distribuiti come una variabile casuale normale.
- La statistica W può essere interpretata come il quadrato del coefficiente di correlazione in un diagramma quantile-quantile cioè calcolato tra quantili empirici e quantili teorici della normale standardizzata.

Test di Normalità di Shapiro-Wilks

Esempio delle altezze degli alberi su scala log

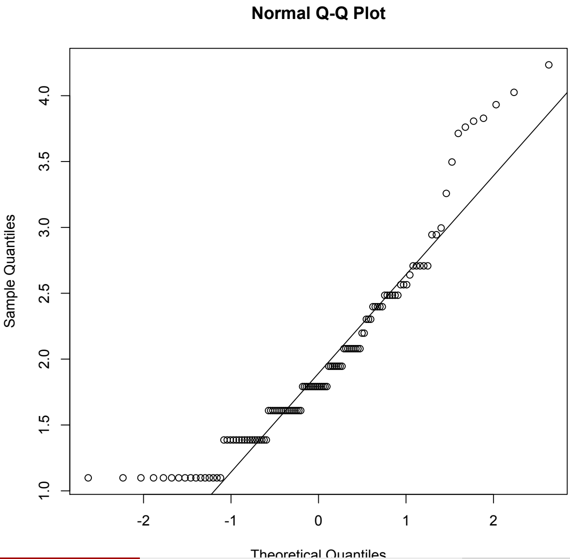
```
pp=seq(0.0001,0.8,length=100) #sequenza dei percentili da calcolare
 qn1=qnorm(pp)#quantili teorici
qa=quantile(alberi.log[,2],prob=pp) #quantili osservati
cor(qa,qn1)^2 #correlazione tra quantili osservati e quantili teori
[1] 0.8540894

shapiro.test(alberi.log[,2])

Shapiro-Wilk normality test

data:  alberi.log[, 2]
W = 0.8979, p-value = 1.865e-07
```

Test di Normalità di Shapiro-Wilks



Test del χ^2 per tabelle di contingenza

Alcohol (oz/day)	Nicotine (mg/day) None	1 to 15	Over 15	
None	105	7	11	123
0.01 to 0.10	58	5	13	76
0.11 to 0.99	84	37	42	163
1.00 or more	57	16	17	90
Tot.	304	65	83	452

Domanda: Il consumo di nicotina e il consumo di alcool dipendono l'uno dall'altro?

H_0: sono indipendenti

H_1: sono dipendenti

Test del χ^2 per tabelle di contingenza

Per poter verificare se H_0 è vera o meno costruisco la *tabella di indipendenza* o *tabella dei valori attesi*:

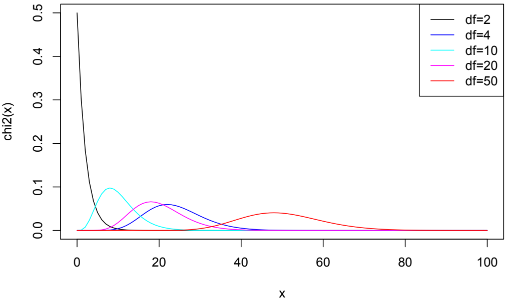
Alcohol (oz/day)	Nicotine (mg/day) None	1 to 15	Over 15	
None	$\frac{123\cdot304}{452}=82.73$	$\frac{123\cdot65}{452}=17.688$	$\frac{123\cdot83}{452}=22.586$	123
0.01 to 0.10	$\frac{76\cdot304}{452}=51.115$	$\frac{76\cdot65}{452}=10.929$	$\frac{76\cdot83}{452}=13.956$	76
0.11 to 0.99	$\frac{163\cdot304}{452}=109.628$	$\frac{163\cdot65}{452}=23.440$	$\frac{163\cdot83}{452}=29.931$	163
1.00 or more	$\frac{90\cdot304}{452}=60.531$	$\frac{90\cdot65}{452}=12.942$	$\frac{90\cdot83}{452}=16.527$	90
Tot.	304	65	83	452

Calcolo la distanza media tra questa tabella e quella osservata:

$$\chi^2 = \sum \frac{(oss. - atteso)^2}{atteso}$$

Questa è la statistica test che nel nostro esempio vale $\chi^2 = 42.25$. Questa statistica va confrontata con la *distribuzione chi-quadrato*

Test del χ^2 per tabelle di contingenza



distribuzione chi-quadro con diversi gradi di libertà. I gradi di libertà sono determinati dal numero di righe (r) e di colonne (c) della tabella che stiamo studiando, ovvero $df = (r-1)(c-1)$. Quindi nel nostro esempio $\chi^2_{(4-1)(3-1)=6}$. Il $p - value = 1.64 \cdot 10^{-07}$ concludiamo che c'è dipendenza.