Clase 8 y 9. Correlación

Simoneta Negrete Yankelevich

R.1

Vamos a ver una serie de datos para ver si existe una relación lineal entre ellos.

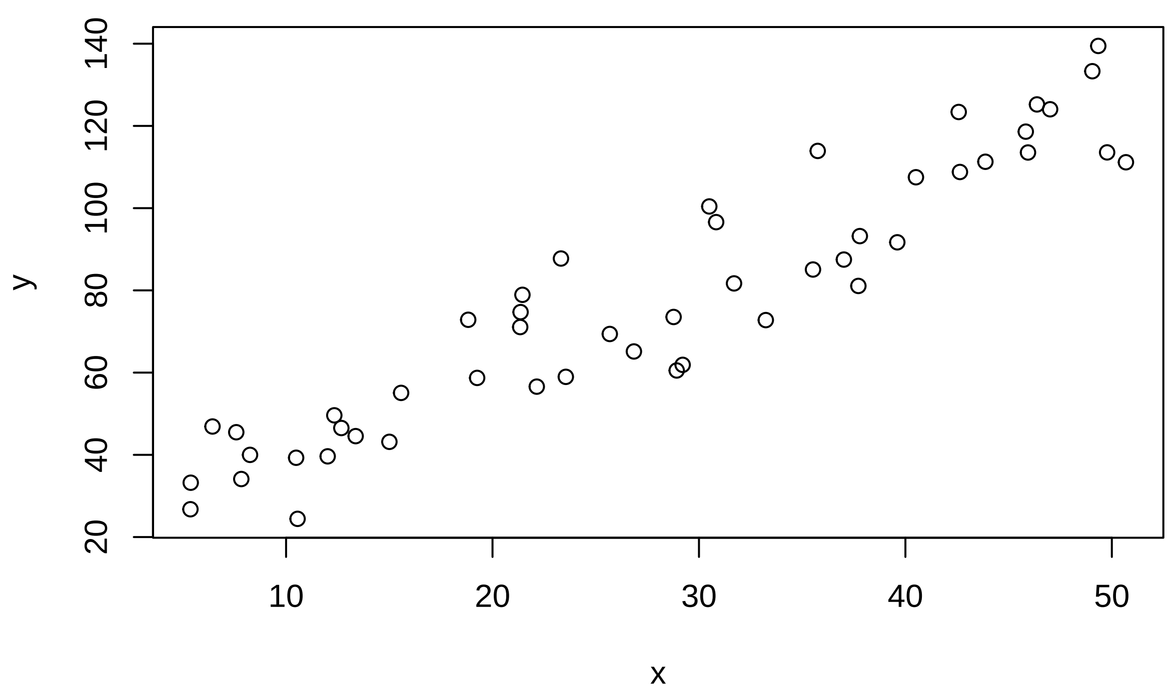
```
data<-read.table("twosample.txt",header=T)
attach(data)
data
```

```
##            x        y   a      b
## 1   5.366516 26.76595 one   male
## 2   6.435778 46.89376 one   male
## 3   7.831232 34.11415 one female
## 4   7.587142 45.49667 one female
## 5   5.380939 33.22162 one   male
## 6   8.254098 39.98920 one female
## 7  10.556489 24.43327 one female
## 8  12.336669 49.61092 one   male
## 9  10.487217 39.29021 one   male
## 10 12.014185 39.63691 one   male
## 11 13.369883 44.54903 two   male
```

```
## 12 12.677217  46.51998   two   male
## 13 15.004940  43.16512   two   male
## 14 15.571449  55.06652   two   male
## 15 19.254838  58.70913   two female
## 16 18.822350  72.85045   two   male
## 17 21.358761  74.71627   two   male
## 18 21.449881  78.92278   two   male
## 19 21.341156  71.08121   two   male
## 20 23.551192  58.97132   two female
## 21 23.315863  87.74729 three   male
## 22 22.142836  56.59044 three female
## 23 26.848780  65.14307 three   male
## 24 28.921049  60.52122 three female
## 25 25.680903  69.39125 three   male
## 26 29.209522  61.89014 three female
## 27 28.770843  73.52917 three   male
## 28 30.500931 100.41251 three female
## 29 31.698786  81.69903 three   male
## 30 30.832011  96.59936 three female
## 31 33.237294  72.77412  four   male
## 32 35.523952  85.07512  four   male
## 33 37.792353  93.20267  four   male
## 34 35.752220 113.91521  four female
## 35 39.607470  91.69301  four female
## 36 37.018701  87.49319  four female
## 37 37.721269  81.07711  four female
## 38 42.640823 108.79472  four female
```

```
## 39 42.582064 123.38632  four female
## 40 40.510316 107.51459  four   male
## 41 43.873717 111.28685  five   male
## 42 45.831377 118.60641  five female
## 43 47.009356 124.04675  five female
## 44 45.938572 113.53865  five female
## 45 49.770829 113.56021  five female
## 46 46.372772 125.22990  five female
## 47 49.054437 133.30422  five female
## 48 50.682444 111.15864  five female
## 49 49.341443 139.45162  five   male
```

```
plot(x,y)
```



Se acuerdan que necesitamos primero para calcular el coeficiente de correlación de pearson? Las varianzas individuales

```
var(x)
```

```
## [1] 199.9837
```

```
var(y)
```

```
## [1] 977.0153
```

¿y que más? La covarianza y estamos hechos

```
var(x,y)
```

```
## [1] 414.9603
```

Ahora calculamos r

```
var(x,y)/sqrt(var(x)*var(y))
```

```
## [1] 0.9387684
```

Ahora hagamoslo en automático

```
cor(x,y)
```

```
## [1] 0.9387684
```

Y ahora hagamos la prueba de hipótesis

Calculamos EE de r

```r
EEr<-((1-(cor(x,y)^2))/(length(x)-2))^0.5
EEr
```

```
## [1] 0.05025759
```

Calculo t de la muestra

```r
te<-cor(x,y)/EEr
te
```

```
## [1] 18.67914
```

Calculo t de tablas

```r
qt(0.975,47)
```

```
## [1] 2.011741
```

Calculo la p

```r
2*(1-pt(18.67914,47))
```

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

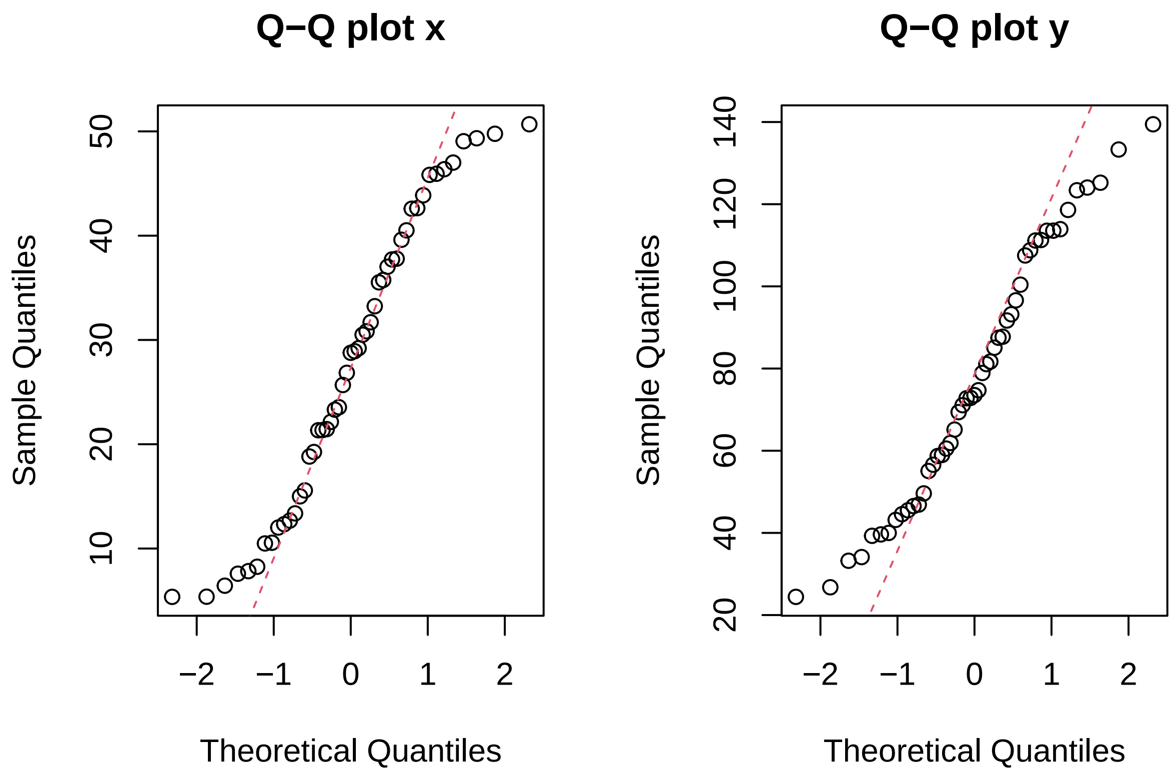
Ahora hagamoslo de manera automática

```r
pearson<-cor.test(x,y)
pearson
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 18.679, df = 47, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8934139 0.9651786
## sample estimates:
##       cor
## 0.9387684
```

¿Que nos falta?. Pues no sabemos si cumplimos con los supuestos. Veamos el de normalidad

```r
par(mfrow=c(1,2))
qqnorm(x, main="Q-Q plot x"); qqline(x, col = 2, lty = 2)
qqnorm(y, main="Q-Q plot y"); qqline(y, col = 2, lty = 2)
```



¿Que opciones tengo?.

1. Hacer una prueba de sesgo y kurtosis para ver si estas desviaciones son significativas

2. Si son significativas, puedo intentar transformaciones o puedo utilizar muchas de las otras pruebas de correlación que son robustas a la violación de este supuesto. *Vean Q y k p.76 y Crawley p.97-102.*

Fin