Correlación en R

Luis Cárceles

2024-04-07

Introducción

Como se calcula la correlación y los test de correlación. Se utilizan los ejemplos del libro: "Probabilidad y estadística" de J.Susan Milton y Jesee C. Arnold

Estimador para el coeficiente de correlación de Pearson:

$$\hat{\rho} = R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Donde: $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ y $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

Prueba de hipótesis

Para la prueba de hipótesis se suele utilizar el estadistico:

$$T_{n-2} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

Donde R es el coeficiente de correlación estimado con la fórmula anterior y n es el número de puntos. T_{n-2} se distribuye como una T de Student con $n-2$ grados de libertad.

Intervalo de confianza.

El intervalo de confianza $100(1-\alpha)\%$ para ρ se calcula mediante las ecuaciones:

$$Límite\ Inferior = \frac{(1+R) - (1-R)exp(2z_{\alpha/2}\sqrt{n-3})}{(1+R) + (1-R)exp(2z_{\alpha/2}\sqrt{n-3})}$$

$$Límite\ Superior = \frac{(1+R) - (1-R)exp(-2z_{\alpha/2}\sqrt{n-3})}{(1+R) + (1-R)exp(-2z_{\alpha/2}\sqrt{n-3})}$$

Donde R es el coeficiente de correlación estimado ($\hat{\rho}$), n es el número de puntos y z se distribuye Normal, con media 0 y desviación 1.

Ejemplos

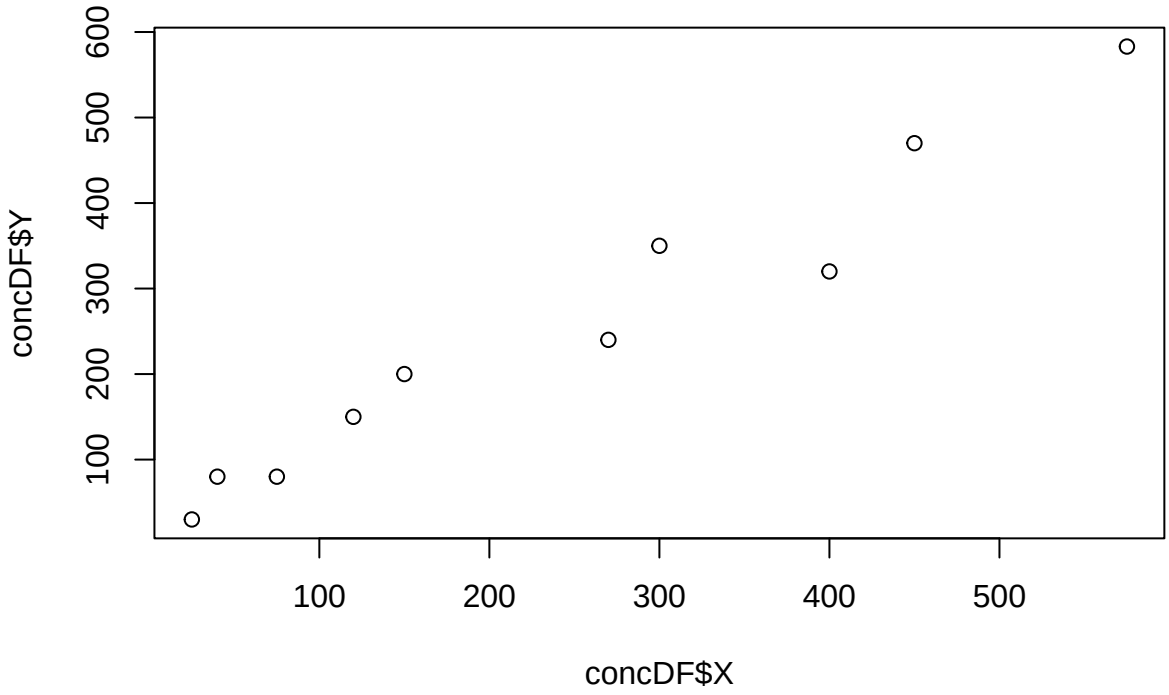
Ejemplo 1:

En un estudio del efecto del efluente de aguas negras en un lago, los investigadores midenla concentración de nitrato en el agua. Un antiguo método manual se ha utilizado para cuantificar esa variable. Sin embargo, se diseñó un nuevo método automatizado. Si existe una correlación positiva alta entre las mediciones tomadas con los dos métodos, se pondrá el uso habitual el automatizado. Se obtienen los datos en el data frame adjunto, sobre la concentración de nitrato, en microgramos de intrato por litro de agua:

```
load("~/10_Data_Analysis/01_The_Math_Sorcerer/E1_Correlacion_Regresion.RData")
concDF
```

```
##       X   Y
## 1    25  30
## 2    40  80
## 3   120 150
## 4    75  80
## 5   150 200
## 6   300 350
## 7   270 240
## 8   400 320
## 9   450 470
## 10 575 583
```

```
plot(concDF$X, concDF$Y)
```



Estimador para el coeficiente de correlación de Pearson

```
Sxx<-sum((concDF$X-mean(concDF$X))^2)
Syy<-sum((concDF$Y-mean(concDF$Y))^2)
Sxy<-sum((concDF$X-mean(concDF$X))*(concDF$Y-mean(concDF$Y)))
R<-Sxy/sqrt(Sxx*Syy)
n<-length(concDF$X)
sprintf("El coeficiente de correlación es: %f", R)
```

```
## [1] "El coeficiente de correlación es: 0.977790"
```

Cálculo de la correlación en R:

```
sprintf("En R: %f", cor(concDF$X, concDF$Y))
```

```
## [1] "En R: 0.977790"
```

En este ejemplo, se estima la correlacion entre X, la medición manual, e Y la medición automática. Se ha obtenido un coeficiente de correlación alto por lo que parece que $\rho \neq 0$. Sin embargo el tamaño de la muestra es pequeño por lo que debe ponerse a prueba:

Prueba de hipótesis:

- H_0: $\rho = 0$
- H_1: $\rho \neq 0$

Calculamos el estadístico:

```
t<-R*sqrt(n-2)/sqrt(1-R^2)
sprintf("El valor del estadístico es: %f", t)
```

```
## [1] "El valor del estadístico es: 13.195483"
```

el valor p para este estadístico es:

```
p_value<-2*pt(q = t, df = n-2 , lower.tail = FALSE)
sprintf("The p-valor del estadístico es: %f", p_value)
```

```
## [1] "The p-valor del estadístico es: 0.000001"
```

Como el p-valor es menor que el 1% de nivel de significación, rechazamos H_0. Hay evidencias suficientes para aceptar que los datos X e Y están correlacionados.

Intervalo de confianza del 95%:

```
alpha<-1-0.95
z<-qnorm(alpha/2, 0, 1, lower.tail=FALSE)
lowLimit<-((1+R)-(1-R)*exp(2*z/sqrt(n-3)))/((1+R)+(1-R)*exp(2*z/sqrt(n-3)))
highLimit<-((1+R)-(1-R)*exp(-2*z/sqrt(n-3)))/((1+R)+(1-R)*exp(-2*z/sqrt(n-3)))
sprintf("El límite inferior del inervalo de confianza es: %f", lowLimit)
```

```
## [1] "El límite inferior del inervalo de confianza es: 0.905832"
```

```
sprintf("El límite superior del inervalo de confianza es: %f", highLimit)
```

```
## [1] "El límite superior del inervalo de confianza es: 0.994909"
```

Si utilizamos el test de correlación integrado en R:

```
cor.test(concDF$X, concDF$Y, alternative="two.sided", method="pearson", conf.level=0.95)
```

```
##
##  Pearson's product-moment correlation
##
## data:  concDF$X and concDF$Y
## t = 13.195, df = 8, p-value = 1.036e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9058321 0.9949085
## sample estimates:
##       cor
## 0.9777899
```