Laboratorio 5: Correlación

Dr. Marco Aurelio González Tagle

26 /08/ 2021

Índice

Instrucciones

Para cada ejercicio

- Examinar la relación que existe entre dos muestras mediante una correlación,
- Explore los datos gráficamente y explique,
- Establezca la Hipótesis nula y la Hipótesis alternativa,
- Aplique la prueba correspondiente,
- Reporte los datos (indicar valor de r, grados de libertad y probabilidad, así como la significancia de la correlación).

Ejercicio 1: El cuarteto de Anscombe

El **cuarteto de Anscombe** comprende cuatro conjuntos de datos que tienen estadísticas descriptivas simples casi idénticas (Cuadro 4), pero tienen distribuciones muy diferentes y parecen muy diferentes cuando se grafican (Figura 1). Cada conjunto de datos consta de once puntos (x, y). Fueron construidos en 1973 por el estadístico Francis Anscombe para demostrar tanto la importancia de graficar los datos antes de analizarlos como el efecto de los valores atípicos y otras observaciones influyentes sobre las propiedades estadísticas.

Descripción de las gráficas:

> El primer gráfico de dispersión (arriba a la izquierda) parece ser una relación lineal simple, correspondiente a dos variables correlacionadas donde y podría modelarse como gaussiana con una media linealmente dependiente de x.
>
> El segundo gráfico (arriba a la derecha) no se distribuye normalmente; mientras que una relación entre las dos variables es obvia, no es lineal y el coeficiente de correlación de Pearson no es significativo. Sería más apropiado una regresión más general y el correspondiente coeficiente de determinación.
>
> En el tercer gráfico (abajo a la izquierda), la distribución es lineal, pero debería tener una línea de regresión diferente (habría sido necesaria una regresión

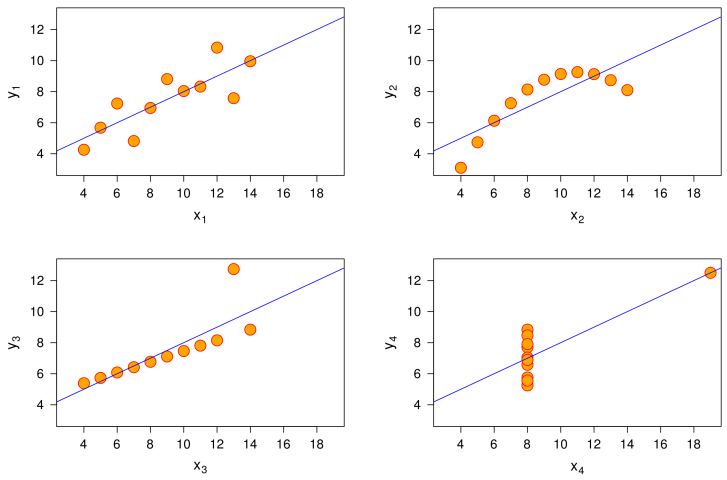

Figura 1: Los cuatro conjuntos son idénticos cuando se examinan utilizando estadísticas de resumen simples, pero varían considerablemente cuando se representan.

robusta). La regresión calculada se compensa con el valor atípico que ejerce suficiente influencia para reducir el coeficiente de correlación de 1 a 0.816.

Finalmente, el cuarto gráfico (abajo a la derecha) muestra un ejemplo en el que un punto en lo alto es suficiente para producir un alto coeficiente de correlación, aunque los otros puntos de datos no indican ninguna relación entre las variables.

Cuadro 1: Propiedades estadísticas de los cuatro grupos de datos.

Propiedad	Valor	Exactitud
Media de x	9	exacto
Varianza de x : σ^2	11	exacto
Media de y	7.50	hasta dos decimales
Varianza de y : σ^2	4.125	± 0.003
Correlación entre x and y	0.816	hasta tres decimales
Línea de regresión	y = 3.00 + 0.500x	dos y tres decimales
Coeficiente de determinación: R^2	0.67	hasta dos decimales

El cuadro 2 se muestran los datos de cada conjunto, la idea es replicarlos en R para afianzar la importancia de la inspección visual cuando se recibe un conjunto de datos.

Cuadro 2: Los conjuntos de datos son los siguientes. Los valores de x son los mismos para los primeros tres conjuntos de datos.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.8

Actividades

- Generar los gráficos de distribución de puntos para cada par de datos

```
# Graficar en un cuadro de 2x2
op = par(mfrow = c(2, 2), mar = c(4.5, 4, 1, 1))
plot(anscombe$x1, anscombe$y1, pch = 20)
plot(anscombe$x2, anscombe$y2, pch = 20)
plot(anscombe$x3, anscombe$y3, pch = 20)
plot(anscombe$x4, anscombe$y4, pch = 20)
par(op)
```

Coeficiente de correlación

Determinar para cada par de datos los coeficientes de correlación r.

¿Alguna sorpresa? Como puedes ver, los cuatro pares de las variables xy tienen básicamente la misma correlación de *0.816*. Pero no todos tienen diagramas de dispersión en los que los puntos se agrupan alrededor de una línea.

El mensaje para llevar a casa es que el coeficiente de correlación puede ser engañoso en presencia de valores atípicos o asociación no lineal. Debido a esto, siempre es importante revisar los datos con un gráfico de dispersión.