Probabilidad y Estadística para Análisis de Datos - Solución al Taller No. 2

J. Mauricio Mejía Castro

11 de marzo de 2021

1. Resumen de Información en Tablas

El Cuadro 1 compila el cálculo de la media, la mediana y el coeficiente de variación para las ocho variables numéricas del conjunto de datos. Además, estos resultados se encuentran agrupados por la variable `class`.
De acuerdo con el Cuadro 1 pueden observarse diferencias significativas de las variables con respecto a la clase. Por ejemplo, es claro que para aquellas mujeres con resultado negativo para diabetes, la media de las variables es menor. Un fenómento similar ocurre con las medianas. No obstante, se observa también que el coeficiente de variación es mayor para las mujeres con resultado negativo para diabetes: existe mayor incertidumbre en los datos de esta clase.

2. Resumen de Información en Gráficas

La Figura 1 contiene un diagrama de dispersión entre el índice de masa corporal y el espesor del pliegue cutáneo ubicado sobre el músculo triceps. La gráfica refleja cierta linealidad difusa entre las dos variables. Cabe destacar también el elevado número de valores inexistentes para la variable `skin`.

Además del grafico de dispersión, se construyeron también los diagrama de caja de la Figura 2, debido a su capacidad para identificar valores atípicos. Allí se podrá apreciar la distribución de los datos, teniendo en cuanta la clase. Es posible ver como el espesor del pliegue cutáneo tiende a ser menor en las mujeres con resultado positivo para diabetes y como estas mismas mujeres tienden a registrar mayores IMC.

class	Negativo para diabetes	Positivo para diabetes
preg_media	3.298	4.86567164179105
plas_media	109.98	141.257462686567
pres_media	68.184	70.8246268656716
skin_media	19.664	22.1641791044776
test_media	68.792	100.335820895522
mass_media	30.3042	35.1425373134328
pedi_media	0.429734	0.5505
age_media	31.19	37.0671641791045
preg_mediana	2	4
plas_mediana	107	140
pres_mediana	70	74
skin_mediana	21	27
test_mediana	39	0
mass_mediana	30.05	34.25
pedi_mediana	0.336	0.449
age_mediana	27	36
preg_coevar	91.4852814621555	76.8904956904246
plas_coevar	23.7690486955388	22.6109272038089
pres_coevar	26.4916628729699	30.3451110181861
skin_coevar	75.7218628648506	79.7670480694412
test_coevar	143.716259582972	138.224936511905
mass_coevar	25.3755420425225	20.6671680464296
pedi_coevar	69.5977754511886	67.6393248963871
age_coevar	37.4083193062878	29.5902152087236

Cuadro 1: Cálculo de estadísticas descriptivas agrupado por clase para todas las variables del conjunto de datos.

3. Exploración de correlaciones

Para esta sección resulta valioso observar lás gráficas de densidad de las variables `mass` y `skin` que se encuentran en la Figura 3.

A partir de estás gráficas es claro que las variables no tienden a distribuirse normalmente. Por consiguiente, no es recomendable utilizar el coeficiente de Pearson para determinar la correclación entre ellas. Es por esta razón que se prefiere el coeficiente de Kendall.

Dado que se desconoce la razón de aquellos valores, que marcan cero para `mass` y `skin`, se opta por seguir incluyendolos en el análisis. En este punto, solo la asesoría de una opinion experta en el conjunto de datos podría

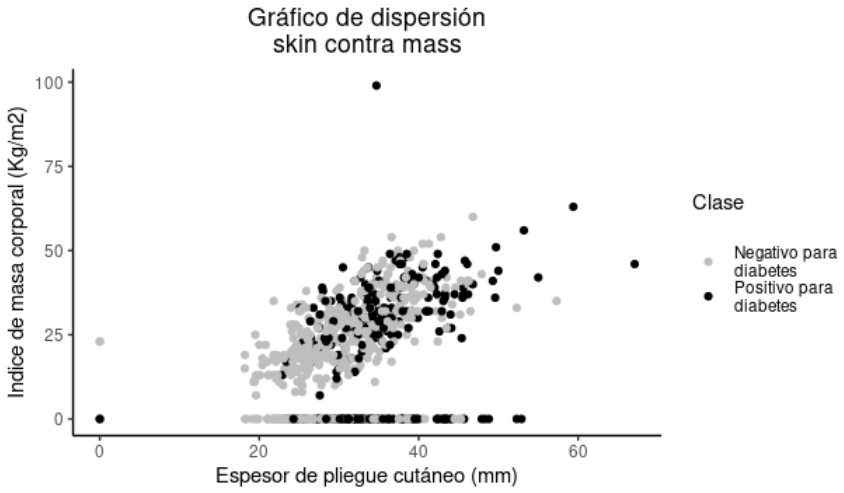

Figura 1: Gráfica de dispersión para las variables `mass` y `skin`.

disuadir esta decisión.

Al ejecutar el test de correlación en R, obtenemos el siguiente resultado:

```
> cor.test(diabetis$mass, diabetis$skin, method = "kendall")

Kendall's rank correlation tau

data:  data.diabetis$mass and data.diabetis$skin
z = 13.18, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.3315319
```

Este test arroja un `p-value` muy bajo. No obstante, no secuentan en este punto con herramientas más formalesque permitan decidir sieste valor es significativamente diferente de cero. Por consiguiente, aun no puede concluirse nada sobre la correlación entre estas variables.

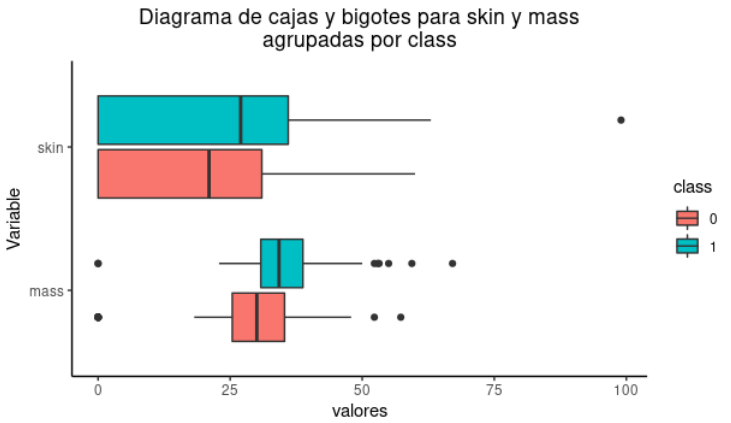

Figura 2: Diagrama de cajas y bigotes para las variables `mass` y `skin`.

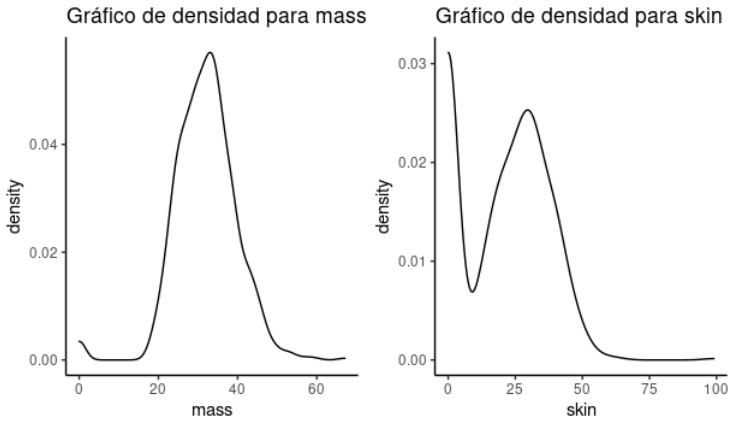

Figura 3: Distribuciones de las variables `mass` y `skin`.