CORRELACIÓN LINEAL

El análisis de correlación lineal se encuentra estrechamente vinculado con el análisis de regresion lineal y ambos pueden ser considerados cómo dos estrategias de análisis de un mismo problema o estudio, en donde lo que cambia es el objetivo del análisis.

El objetivo en un análisis de correlación se centra en medir y evaluar el grado de asociación lineal que hay entre dos variables cuantitativas. Acá no interesa si una variable es la causa de otra variable, a diferencia del análisis de regresión. En el análisis de correlación no existe relación causa – efecto entre las variables. En en análisis de correlación no identificamos a una variable dependiente ni una variable independiente, como si se debe hacer en el análisis de regresión, sino que consideraremos a los dos variables a estudiar como x1 y x2.

En cambio, el análisis de regresión lineal estudia la relación que existe entre dos o más variables cuantitativas y es capaz de identificar un modelo (o función) que relaciona las variables. Por lo tanto, el objetivo en un análisis de regresión lineal es predecir la media de una variable dependiente en función de una o más variables independientes (Regresión lineal simple contempla una variable dependiente y una variable independiente, Regresión lineal múltiple contempla una variable dependiente y dos o más variables independientes).

Veamos el siguiente problema:

Se realizó un estudio con 21 variedades de ajíes, en donde se midieron características de: altura de planta, materia seca, sólidos solubles, número de semillas por fruto, diámetro ecuatorial del fruto, longitud del fruto, peso del fruto y número de frutos por planta (Archivo Ajies)

CORRELACIÓN LINEAL

Variedad	Altura de planta	Materia Seca	Solidos solubles	N semillas	Diametro ecuatorial del fruto	Longitud fruto	Peso fruto	N frutos por planta
1	50,60	18,54	11,84	152,43	3,17	11,33	25,85	4,17
2	51,60	12,90	9,99	152,17	2,98	17,73	43,43	8,30
3	43,80	12,11	9,14	220,00	3,92	15,22	67,06	4,70
4	37,00	13,59	9,72	169,14	3,70	15,49	50,33	4,27
5	47,40	12,04	9,12	243,83	4,09	16,61	73,26	6,00
6	41,00	11,15	8,07	211,40	4,34	15,89	71,12	4,40
7	45,20	14,56	10,31	169,80	2,79	14,84	34,58	7,85
8	54,17	14,35	9,90	212,50	2,96	15,07	36,39	10,30
9	46,25	13,96	9,41	51,00	1,45	6,50	7,10	22,00
10	26,30	10,01	7,98	64,50	2,35	7,20	21,28	9,79
11	13,50	14,79	9,45	140,43	2,48	10,59	18,74	3,92
12	39,70	14,10	9,80	117,40	2,76	12,13	22,28	3,70
13	37,40	13,66	9,86	161,00	3,10	13,86	39,37	4,58
14	39,90	12,21	10,39	115,33	2,61	12,99	31,16	4,56
15	38,20	11,59	8,24	153,50	3,63	16,86	60,88	3,91
16	35,00	11,08	7,79	164,17	4,18	5,20	29,26	11,42
17	29,60	11,72	8,75	183,86	3,59	7,10	31,27	10,67
18	41,14	13,65	9,92	110,33	2,65	13,87	31,46	4,33
19	23,40	8,25	7,52	204,83	6,21	10,18	105,78	3,08
20	24,20	8,62	6,86	210,83	7,38	9,25	155,19	4,10

Lo que iteresa en este estudio es determinar el grado de asociación lineal que hay entre las variables evaluadas. Cuando las variables que queremos asociar son cuantitativas el método estadístico más usado es el análisis de correlación. Cuando las variables son cualitativas (o categorizadas), el análisis de tablas de contingencia es la estrategia usual a seguir.

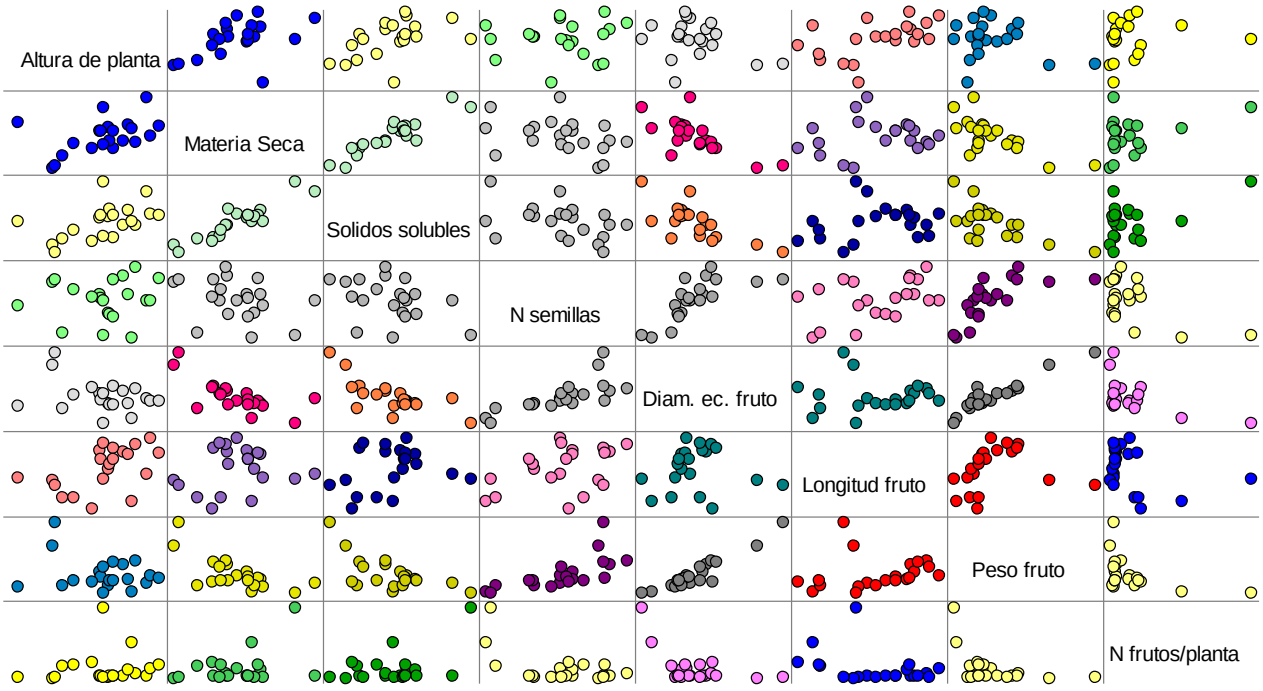
Aca el objetivo no estaba centrado en querer estimar o predecir a una variable dependiente en función de una variable independiente (cuyo objetivo corresponde a un análisis de regresión lineal).

CORRELACIÓN LINEAL

a) Construya una matriz de diagrama de dispersión para describir las correlaciones entre las variables estudiadas.

Es útil cuando se tienen muchas variables en un estudio, graficar las variables unas versus otras mediante diagramas de dispersión. Las matrices de diagramas de dispersión permiten tener una imagen simultánea de todos estos graficos en un solo paso.

Para realizar el gráfico llamado matriz de diagrama de dispersión debemos ir a Gráficos-Matriz de diagramas de dispersión, y en variables agregar a todas las variables que queremos graficar (en este caso desde altura de plantas hasta número de frutos por planta). La columna Variedad del archivo de datos no es una variable evaluada, sino que es sólo un criterio de clasificación que indica a que variedad de ajies pertenecen cada uno de los datos evaluados.



Los gráficos que observaremos en esta matriz de diagramas de dispersión son solamente los 28 graficos que se ubican por sobre la diagonal que contiene el nombre de las variables. Los graficos que se encuenran por bajo de la diagonal con nombres de variables son simplemente un “espejo” de los graficos ubicados por sobre la diagonal.

Así por ejemplo, el segundo gráfico de la primera fila, desde izquierda a derecha (en amarillo) corresponde a la asociación entre las variables Altura de planta y sólidos solubles.

El tercer grafico de la cuarta fila, desde izquierda a derecha (en morado oscuro), corresponde a la asociación entre las variables Número de semillas y peso de frutos.

Aunque los gráficos sirven para anticipar los resultados del análisis, la cuantificación de la asociación es un paso esencial y para ello debemos calcular una medida de correlación. La cuantificación de estas asociaciones las podemos observar en el punto c) de esta guía.

b) Establezca detalladamente las hipótesis que interesa contrastar en el análisis de correlación.

Las hipótesis a contrastar corresponde a:

Ho: $\rho = 0$
Ha: $\rho \neq 0$

ρ (que se lee como "rho"): corresponde al coeficiente de correlación lineal de Pearson poblacional

En las pruebas de hipótesis se usan las letras griegas, ya que en las pruebas de hipótesis nos referimos a los parámetros poblacionales y no a sus estimadores.

Como no contamos con los datos de la población completa, pero si con una pequeña muestra (21 observaciones) deberemos "estimar" ese parámetro poblacional.

El estimador de ρ (rho), se simboliza con la letra latina "r", por lo tanto:

r: corresponde al coeficiente de correlación lineal muestreal

En resumen recordar que:
Parámetros: es un valor supuesto de una población
Estimadores: son valores numéricos calculados sobre una muestra

El coeficiente de correlación lineal de Pearson es una medida de la magnitude de la asociación lineal entre dos variable cuantitativas y que no depende de las unidades de medida de las variables originales.

En caso de aceptar la Ho la conclusión sería que las variables no estan asociadas o correlacionadas linealmente entre si y en el caso de rechazar la Ho, la conclusion sería que si existe correlación lineal entre las variables.

c) Para los siguientes pares de variables:

- Materia seca y sólidos solubles
- Diámetro ecuatorial de fruto y sólidos solubles
- Número de frutos por planta y altura de plantas

Calcular el coeficiente de correlación lineal de Pearson y concluya por una hipótesis. Elabore un cuadro con los coeficientes de correlación, indicando junto al valor si éste es significativo o no lo es.

Para realizar lo anterior debemos seleccionar el menú – Análisis de correlación – Coeficiente de correlación – Seleccionar a las variables: Materia seca, sólidos solubles, diámetro ecuatorial de fruto, número de frutos por planta y altura de plantas. Luego aceptar. En la selección del coeficiente de correlación seleccionamos el coeficiente de correlación de Pearson (Nótese que la presentación de los resultados se puede solicitar en forma matricial o en forma de lista)

Como habíamos dicho anteriormente la cuantificación de las asociaciones de las variables evaluadas las podemos observar en la Tabla de Resultados (ya sea en formato de forma matricial o forma de lista)

Presentacion de Resultados en forma matricial (Salida de Infostat)

Coeficientes de correlación

Correlación de Pearson: Coeficientes|probabilidades

	Materia Seca	Solidos solubles	Diam. ec. fruto	N frutos/planta	Altura de planta
Materia Seca	1,00000000	0,00000001	0,00034665	0,09838030	0,03779118
Solidos solubles	0,91690676	1,00000000	0,00023292	0,03905144	0,02333261
Diam. ec. fruto	-0,70619215	-0,72001925	1,00000000	0,01233277	0,14978440
N frutos/planta	0,37038105	0,45327824	-0,53562649	1,00000000	0,59432546
Altura de planta	0,45590496	0,49246176	-0,32559504	0,12332445	1,00000000

En este Cuadro, en la diagonal principal, conformada por puros unos, se observan las correlaciones de cada variable con sí misma. Este coeficiente es siempre 1 y no tiene ningún valor interpretativo.

En la parte inferior de la diagonal conformada por unos, están los valores de los coeficiente de correlación lineal de Pearson (r) calculados en nuestra muestra que nos indican el grado de asociación que hay entre la variables, en donde un número negativo indica una asociación negativa (es decir que si una variable crece la otra variable disminuye), un número positivo indica una asociación positiva (es decir que si una variable crece la otra variable también crece), y valores cercanos a cero indican que no existe correlación lineal entre las variables.

El coeficiente de correlación lineal de Pearson (r) es un coeficiente cuyos valores varían entre -1 y 1 y el signo indica la dirección de la asociación.

En la parte superior de la diagonal se muestra el valor de probabilidad (p-valor) de la prueba de hipótesis realizada. Al trabajar con un nivel de significancia del 5%, se rechaza la hipótesis nula (es decir se concluye que existe correlación lineal entre ambas variables, la correlación es estadísticamente significativa) si el valor p es $\leq 0{,}05$ para las hipótesis: Ho: $\rho = 0$ vs Ha: $\rho \neq 0$

CORRELACIÓN LINEAL

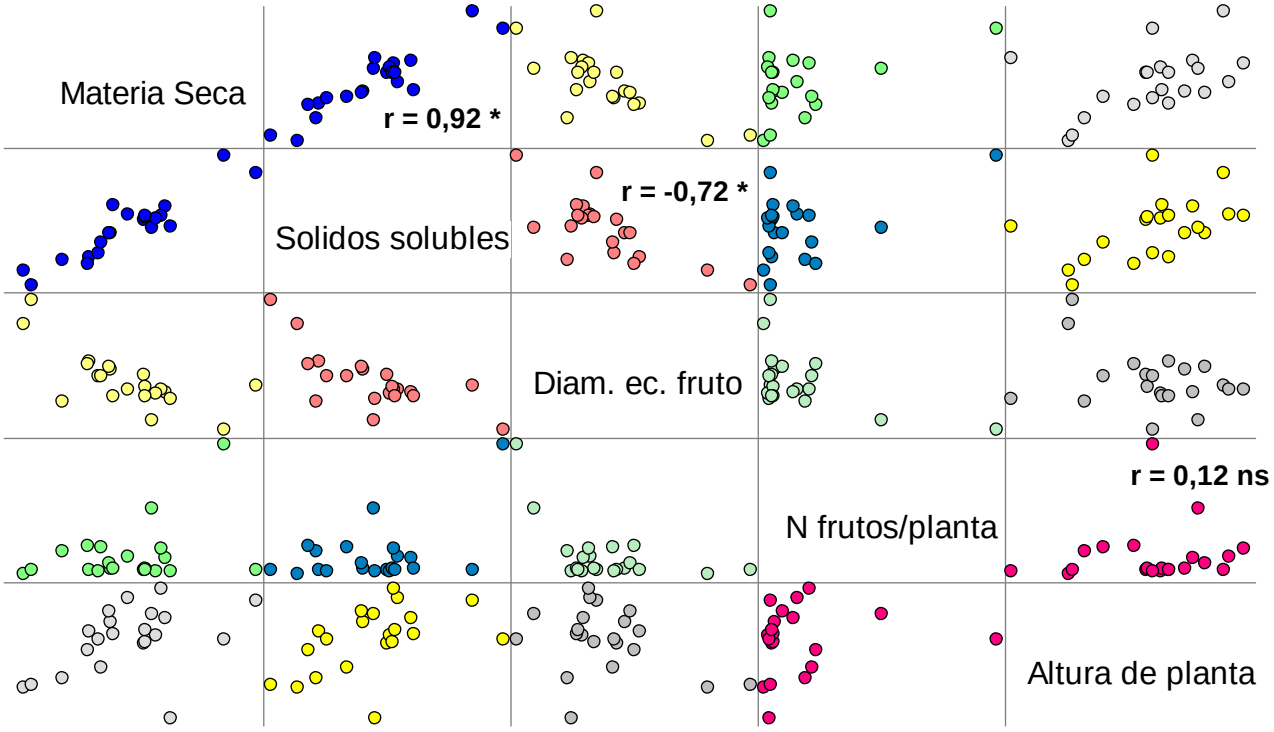
Conclusiones:

Se observa que la correlación entre las variables Materia Seca y Sólidos solubles es alta, positiva (r = 0,92) y estadísticamente significativa (p valor=0,00000001).

La correlación entre las variables Diámetro ecuatorial de fruto y sólidos solubes es alta, negativa (r= - 0,72) y estadísticamente significativa (p valor=0,00023292).

La correlación entre las variables Número de frutos por planta y Altura de planta es baja, positiva (r = 0,12) pero estadísticamente no significativa (p valor = 0,5943).

En la presentación de un Informe los resultados se podrían presentar de la siguiente manera o bien elaborando un Cuadro con los coeficientes de correlación, indicando junto al valor si éste es significativo o no lo es:



Presentacion de Resultados en forma de lista (Salida de Infostat)

Infostat da la opción de presentar los resultados en forma de lista

Coeficientes de correlación

Correlación de Pearson

Variable(1)	Variable(2)	n	Pearson	p-valor
Materia Seca	Materia Seca	21	1,00	<0,0001
Materia Seca	Solidos solubles	21	0,92	<0,0001
Materia Seca	Diam. ec. fruto	21	-0,71	0,0003
Materia Seca	N frutos/planta	21	0,37	0,0984
Materia Seca	Altura de planta	21	0,46	0,0378
Solidos solubles	Materia Seca	21	0,92	<0,0001
Solidos solubles	Solidos solubles	21	1,00	<0,0001
Solidos solubles	Diam. ec. fruto	21	-0,72	0,0002
Solidos solubles	N frutos/planta	21	0,45	0,0391
Solidos solubles	Altura de planta	21	0,49	0,0233
Diam. ec. fruto	Materia Seca	21	-0,71	0,0003
Diam. ec. fruto	Solidos solubles	21	-0,72	0,0002
Diam. ec. fruto	Diam. ec. fruto	21	1,00	<0,0001
Diam. ec. fruto	N frutos/planta	21	-0,54	0,0123
Diam. ec. fruto	Altura de planta	21	-0,33	0,1498
N frutos/planta	Materia Seca	21	0,37	0,0984
N frutos/planta	Solidos solubles	21	0,45	0,0391
N frutos/planta	Diam. ec. fruto	21	-0,54	0,0123
N frutos/planta	N frutos/planta	21	1,00	<0,0001
N frutos/planta	Altura de planta	21	0,12	0,5943
Altura de planta	Materia Seca	21	0,46	0,0378
Altura de planta	Solidos solubles	21	0,49	0,0233
Altura de planta	Diam. ec. fruto	21	-0,33	0,1498
Altura de planta	N frutos/planta	21	0,12	0,5943
Altura de planta	Altura de planta	21	1,00	<0,0001

Estos resultados son los mismos que lo que discutimos anteriormente en la presentación de resultados en forma matricial. En este Cuadro, las dos primeras columnas corresponden a las dos variables consideradas en el análisis de correlación (x1 y x2), la columna “n” corresponde a la muestra evaluada (en este caso n=21 observaciones), la columna llamada “Pearson” corresponde al valor del coeficiente de correlación lineal de Pearson (r) y la última columna corresponde al p-valor.