1 Correlación

Considera un par de variables aleatorias X e Y. Por ejemplo, X podría ser la masa de un objeto medida en kilogramos, e Y podría ser la masa del obejto medida en gramos o X podría ser la estatura de una persona, e Y podría ser el tamaño de su zapato, o X podría ser el la colocación alfabética de la segunda letra en el apellido de una persona ($A = 1, B = 2$, etc.), e Y podría ser su nivel de colesterol.

Uno de los principales problemas que abordar el concepto de correlación junto con la regresión es el grado en que el conocimiento de X ayuda a predecir Y (o viceversa), o manera equivalente cual es grado en que dos variables están correlacionadas.

1.1 Perfecta correlación

Un ejemplo de correlación perfecta es la masa de un objeto expresada en kilogramos o gramos. Si conocemos la masa X en kilogramos, entonces también conocemos la masa Y en gramos. Simplemente necesitamos multiplicar por 1000. Es decir, Y = 1000X. Un kilogramo equivale a 1000 gramos, 2.73 kilogramos equivale a 2730 gramos, etc. El conocimiento de la masa en kilogramos nos permite establecer exactamente cuál es la masa en gramos. Lo contrario también es cierto, por supuesto. El conocimiento de la masa en gramos nos permite establecer exactamente cuál es la masa en kilogramos. Simplemente dividimos por 1000.)

Si tomamos un grupo de objetos y determinamos sus masas en kilogramos y gramos, y luego dibujamos los resultados, obtendremos algo cque se muestra en la siguiente figura. (Para el propósito actual, asumiremos que cualquier error de medición es insignificante). Todos los puntos se encuentran en una línea recta. Esto es consecuencia de la perfecta correlación.

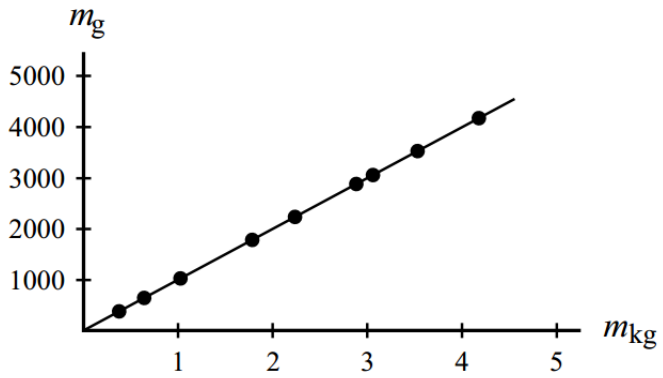

Figure 1: La masa en gramos está perfectamente correlacionada con la masa en kilogramos.

1.2 Alguna correlación

Un ejemplo de correlación no nula pero imperfecta es el segundo ejemplo mencionado anteriormente, que involucra estatura y tamaño de zapatos. (Los tamaños de zapatos para hombres y mujeres usan diferentes escalas, así que veamos los tamaños para hombres aquí. Además, algunos tamaños de fabricantes son grandes o pequeños, pero ignoramos ese problema). Ciertamente no esperamos una correlación perfecta entre la estatura y el tamaño zapato, porque eso significaría que podríamos predecir exactamente el tamaño del zapato de una persona según la estatura (o viceversa).

Esto no es posible, por supuesto, porque todas las personas que miden seis pies de estatura ciertamente no tienen el mismo tamaño de zapato. Además, no puede haber una correlación perfecta porque los

tamaños de los zapatos usan una escala discreta, mientras que las estaturas son continuas.

¿Pero hay al menos alguna correlación? Es decir, ¿el conocimiento de la estatura de una persona nos permite adivinar mejor el tamaño de su zapato, en comparación con nuestra suposición si no tuviéramos conocimiento de la estatura?.

En la siguiente figura se muestra un diagrama de dispersión de algunos datos. (Se hizo una muestra de ciertos estudiantes anotando su estatura y el tamaño de sus zapatos. La estatura se mide en pulgadas. Como el tamaño de los hombres y mujeres usan escalas diferentes, solo se usó los datos de 26 estudiantes varones).

De los datos, el tamaño promedio de un zapato de las 26 personas es de 10.4, mientras que el tamaño promedio de zapato de una persona 6 pies es de 11.4. Por lo tanto, si se desea hacer una estimación del tamaño de un zapato de una persona de 6 pies, se podriá hacerlo mejor si adivina el valor de 11.4 en lugar de 10.4.

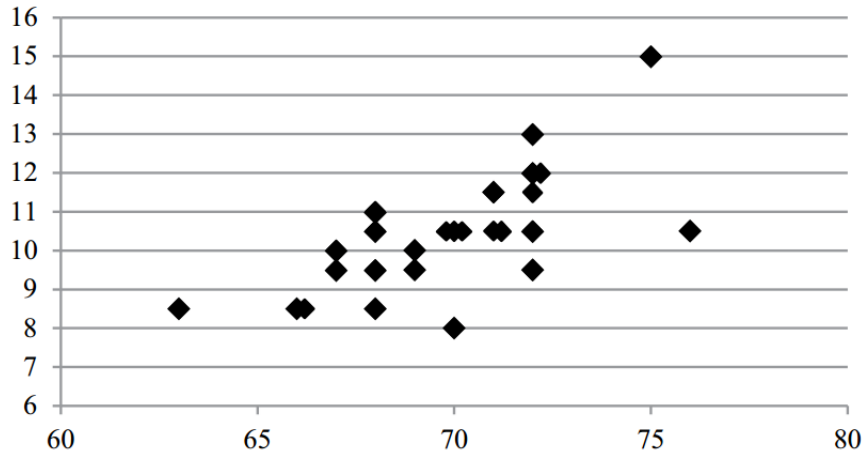

Figure 2: Un diagrama de dispersión del tamaño de un zapato frente a la altura (en pulgadas)

1.3 Correlación cero

Un ejemplo de correlación cero es el tercer ejemplo mencionado anteriormente, que incluye la colocación alfabética ($A = 1$, $B = 2$, etc.) de la segunda letra del apellido, junto con el nivel de colesterol. Es altamente dudoso que haya mucha correlación aquí.

¿Qué la segunda letra de un apellido, ayudaría a predecir el nivel de colesterol?. En el mejor de los casos, ciertos nombres (Murphy, Smith, Li, etc.) son comunes en ciertas etnias y es indudable que diferentes etnias tienen niveles de colesterol ligeramente diferentes (en promedio) debido a genes y dietas diferentes. Pero en general este efecto es pequeõ y se elimina por otros efectos. Entonces, por el bien del argumento, asumiremos que no hay correlación aquí, cuando se considera esta pregunta.

Sin embargo, este ejemplo debería convencerte de que pequeñas correlaciones (o quizás incluso grandes) pueden aparecer en situaciones en las que a primera vista es difícil imaginar una correlación.

Los primeros dos de los ejemplos anteriores implican una correlación positiva; un aumento en X corresponde a un aumento en Y (en promedio). La línea (o puntos en general) de puntos en el diagrama de dispersión tiene una pendiente ascendente. También es posible tener una correlación negativa, donde un aumento en X corresponde a una disminución en Y (en promedio).

La línea (o puntos en general) de puntos en el diagrama de dispersión tendrá una pendiente descendente. Un ejemplo de correlación negativa es la ingesta de vitamina C y la incidencia de escorbuto. Cuanta más vitamina C tomes, menos probabilidades tendraś de tener escorbuto.

Ten en cuenta que la correlación no implica necesariamente la causalidad. En el caso de la vitamina C y el escorbuto, resulta que hay causalidad; tomar más vitamina C ayuda a evitar que tengas escorbuto. Pero en el caso de la estatura y el tamaño del zapato, no es que ser alto hace que tus pies sean más grandes,

más que tener pies grandes te hace ser más alto. (La situación es simétrica, así que si quiere argumentar la causalidad, será difícil decir cuál está causando cuál). En cambio, lo que está sucediendo es que hay una tercera cosa, a saber, la genética (y la dieta también), que hace que tanto la estatura como el tamaño de los pies sean más grandes o más pequeños (en promedio).

Otro ejemplo en este sentido consiste en la cantidad de veces que las personas en un pueblo determinado en un día determinado se ponen las gafas de sol, junto con la cantidad de veces que se aplican protector solar. Hay una correlación positiva entre estas dos cosas, pero ninguna de ellas causa la otra. En su lugar, ambos son causados por una tercera cosa: el sol!.

Trataremos solo la correlación lineal, aunque ciertamente hay ejemplos de correlación no lineal. Un ejemplo simple es la relación entre el área de un cuadrado y la longitud de su lado: área $=$ (longitud del lado)^2. Esta relación es cuadrática, no lineal.

Otro ejemplo es la relación entre el ingreso laboral y la edad de una persona. Los niños de tres años no ganan mucho trabajando en un trabajo y tampoco lo hacen los de 100 años (generalmente). Por lo tanto, la gráfica del ingreso promedio en función de la edad debe comenzar en cero, luego aumentar a un máximo y luego volver a cero.

2 Un modelo para la correlación

Ahora tratemos de entender la manera qen que se pueden correlacionar dos variables aleatorias. Esta comprensión nos llevará al coeficiente de correlación ρ. Asumiremos en la presente discusión que las dos variables aleatorias tienen distribuciones normales.

Este supuesto no es necesario, nuestros resultados matemáticos serán válidos para cualquier distribución. De hecho, cuando se trata de datos reales del mundo real, a menudo ocurre que una o ambas variables no están normalmente distribuidas. Sin embargo, debido al teorema del límite central, muchas variables aleatorias de la vida real se distribuyen de manera aproximadamente normal.

Consideramos una variable aleatoria distribuida normalmente con media cero y desviación estándar σ_X:

$$X : \mu = 0 \quad \sigma = \sigma_X \tag{1}$$

Hemos elegido la media igual a cero solo para que nuestros cálculos sean más sencillos. Todos los resultados se pueden generalizar para cualquier media.

Consideremos otra variable aleatoria Y que está correlacionada (hasta cierto punto) con X. Con esto queremos decir que Y está parcialmente determinada (de manera lineal) por X y parcialmente determinada por otra variable aleatoria Z (que se supone que está distribuida normalmente) que es independiente de X. Z puede a su vez ser la suma de muchas otras variables aleatorias, todas independientes de X. Estamos agrupando el efecto de todas estas variables en una variable Z. Podemos ser cuantitativos acerca de la dependencia de Y en X y Z escribiendo Y como,

$$y = mX + Z \tag{2}$$

donde m es un factor numérico. Para mantener las cosas simples, asumiremos que la media de Z también es cero. Entonces, si la desviación estándar de Z es σ_z, tenemos

$$Z : \mu = 0 \quad \sigma = \sigma_Z \tag{3}$$

Debes tener en cuenta que si tomamos la media de la ecuación (2), vemos que los diversos medias están relacionados por,

$$\mu_Y = m\mu_X + \mu_Z \tag{4}$$

Ya que estamos asumiendo $\mu_X = \mu_Z = 0$ aquí, implica que μ_y también es igual a cero. En la ecuación (2), estamos produciendo Y a partir de dos distribuciones conocidas (e independientes) X y Z. Para ser explícitos, el significado de la ecuación(2) es el siguiente.

Elije un valor de x de la variable aleatoria X y multiplica el resultado por m para obtener mx. Luego elije un valor z de la variable aleatoria Z y agréguelo a mx para obtener $y = mx + z$. Este es el valor pedido de y. Podemos etiquetar este par ordenado de valores (X, Y) como (x_1, y_1). Luego repetimos el proceso con nuevos valores de X y Z para obtener un segundo par (X, Y), (x_2, y_2). Y así sucesivamente, para tantos pares como queramos.

Como ejemplo, Y podría ser el peso medido de un objeto, X podría ser el verdadero peso y Z podría ser el error introducido por el proceso de medición (lectura de la escala, comportamiento de la escala en función de una ubicación ligeramente ladeada del objeto, etc.). Es posible que estas variables no tengan distribuciones normales, pero, nuevamente, esa suposición no es crítica en nuestra discusión. En este ejemplo, $m = 1$.

Debemos mencionar que aunque la ecuación (2) es el punto de partida para obtener la mayoría de los resultados de correlación, pero rara vez es el punto de partida en la práctica. Es decir, rara vez se te dan las distribuciones X y Z subyacentes. En su lugar, siempre se le dan algunos datos y necesitas calcular el coeficiente de correlación ρ.

Para ver qué tipo de correlación (2) produce entre X e Y, consideremos dos casos especiales, para obtener una idea general de los efectos de m y Z.

2.1 Correlación perfecta ($\sigma_Z = 0$)

Si la desviación estándar de Z es $\sigma_z = 0$, entonces Z siempre toma el valor $z = 0$, porque suponemos que la media de Z es cero. Así que la ecuación (2) se reduce a $Y = mX$. Es decir, Y es un número fijo m veces X; todos los valores de x e y están relacionados por $y = mx$.

Esto significa que todos los puntos (x, y) en un diagrama de dispersión se encuentran en una línea recta $y = mx$, como se muestra en la siguiente figura de puntos aleatorios generados numéricamente a partir de una distribución normal X.

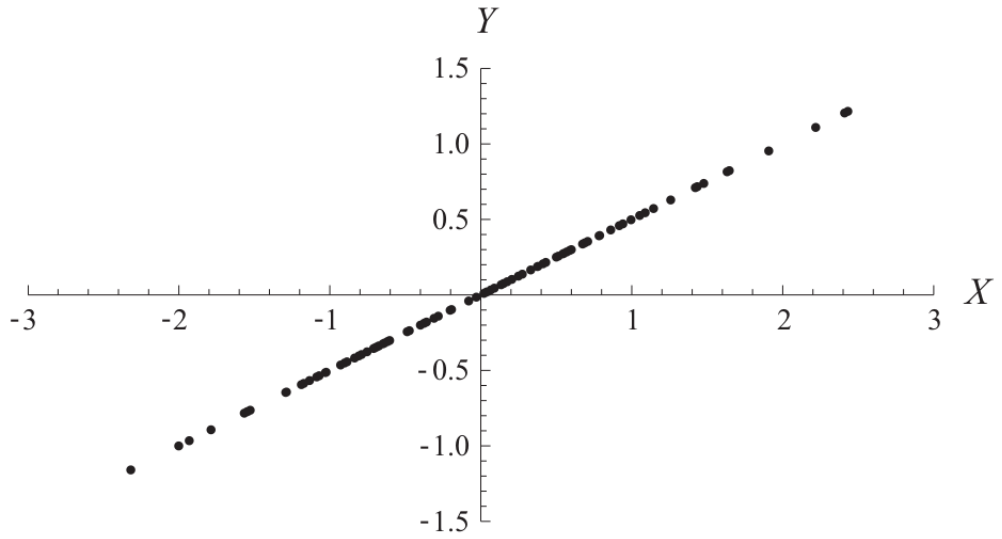

Figure 3: Perfecta correlación

Hemos elegido arbitrariamente $m = 0.5$ y $\sigma_X = 1$. En el presente caso de una línea recta, decimos que X e Y están perfectamente correlacionados (o completamente). El valor de Y está completamente determinado por el valor de X. No hay una variable aleatoria adicional Z para desordenar esta determinación completa.

En el caso donde σ_Z es pequeño pero no nulo, obtenemos una correlación fuerte pero no perfecta. La siguiente figura muestra una gráfica de puntos en el caso donde σ_Z es igual a $(0.1)\sigma_X$.

Hemos elegido de nuevo $m = 0.5$ y $\sigma_X = 1$ (y por lo tanto, $\sigma_z = 0.1$). Hemos generado valores aleatorios de cada una de las distribuciones normales X y Z y luego formando $Y = mX + Z$. En el presente caso de σ_z pequeõ, el conocimiento de X es muy útil para predecir Y, aunque no predice Y exactamente.

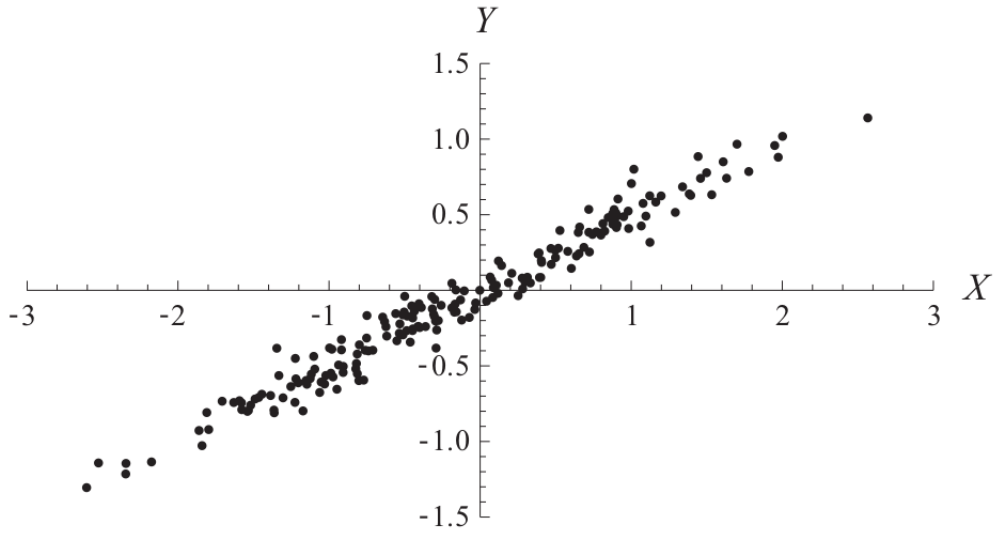

Figure 4: Fuerte correlación

2.2 Correlación cero ($m = 0$)

Si $m = 0$, entonces la ecuación (2) se reduce a $Y = Z$. Y dado que Z es independiente de X, esto significa que Y también es independiente de X. La siguiente figura muestra una gráfica de puntos en el caso donde $m = 0$. Hemos elegido arbitrariamente $\sigma_X = 2$ y $\sigma_Z = 1$. Hemos generado los puntos seleccionando valores aleatorios de cada una de las distribuciones normales X y Z y luego estableciendo Y igual a Z.

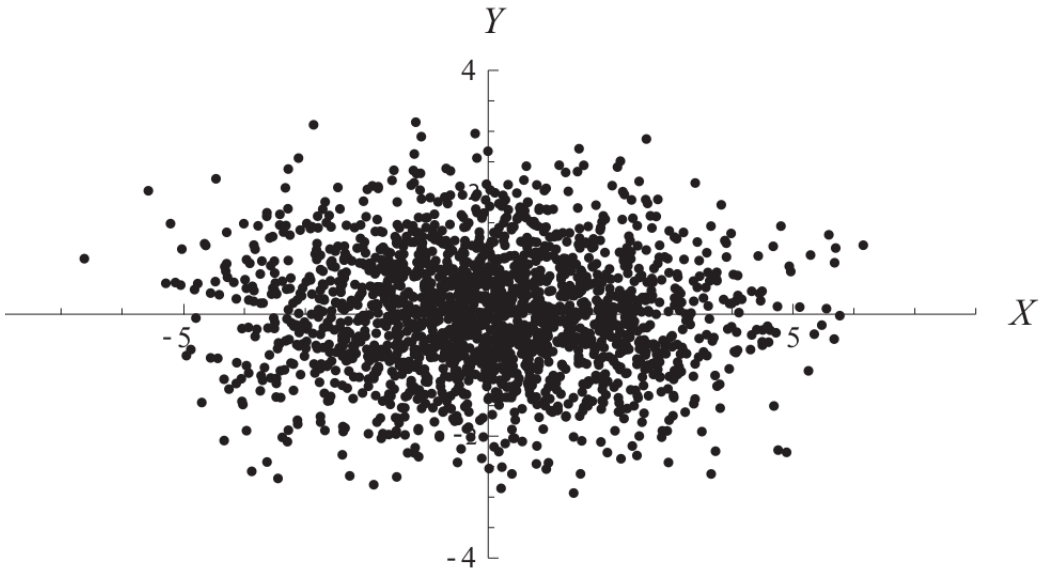

Figure 5: Correlación cero

De la figura se observa que X e Y no están correlacionados completamente. La distribución para Y es

independiente del valor de X. Es decir, para cualquier valor dado de X, los valores de Y se distribuyen normalmente alrededor de $Y = 0$, con la misma desviación estándar (que es igual a σ_Z). En otras palabras, la probabilidad (o más bien, la densidad de probabilidad) de obtener un cierto valor de Y, dado un valor particular de X, es independiente del valor de X. Esta probabilidad viene dada por la distribución normal para Z, ya que $Y = Z$ en el presente caso donde $m = 0$.

Si imaginamos dibujar franjas sombreadas verticales en dos valores diferentes de X, como se muestra en la siguiente figura, entonces las distribuciones de los valores Y en estas dos franjas son las siguientes. Igual, excepto por un factor de escala global. Este factor de escala es simplemente la probabilidad (o más bien, la densidad de probabilidad) de obtener cada uno de los valores dados de X.

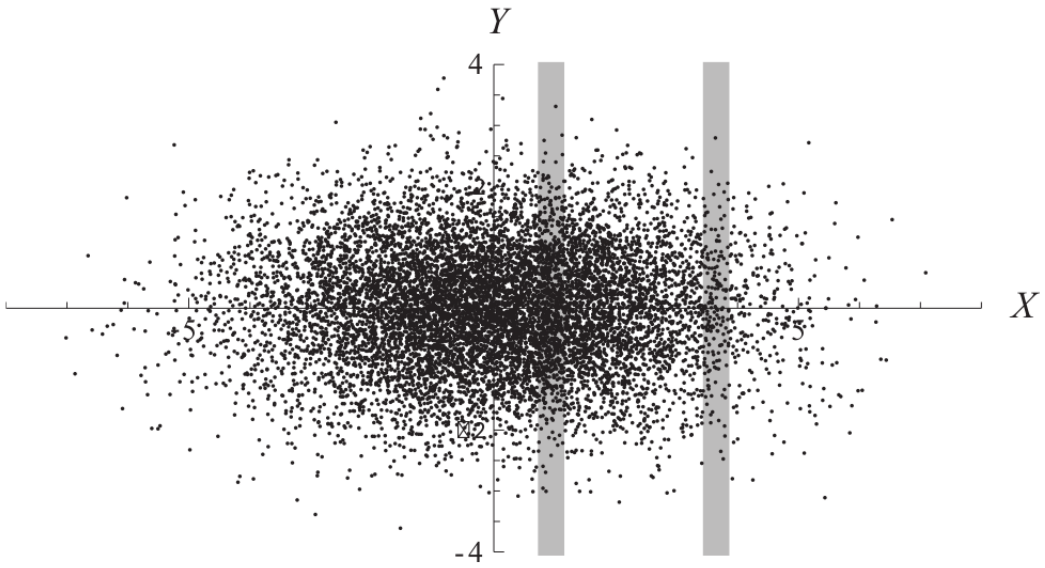

Figure 6: Si $m = 0$, la distribución de los valores de Y dentro de una franja vertical es independiente (aparte de un factor de escala global) de la ubicación de la franja.

Valores más grandes de $|X|$ es menos probable, debido al factor$e^{-x^2/2\sigma_X^2}$ en la distribución normal. Así que hay menos puntos en la franja derecha. Pero dado un valor de X, la distribución de probabilidad para Y (en este caso $m = 0$) es simplemente la distribución de probabilidad para Z, que es independiente de X.

En el caso de que m sea pequeño pero distinto de cero, obtenemos una correlación débil. La siguiente figura muestra una gráfica de puntos en el caso en que $m = 0.2$ y nuevamente con $\sigma_X = 2$ y $\sigma_Z = 1$. En este caso, el conocimiento de X ayuda un poco a predecir el valor de Y. No ayuda mucho en la región cercana al origen, el gráfico no muestra mucha inclinación (parece básicamente la misma que figura de cero correlación cerca del origen). Pero para valores más grandes de X, hay un sesgo claro en los valores de Y. Más puntos se encuentran sobre el eje X en el lado derecho de la gráfica y más puntos se encuentran debajo del eje X en el lado izquierdo.

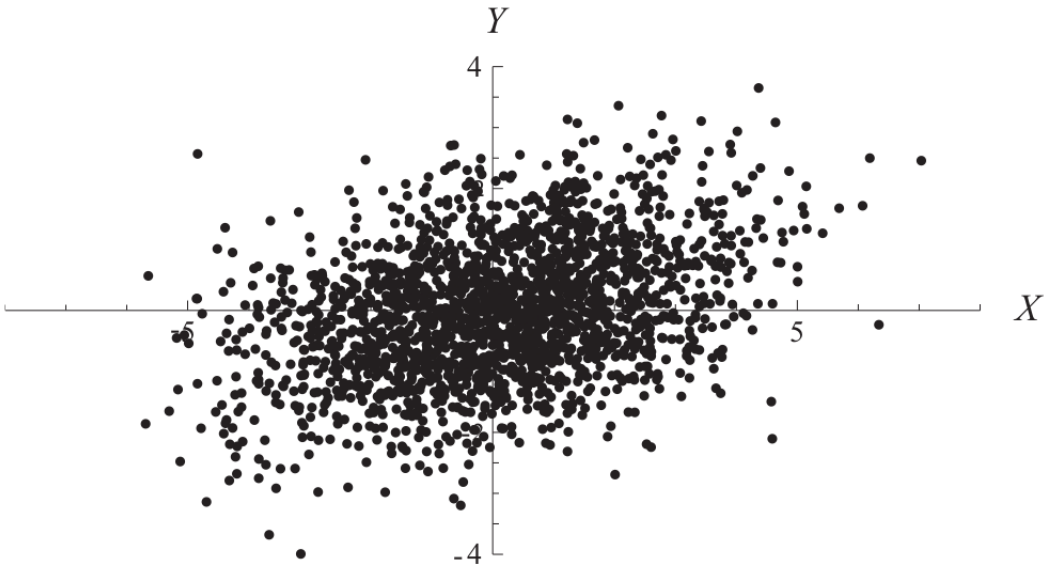

Figure 7: Correlación débil

2.3 Anotaciones

1. Todos los diagramas de dispersión anteriores se centran en el origen porque asumimos que las medias de X y Z son cero, lo que implica que la media de Y es cero. Si Z en su lugar tuviera un valor medio distinto de cero μ_Z, entonces la mancha de puntos se desplazaría hacia arriba en μ_Z. Si X tuviera una media μ_X distinta de cero, entonces la mancha de puntos se desplazaría hacia la derecha en μ_X y también por $m\mu_X$.

2. En las discusiones anteriores, tratamos a X como la variable independiente y a Y como la variable dependiente, y analizamos en qué medida X determinó Y. Sin embargo, si alguien te da uno de los diagramas de dispersión anteriores, podrías razonablemente inclinar tu cabeza hacia un lado y considerar que X es una 'función' de Y y luego observar en qué medida Y determina X.

3. Notamos en la `Figure 6` que la distribución relativa de los valores de Y dentro de una franja vertical es independiente de la ubicación de la franja. Este hecho se mantiene no solo en la `Figure 6` donde hay una correlación cero, sino también (en un sentido ligeramente modificado) en el caso de una correlación distinta de cero, incluso cuando hay una correlación fuerte como en la `Figure 4`. Aunque podría parecer que la dispersión (la desviación estándar) de los valores de Y se reduce en las colas de la gráfica en la `Figure 4`, la dispersión es de hecho la misma para todos los valores de X. La expresión $Y = mX + Z$ dice que para cualquier valor dado de X, los valores de Y se centran en mX (en lugar de cero; esta es la ligera modificación mencionada anteriormente) y tienen la misma desviación estándar de σ_Z alrededor de este valor. La propagación parece ser mayor en el medio del gráfico, pero solo porque hay más puntos allí.

3 Coeficiente de correlación

Ahora mostraremos cómo producir el coeficiente de correlación ρ a partir de cantidades asociadas con el lado derecho de la ecuación (2), a saber m, σ_x, y σ_z. Para hacer esto, necesitaremos determinar la desviación estándar de $Y = mX + Z$. Sabemos que mX tiene una desviación estándar $m\sigma_X$ y Z tiene una desviación estándar de σ_Z. Y sabemos que la desviación estándar de la suma de dos variables independientes (como X y Z son) se obtiene sumando las dos desviaciones estándar en cuadratura. (Las variables no necesitan ser normales para que esto sea cierto). Por lo tanto, Y se escribe mediante [1]:

[1] En estas notas se usará r o ρ para indicar el coeficiente de correlación

$$Y : \mu = 0, \quad \sigma_Y = \sqrt{m^2\sigma_X^2 + \sigma_Z^2}. \tag{5}$$

El valor $\mu = 0$ sigue de la ecuación (4), ya que asumimos que $\mu_X = \mu_Z = 0$. Revisemos algunos casos limitantes de la ecuación (5). En un extremo donde $\sigma_z = 0$ (correlación completa entre X e Y), tenemos $\sigma_Y = m\sigma_X$.

Para valores generales de m y σ_z, definimos el coeficiente de correlación ρ como la fracción de σ_T que se puede atribuir a X (suponiendo un modelo lineal). Dado que la parte de σ_Y que se puede atribuir a X es $m\sigma_X$, esta fracción es,

$$r = \rho = \frac{m\sigma_X}{\sigma_Y} = \frac{m\sigma_X}{\sqrt{m^2\sigma_X^2 + \sigma_Z^2}} \tag{6}$$

De manera equivalente, ρ^2 es igual a la fracción de la varianza de Y que se puede atribuir a X. El uso de la expresión para ρ en la ecuación anterior requiere conocimiento de m, junto con σ_x y σ_Y o σ_Z. Si tenemos m, las distribuciones X y Z subyacentes que conforman Y, entonces podemos usar la ecuación anterior para encontrar ρ. Pero como se mencionó anteriormente, generalmente solo nos dan una colección de puntos de datos en el plano $x - y$, sin que nos den las distribuciones de X y Z exactas. ¿Como encontramos ρ en ese caso?

4 Covarianza

Para encontrar ρ si se nos da una colección de puntos de datos, necesitamos definir la covarianza de dos variables aleatorias.

La `covarianza` de las variables aleatorias X y Y es la cantidad denotada por $cov(X, Y)$ y es dada por:

$$cov(X, Y) = \mathbb{E}([X - \mathbb{E}(X)][Y - \mathbb{E}(Y)]) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{7}$$

siempre que las esperanzas existan.

Podemos agregar algunas propiedades más, de la varianza y covarianza:

- $\text{Var}(X + Y) = \text{Var}(X) + 2cov(X, Y) + \text{Var}(Y)$
- Si X y Y son independientes : $cov(X, Y) = 0$ y $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.

La covarianza se utiliza a menudo como una medida de la dependencia de X e Y y la razón de esto es que $cov(X, Y)$ es un sólo un número (en lugar de un objeto complicado como la función densidad conjunta) que contiene información útil sobre el comportamiento conjunto de X e Y.

Por ejemplo, si $cov(X, Y) > 0$, entonces $X - \mathbb{E}(X)$ y $Y - \mathbb{E}(Y)$ pueden tener una buena chance (en algún sentido) de tener el mismo signo.

La principal desventaja de la covarianza como medida de dependencia de X e Y es que no es invariante en la escala: Si X y Y estaán medidas en pulgadas y U y V tienen las misma medida en centímetros (tal que $U = \alpha X$ y $V = \alpha Y$, donde $\alpha \sim 2.54$), entonces $cov(U, V) \sim 6cov(X, Y)$ a pesar del hecho de que el par (X, Y) y (U, V) miden la misma cantidad.

Para tratar con esta `reescala` en la covarianza se define lo siguiente:

`El coeficiente de correlación` de las variables aleatorias X y Y es la cantidad $\rho(X, Y)$ definida por:

$$\rho(X, Y) = \frac{cov(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{cov(X, Y)}{\sigma_X\sigma_Y}. \tag{8}$$

siempre que la última cantidad exista y $\text{Var}(X)\text{Var}(Y) \neq 0$.

$\text{cov}(x, y)$ también se puede definir para un conjunto de puntos de datos. Es solo que en lugar de hablar sobre el valor esperado de XY (asumiendo que las medias son cero), hablamos sobre el valor promedio de los productos $x_i y_i$, donde se toma el promedio de todos los puntos de datos (x_i, y_i) . Si tenemos n puntos (x_i, y_i), entonces la covarianza en el caso general de medias distintas de cero es

$$\text{cov}(x, y) = \frac{1}{n} \sum (x_i - \overline{x})(y_i - \overline{y}) \quad \text{para puntos datos} \tag{9}$$

4.1 Anotaciones

1. El coeficiente de correlación ρ es independiente de las medias de X, Y y Z. Esto sigue del hecho de que ninguna de las cantidades m, σ_X, σ_Y, σ_Z, o cov(X, Y) depende de las medias. Cambiar las medias simplemente desplaza toda la mancha de puntos alrededor del plano $X - Y$.

2. La `ecuación 8` es simétrica en X e Y. Esto significa que si cambiamos el valor independiente y las variables dependientes en un diagrama de dispersión e imaginamos que X es parcialmente dependiente de Y (en lugar de que Y sea parcialmente dependiente de X), entonces el coeficiente de correlación es el mismo. Esto no es terriblemente obvio, dada la falta de simetría en la relación en la `ecuación 2`, donde Z es independiente de X, no Y.

5 Ejemplos con varios valores del coeficiente de correlación

La figura muestra ejemplos de gráficos de dispersión para seis valores diferentes de r. Todos los gráficos tienen $\sigma_X = 2$ y $\sigma_Y = 1$ y hay alrededor de 1000 puntos en cada uno. Ten en cuenta que se necesita una r considerable para obtener un diagrama de dispersión que se ve significativamente diferente del caso $r = 0$. La gráfica de $r = 0.3$ se ve aproximadamente igual. Las gráficas en esta figura le dan un sentido visual de lo que significa un r en particular, por lo que debes tenerlo en cuenta cada vez que se le te un valor de r.

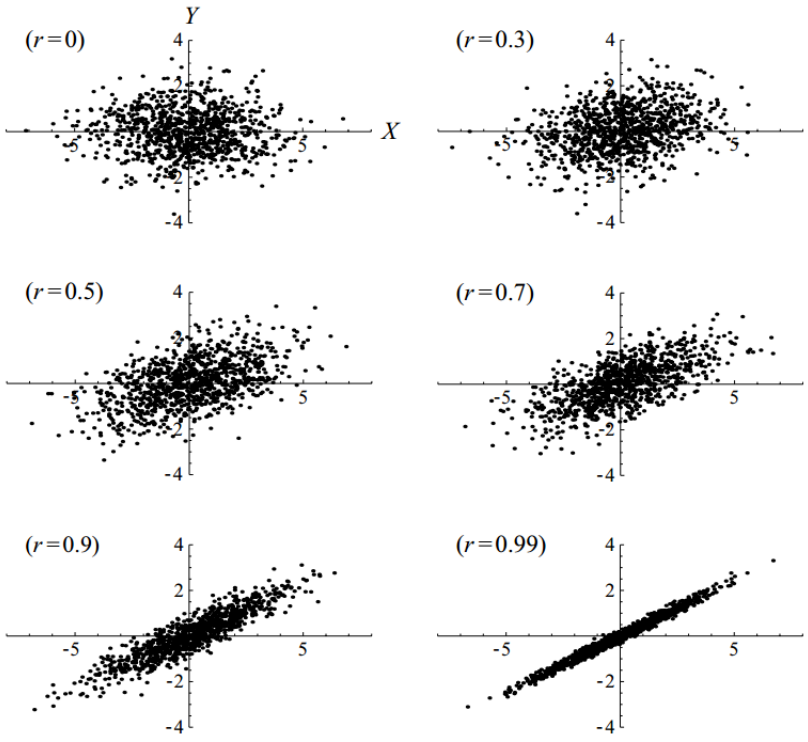

Figure 8: Gráficos de dispersión para varios valores del coeficiente de correlación r.

¿Qué se considera un valor 'bueno' o 'alto' de r? Bueno, eso depende de los datos con los que estés tratando. Si eres un científico social y encuentra una correlación de $r = 0.7$ entre una determinada característica y por ejemplo, el número de meses que una persona ha estado desempleada, ese es un resultado muy importante. Acabas de encontrar una característica que ayuda sustancialmente a predecir la duración del desempleo.

(Pero ten en cuenta que la correlación no implica necesariamente una causación. Aunque has encontrado algo que ayuda a predecir, puede que no sirva para explicar). Sin embargo, si eres un físico y encuentras una correlación de $r = 0.7$ entre la distancia d que un objeto cae (en el vacío, se cae del reposo) y el cuadrado del tiempo de caída t, entonces ese es un resultado terrible. Algo ha salido gravemente mal, porque los puntos de datos deben (al menos hasta pequeños errores experimentales) estar en la línea recta dada por $d = (g/2)t^2$, donde g es la aceleración debida a la gravedad.

Todos gráficos en la `figure 8` tienen valores positivos de r. Las gráficas para valores negativos se ven iguales, excepto que las manchas de puntos tienen pendiente hacia abajo. Por ejemplo, en la siguiente figura se muestra un diagrama de dispersión con $r = -0.7$. Como r es negativo implica que m lo es también, entonces la `ecuación 2` nos dice que un aumento en X produce una disminución en Y (en promedio). De ahí la pendiente negativa.

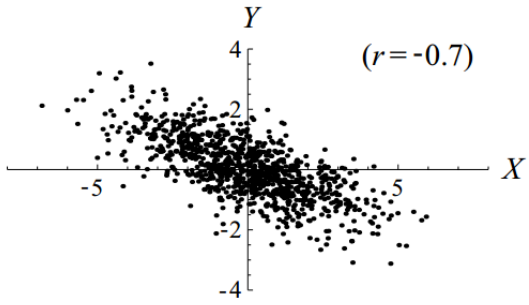

Figure 9: Gráfico de dispersión con correlación negativa.

En las figuras 3 a la 7, los tres parámetros especificados que se usaron para generar numéricamente los gráficos fueron:

$$\sigma_X, \quad \sigma_Z, \quad m, \tag{10}$$

mientras que las figuras 8 y 9, los parámetros especificados fueron:

$$\sigma_X, \quad \sigma_Y, \quad r. \tag{11}$$

Ambos conjuntos de parámetros contienen la misma información, expresada de diferentes maneras (aunque ambos conjuntos contienen σ_X). Es fácil ir de un conjunto a otro. Dado el conjunto en la ecuación (10), los valores σ_Y y r en la ecuación (11) se puede encontrar a través de la ecuación (6):

$$\sigma_Y = \sqrt{m^2\sigma_X^2 + \sigma_Z^2} \quad \text{y} \quad r = \frac{m\sigma_X}{\sqrt{m^2\sigma_X^2 + \sigma_Z^2}} \tag{12}$$

Por ejemplo, la `figure 7` se generó a partir de los parámetros $m = 0.2$, $\sigma_X = 2$, y $\sigma_Z = 1$. Así que puedes mostrar rápidamente que la ecuación (12) da $\sigma_y = 1.08$ y $r = 0.37$.

Para ir al otro lado, la expresión anterior para σ_Y puede reescribirse como $\sigma_Z^2 = \sigma_Y^2 - m^2\sigma_X^2$. Pero la ecuación (6) nos dice que $m^2\sigma_X^2 = r^2\sigma_Y^2$, entonces obtenemos $\sigma_Z^2 = (1 - r^2)\sigma_Y^2$. Por lo tanto, dado

el conjunto de parámetros en la ecuación (11), los valores de σ_Z y m en la ecuación (10) se puede calcular a través de,

$$\sigma_Z = \sigma_Y\sqrt{1-r^2} \quad \text{y} \quad m = \frac{r\sigma_Y}{\sigma_X} \tag{13}$$

Por ejemplo $r = 0.3$ en la `figure 8` fue generado desde los parámetros $r = 0.3$, $\sigma_X = 2$ y $\sigma_Y = 1$. Así que la ecuación (13), produce $\sigma_Z = 0.95$ y $m = 0.15$.

Generalmente se describen los diagramas de dispersión en términos de r (y σ_X y σ_Y) en lugar de m (y σ_X y σ_Z). Pero siempre se puede alternar entre r y m usando las ecuaciones (12) y (13). Sin embargo, de ninguna manera estamos terminados con m. Esta cantidad es extremadamente importante, ya que es la pendiente de la **línea de regresión**.