# CBS
## Colegio Bautista Shalom

# Estadística III
# Sexto PCOC
# Tercer Bimestre

**Contenidos**

**ESTADÍSTICA DESCRIPTIVA**

- ✓ MEDIANA DE LAS DESVIACIONES ABSOLUTAS (MAD).
- ✓ DISTANCIA O RANGO INTERCUARTIL.
- ✓ GRÁFICO DE CAJA (BOX-PLOT).
- ✓ RELACIONES ENTRE VARIABLES NUMÉRICAS.
- ✓ GRÁFICO DE DISPERSIÓN (SCATTER PLOT).
- ✓ COEFICIENTE DE CORRELACIÓN.
- ✓ COEFICIENTE DE CORRELACIÓN DE PEARSON.
- ✓ COEFICIENTE DE CORRELACIÓN DE SPEARMAN.
- ✓ GRÁFICOS ENGAÑOSOS.

**NOTA:** conforme avances en tu aprendizaje tu catedrático(a) te indicará la actividad o ejercicio a realizar. Sigue sus instrucciones.

**MEDIANA DE LAS DESVIACIONES ABSOLUTAS (MAD)**

La MAD (median absolute deviations) es otra medida de dispersión que pretende dar una idea resumen de "distancias a un punto central" tal como ocurre con el desvío estándar. Pero ¿en qué difiere del desvío estándar?

- Considera la mediana como punto central de la distribución para calcular las desviaciones.
- Toma el valor absoluto de las desviaciones para eliminar el signo (en vez de elevar al cuadrado como hacemos al calcular el desvío estándar).
- Toma la mediana de las distancias (en vez de promediar como hacemos con s).

Definimos la *MAD* de una muestra $X_1, X_2, \ldots, X_n$ como

$$MAD = mediana\,(|X_i - \tilde{X}|)$$

¿Cómo calculamos la MAD?

1. Ordenamos los datos de menor a mayor.
2. Calculamos la mediana.
3. Calculamos la distancia de cada dato a la mediana.
4. Despreciamos el signo de las distancias y las ordenamos de menor a mayor. 5. Buscamos la mediana de las distancias sin signo.

Propiedades de la MAD - Si la distribución es acampanada y simétrica la MAD y el desvío estándar s se relacionan del siguiente modo: $s \cong 1.48\,\text{MAD}$

La MAD es una medida de dispersión muy robusta a la presencia de datos outliers.

Ejemplo Consideremos los siguientes datos ordenados (n = 13).

| *Posición* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Datos* | 104 | 112 | 134 | 146 | 155 | 168 | 170 | 195 | 246 | 302 | 338 | 412 | 678 |

1. Como $n = 13$ la mediana es el dato que ocupa la posición $(13+1)/2 = 7 \Rightarrow \tilde{X} = 170$.
2. Calculamos las diferencias a la mediana

   $-66,\ -58,\ -36,\ -24,\ -15,\ -2,\ 0,\ 25,\ 76,\ 132,\ 168,\ 242,\ 508$
3. Despreciamos el signo de las distancias y las ordenamos de menor a mayor

   $0,\ 2,\ 15,\ 24,\ 25,\ 36,\ 58,\ 66,\ 76,\ 132,\ 168,\ 242,\ 508$
4. Tenemos $n = 13$ diferencias, por lo tanto la mediana es la diferencia que ocupa el séptimo lugar, en consecuencia

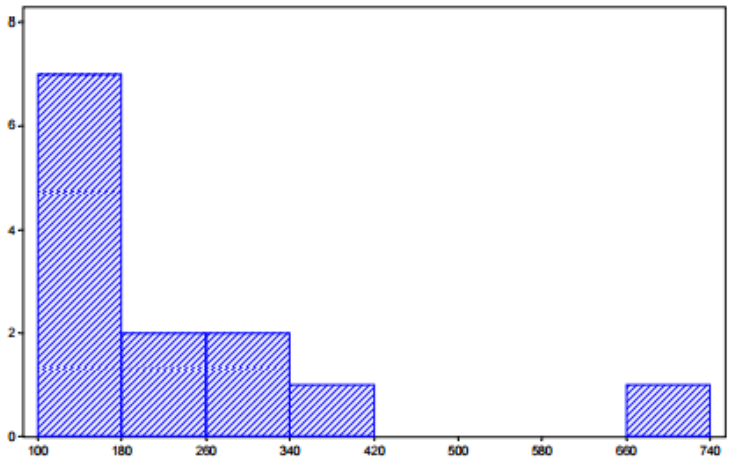
$$\text{MAD} = 58$$

Si la distribución fuera simétrica esperaríamos que el desvío estándar fuera:

$$s \cong 1.48\,\text{MAD} = 1.48 \cdot 58 = 85.8$$

pero para estos datos s = 160.48.

Esta gran diferencia nos dice que la distribución es asimétrica. El histograma de estos datos, que se presenta en la figura siguiente confirma este hecho.



**DISTANCIA O RANGO INTERCUARTIL**

El rango intercuartil o distancia intercuartil ($D_I$) de un conjunto de datos es la distancia entre los dos cuartiles:

$$D_I = C_S - C_I$$

Indica el rango donde se encuentra aproximadamente el 50% "central" de las observaciones.

Propiedades:

- Si todos los datos son iguales $D_I = 0$. Pero $D_I$ puedes ser igual a cero aún cuando no todos los datos sean iguales.

  *Ejemplo* 5 12 12 12 12 12 20 $n = 7$ $C_I = 12$ $C_S = 12$ $D_I = 0$
- Es una medida robusta de dispersión.
- Cuando la distribución es simétrica y acampanada la relación entre la distancia intercuartil y el desvío estándar es la siguiente

$$D_I \cong \frac{4}{3}\, s$$

Para distribuciones muy asimétricas $s > D_I$

Ejemplo Consideremos nuevamente los datos siguientes.

| *Posición* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Datos* | 104 | 112 | 134 | 146 | 155 | 168 | 170 | 195 | 246 | 302 | 338 | 412 | 678 |

Posición del Cuartil Inferior = (13+1)/4 = 3.5 $\Rightarrow C_I = \frac{134+146}{2} = 140$

Posición del Cuartil Superior = 3.(13+1)/4 = 10.5 $\Rightarrow C_S = \frac{302+338}{2} = 320$

$$D_I = C_S - C_I = 320 - 140 = 80$$

Concluimos que el 50% central de los datos se encuentra en una distancia de 80 unidades.
Para estos datos s = 160.5.
Si la distribución fuera simétrica esperaríamos que $D_I \cong 0.75\ s = 0.75 \cdot 160.5 = 120$. Sin embargo, $D_I = 80$, lo que nos indica que la distribución es asimétrica.

**GRÁFICO DE CAJA (BOX-PLOT)**

Concluimos este capítulo presentando un gráfico propuesto por Tukey para presentar datos numéricos, especialmente útil para comparar distribuciones de varios conjuntos de observaciones.

Está basado en medidas robustas de posición y dispersión. ¿Cómo se construye un box-plot?

1. Ordenar los datos de menor a mayor
2. Calcular la mediana, el cuartil inferior, el cuartil superior y la distancia intercuartil.
3. Calcular cotas que nos permitirán decidir si un dato es outlier:

- $2^a$ cota inferior $= C_I - 3\ D_I$
- $1^a$ cota inferior $= C_I - 1.5\ D_I$
- $1^a$ cota superior $= C_S + 1.5\ D_I$
- $2^a$ cota superior $= C_S + 3\ D_I$

Cualquier dato que caiga entre la 1ª y 2ª cota inferior o entre la 1ª y 2ª cota superior será declarado outlier. Cualquier dato que caiga por fuera de la 2ª cota inferior o la 2ª cota superior será declarado outlier severo.

4. Dibujar una escala que cubra el rango de variación de los datos y marcar la mediana y los cuartiles. Dibujar una caja que se extienda entre los cuartiles y marcar en ella la posición de la mediana.
5. Partiendo del cuartil inferior trazar una línea (bigote) que llegue hasta el último dato contenido "dentro" de la 1ª cota inferior. Partiendo del cuartil superior trazar una línea (bigote) que llegue hasta el último dato contenido "dentro" de la 1ª cota superior.
6. Marcar la posición de los outliers con un símbolo (por ejemplo, *) y de los outliers severos con otro símbolo (por ejemplo, ○).

Ejemplo:

Consideremos nuevamente los datos siguientes.

| *Posición* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Datos* | 104 | 112 | 134 | 146 | 155 | 168 | 170 | 195 | 246 | 302 | 338 | 412 | 678 |

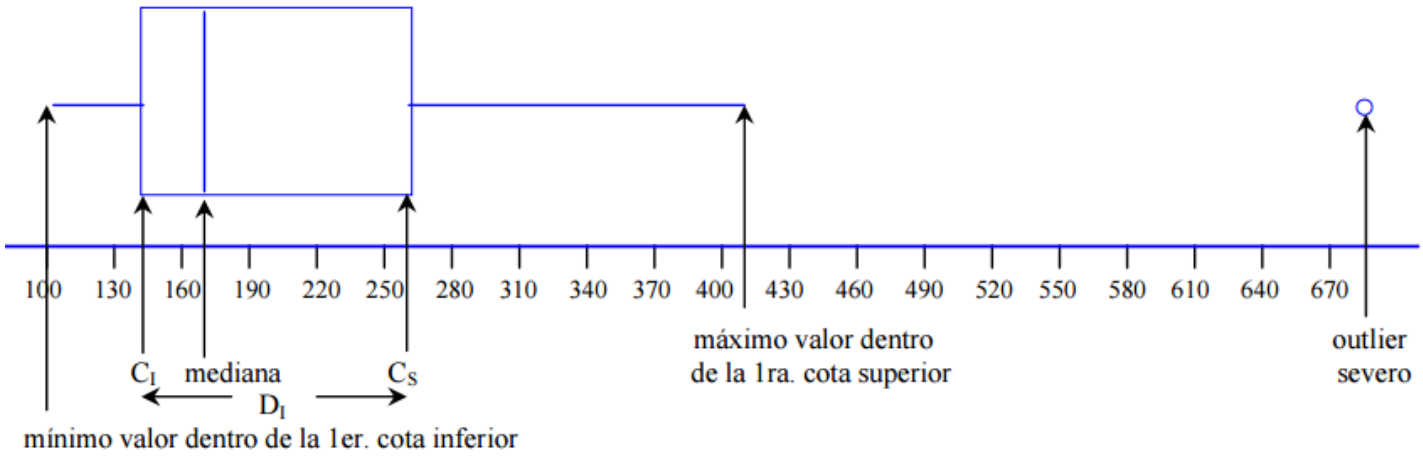
De los ejemplos anteriores sabemos que:

$$C_I = 140 \qquad \tilde{X} = 170 \qquad C_S = 320 \qquad D_I = 320 - 140 = 80$$

Calculamos las cotas:

$$2^a \text{ cota inferior} = C_I - 3\,D_I = 140 - 3 \cdot 80 = -100$$
$$1^a \text{ cota inferior} = C_I - 1.5\,D_I = 140 - 1.5 \cdot 80 = 20$$
$$1^a \text{ cota superior} = C_S + 1.5\,D_I = 320 + 1.5 \cdot 80 = 440$$
$$2^a \text{ cota superior} = C_S + 3\,D_I = 320 + 3 \cdot 80 = 580$$

El gráfico de caja resultante se muestra en la figura siguiente.




¿Qué se observa?

- Un dato outlier.
- La distribución de los datos es asimétrica hacia la derecha, la mitad inferior de los datos se distribuye en un rango mucho menor que la mitad superior.

¿Qué características de la distribución de los datos se manifiestan en un box-plot?

- Muestra los cinco números resúmenes
- Muestra una medida de posición robusta ⇒ MEDIANA
- Muestra una medida de dispersión robusta ⇒ DISTANCIA INTERCUARTIL
- Permite estudiar la simetría de la distribución
- Nos da un criterio de detección de datos outliers

Los distintos paquetes estadísticos dibujan box-plots que no siempre se basan en los criterios que hemos detallado aquí, algunos cambian el modo de calcular los cuartiles; otros, por ejemplo, ofrecen opciones de indicar la media y no la mediana en la caja.

Estos gráficos son muy útiles para comparar varias distribuciones. La Figura siguiente muestra los datos correspondientes a los resultados de una encuesta que se tomó en cuatro poblaciones diferentes las que se identifican de 1 a 4. La variable que se registró es el grado de satisfacción con el desempeño de los gobernantes en el último año (puntaje de 0 a 100).

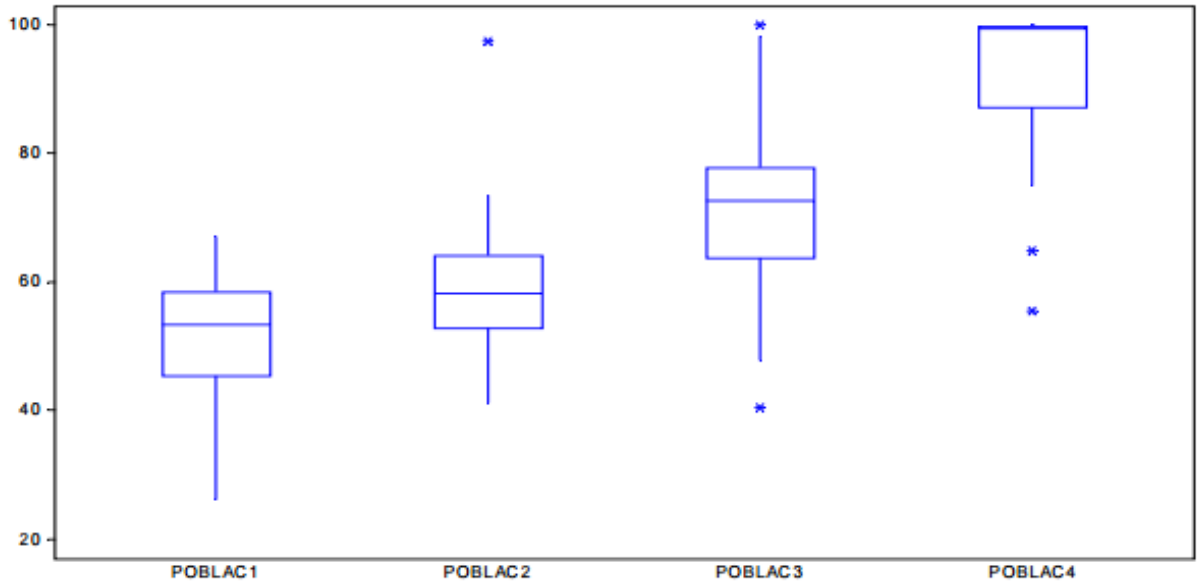
¿Qué concluimos a partir de este gráfico?

La satisfacción de los habitantes de las distintas poblaciones difiere en posición (la mediana cambia notablemente) y en dispersión (la población 3 presenta mayor dispersión que las demás).

Las distribuciones tienen diferentes formas (Población 4 muy asimétrica, mucha gente está totalmente de acuerdo con el desempeño de sus gobernantes, mientras que en las demás la distribución es simétrica).

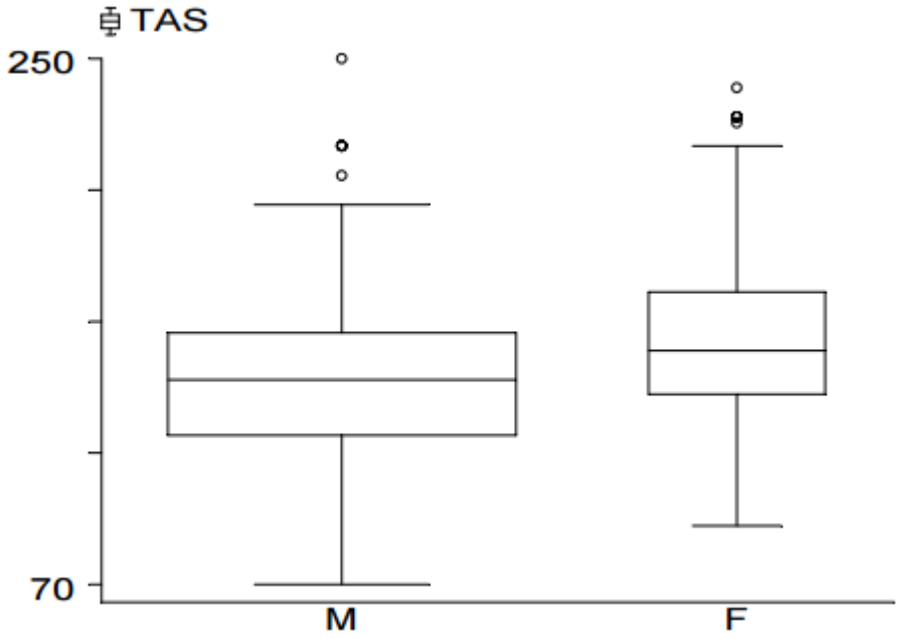
Podemos observar además que el cuartil inferior (percentil 25) del puntaje en la Población 3 es aproximadamente 63 y coincide con el cuartil superior de la Población 2, es decir, en la población 3 el 75% de los encuestados asignaron puntajes de 63 o más, en tanto que en la Población 2 sólo el 25% asignaron puntajes de 63 o más.

Del mismo modo, podemos observar que los encuestados de la Población 4 tienen un grado de satisfacción más alto que prácticamente todos los encuestados en las demás poblaciones.




Box-plot con ancho de la caja proporcional al número de observaciones.

Una variación útil de este gráfico consiste en representar las cajas con ancho proporcional al número de observaciones. Como ejemplo se presenta la distribución de presión arterial sistólica en personas adultas de ambos sexos, que concurrieron espontáneamente a medir su presión. Se dispone de 934 mediciones en varones y 477 mediciones en mujeres.




**RELACIONES ENTRE VARIABLES NUMÉRICAS**

Tal como ocurre en el caso univariado, el análisis de datos bivariados (dos variables medidas o registradas en el mismo individuo) comienza con el estudio del patrón o la estructura subyacente en los datos. Generalmente cuando se estudia la relación entre dos variables, registradas sobre el mismo individuo, una de ellas se considera variable de respuesta (efecto o resultado) y la otra se considera la variable independiente o explicatoria (potencial factor que afecta la variable respuesta).

En este enfoque el objetivo es analizar si existe relación entre ambas, y de ser posible, estudiar la naturaleza y la fuerza de la relación que las liga. Denotando por X la variable independiente y por Y la variable dependiente, un esquema simplificado de las situaciones que podemos encontrar se propone en la tabla siguiente.

| | **Independiente (X)** | **Dependiente (Y)** | **Ejemplo** |
|---|---|---|---|
| A) | Categórica | Categórica | X = hábito de fumar (no / <10 cig / ≥ 10)<br>Y = cáncer de pulmón (si / no) |
| B) | Categórica | Numérica | X = hábito de fumar<br>Y = nivel de colesterol sérico |
| C) | Numérica | Categórica | X = nivel de colesterol sérico<br>Y = infarto de miocardio (si / no) |
| D) | Numérica | Numérica | X = nivel de colesterol<br>Y = presión arterial |

En cada una de estas situaciones el enfoque analítico y el modo de resumen y presentación habitual de los datos cambia. Brevemente, el modo de resumir los datos en cada situación se presenta a continuación.

- **A)** Tablas de doble entrada y medidas de asociación (riesgo relativo, odds ratio, etc.).
- **B)** Medidas resúmenes de nivel de colesterol para cada grupo definido por hábito de fumar o box-plots para cada grupo.
- **C)** Un posible modo de resumir es categorizar la variable numérica y presentar la proporción de casos positivos (infarto de miocardio) en los distintos grupos definidos por nivel de colesterol.
- **D)** Gráficos de dispersión y medidas de correlación.

En cualquier caso, interesa estudiar si existe asociación entre las dos variables, pero el modo de medir asociación o efecto difiere.

En este capítulo consideraremos únicamente el problema de representar gráficamente dos variables numéricas y el modo de resumir la fuerza de la asociación entre dos variables numéricas. Finalmente consideraremos el caso en que la variable independiente es el tiempo, que merece un tratamiento especial y se conoce como análisis de series de tiempo.

**GRÁFICO DE DISPERSIÓN (SCATTER PLOT)**

Es un gráfico muy simple y útil para estudiar relaciones entre dos variables cuantitativas.

Se dibuja un sistema de coordenadas cartesianas en el que se representan los valores que toman las dos variables para cada sujeto o unidad de análisis. Se acostumbra a asignar la variable independiente al eje horizontal (comúnmente denominado eje X) y la variable dependiente al eje vertical (eje Y).

La nube resultante de puntos permite evaluar si existe relación entre las dos variables y la naturaleza de tal relación. Si es lineal, curvilínea, exponencial, logarítmica, cíclica, creciente, decreciente, etc. o si no hay relación aparente entre las variables. Para interpretar un gráfico de dispersión debe mirarse el patrón general que siguen los puntos. Este patrón debería revelar la dirección, forma y fuerza de la relación entre las dos variables.

Consideraremos algunos ejemplos.

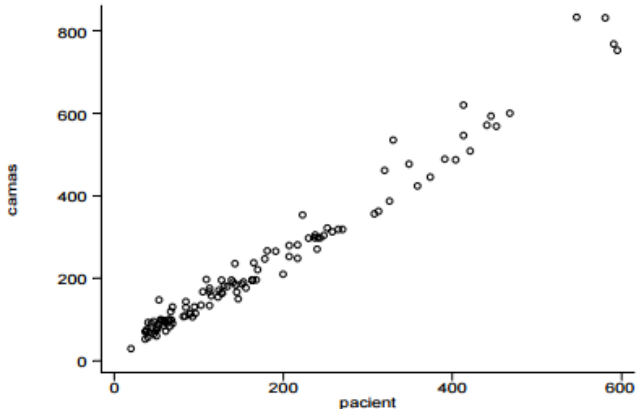
Los gráficos de la Figura 1 corresponden a datos de una muestra aleatoria de 56 hospitales participantes en el proyecto SENIC (Study on the Efficacy of Nosocomial Infection Control). El objetivo fundamental del Proyecto era determinar si los programas de vigilancia y control de infecciones habían reducido la tasa de infección hospitalaria en los Estados Unidos.

En a) hemos representado el número promedio de camas en el hospital durante el período de estudio y el número promedio de pacientes hospitalizados por día durante el período de estudio.

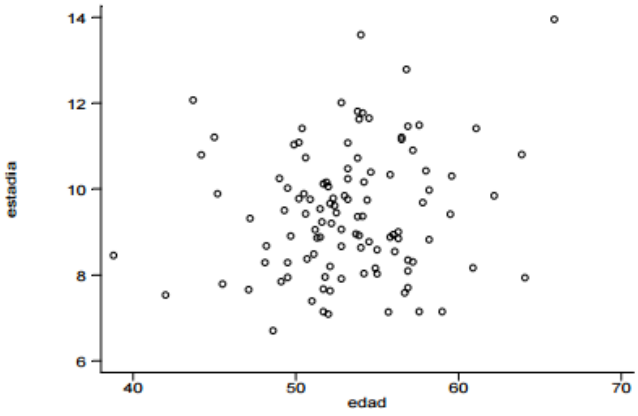
El gráfico b) muestra la relación entre duración promedio de la estadía de todos los pacientes en el hospital (en días) y edad promedio de todos los pacientes del hospital (en años).

**Figura 1. Gráficos de dispersión.**

a) Número de pacientes versus número de camas en hospitales.




b) Tiempo de internación versus edad del paciente.

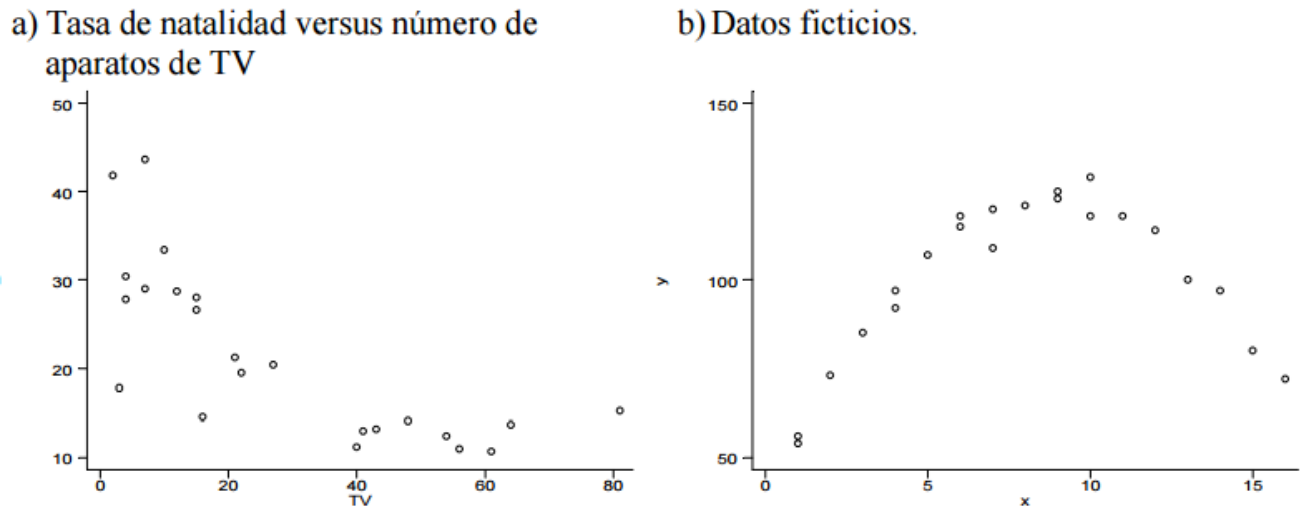



¿Qué nos dicen los gráficos de la Figura 1 acerca de la relación entre las variables?

**Figura 1 a)** Número de camas y número de pacientes están fuertemente relacionados. Cuando una variable aumenta la otra también aumenta, es decir, entre ambas variables existe una asociación positiva. Además podemos proponer que la relación entre ambas variables es lineal ya que una línea recta aproximaría bastante bien la tendencia general de la nube de puntos.

**Figura 1 b)** No parece haber relación entre el tiempo de internación y la edad del paciente. Si nos ubicamos en alguna edad particular, digamos 50 años, podemos encontrar pacientes cuya internación tuvo una duración de cualquier magnitud. La nube de puntos no presenta una tendencia particular.

**Figura 2. Gráficos de dispersión**




En la **Figura 2 a)** hemos representado la tasa de nacimiento cruda (número de nacimientos cada 1000 habitantes) y el número de televisores cada 100 habitantes para 26 naciones (desarrolladas y en vías de desarrollo). Fuente: Statistical Abstract of the United States, 1995 and Human Development Report, 1995, Oxford University Press.

En la **Figura 2 b)** se muestran datos ficticios de dos variables X e Y.
¿Qué nos dicen los gráficos de la **Figura 2** acerca de la relación entre las variables?

**Figura 2 a).** La tasa de natalidad está inversamente relacionada con el número de televisores cada 100 habitantes. Cuando el número de televisores aumenta, la tasa de natalidad disminuye. Además, el decrecimiento no es lineal (una línea recta no es un buen modelo para el tipo de relación que se observa entre las dos variables). Cuando el número de televisores es bajo (cercano a cero), un aumento de 20 televisores por cada 100 habitantes produce una importante disminución de la tasa de natalidad, mientras que si el número de televisores es alto (más de 40), un aumento de la misma magnitud en el número de televisores produce una disminución despreciable en la tasa de natalidad. La relación entre las dos variables podría describirse como exponencial negativa.

**Figura 2 b).** X e Y están fuertemente relacionadas, podemos proponer que la relación entre ambas es curvilínea. No podemos hablar de dirección de la relación ya que es en parte creciente y en parte decreciente.

Al estudiar la relación entre dos variables CUANTITATIVAS.

En general interesa:

- ✓ Investigar si existe asociación entre las dos variables.
- ✓ Cuantificar la fuerza de la asociación, a través de una medida de asociación denominada coeficiente de correlación.
- ✓ Estudiar la forma de la relación y en lo posible proponer un modelo matemático para la relación.
- ✓ Predecir una variable a partir de la otra usando el modelo propuesto (REGRESIÓN).

Un MODELO MATEMÁTICO es una función matemática que propone la forma de relación entre la variable dependiente (Y) y la o las variables independientes.

La función más simple para la relación entre dos variables es la FUNCIÓN LINEAL.

$$Y = a + b \cdot X$$

Un **MODELO DETERMINÍSTICO** supone que, bajo condiciones ideales, el comportamiento de la variable dependiente puede ser totalmente descripto por una función matemática de las variables independientes (o por un conjunto de ecuaciones que relacionen las variables). Es decir, en condiciones ideales el modelo permite predecir SIN ERROR el valor de la variable dependiente.

Ejemplo: Ley de la Gravedad. Podemos predecir exactamente la posición, en cada instante de tiempo, de un objeto que cae libremente en el vacío.

Un **MODELO ESTADÍSTICO** permite incorporar un componente aleatorio en la relación. Debido a este componente aleatorio, las predicciones obtenidas a través de modelos estadísticos tendrán asociado un error de predicción.

Ejemplo: Relación de la altura con la edad en niños.

Niños de la misma edad seguramente no tendrán la misma altura. Sin embargo, a través de un modelo estadístico es posible concluir que la altura aumenta con la edad. Es más, podríamos predecir la altura de un niño de cierta edad y asociarle un error de predicción que tiene en cuenta los errores de medición y la variabilidad entre individuos. En problemas biológicos, trabajando en "condiciones ideales" es posible evitar los errores de medición, pero no la variabilidad individual, por eso es indispensable incluir el componente aleatorio en los modelos estadísticos.

**COEFICIENTE DE CORRELACIÓN**

El grado de asociación entre dos variables numéricas puede ser resumido en un estadístico denominado.

**COEFICIENTE DE CORRELACIÓN**.

Presentaremos en primer lugar el coeficiente de correlación de Pearson, que mide el grado de asociación lineal entre dos variables y posteriormente un estadístico basado en rangos que estima la correlación sin hacer supuestos sobre el tipo de relación entre las variables.

**COEFICIENTE DE CORRELACIÓN DE PEARSON**

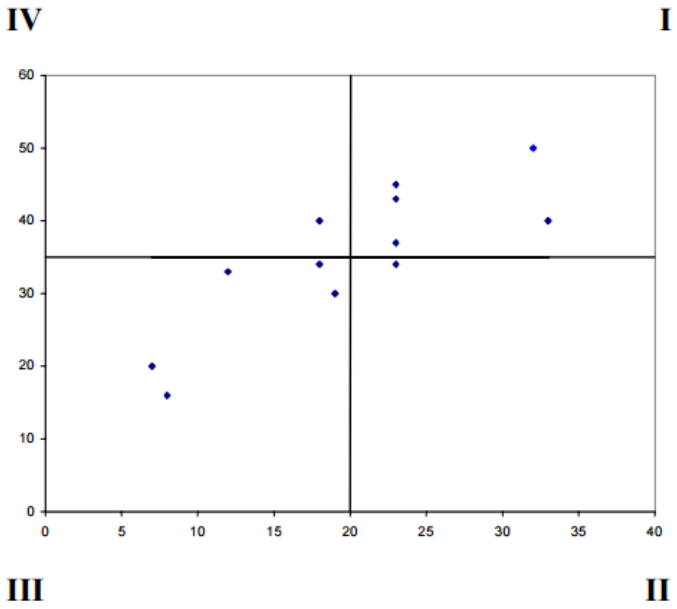
Supongamos que tenemos dos variables (X, Y) registradas en cada una de los n sujetos de una muestra. Sean ($X_i$, $Y_i$) las observaciones realizadas para cada variable en el sujeto i- ésimo. Definimos la covarianza muestral entre X e Y como:

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1} \quad \text{donde} \quad \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{e} \quad \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}.$$

La covarianza es el "promedio" de los productos de las desviaciones de las variables respecto de las correspondientes medias. ¿Cómo se interpreta la covarianza?

En la **Figura 3** se representa una nube de puntos correspondiente a los distintos pares (X, Y) observados en una muestra. Trazamos rectas paralelas a los ejes de coordenadas que pasan por $\overline{X}$ e $\overline{Y}$ y dividimos el plano en cuatro cuadrantes I, II, III y IV.

**Figura 3**



Consideremos ul punto en el Cuadrante I.: la diferencia $(X - \overline{X}) > 0$ y la diferencia $(Y - \overline{Y}) > 0$ y lo mismo ocurre con el signo de las diferencias para cualquier punto ubicado en este cuadrante. Por lo tanto, el producto $(X - \overline{X})(Y - \overline{Y}) > 0$ . Usando el mismo razonamiento para puntos ubicados en los demás cuadrantes obtenemos la siguiente tabla.

| Cuadrante | $(X - \overline{X})$ | $(Y - \overline{Y})$ | $(X - \overline{X})(Y - \overline{Y})$ |
|---|---|---|---|
| I | + | + | + |
| II | + | – | – |
| III | – | – | + |
| IV | – | + | – |

Por lo tanto:

- Si la mayoría de los puntos se encuentran en los cuadrantes I y III la covarianza se construirá básicamente con sumandos positivos y por lo tanto será positiva. Este es el caso de los datos de la Figura 3 en la que la Cov(X, Y) = 738.
- Si la mayoría de los puntos se encuentran en los cuadrantes II y IV la mayoría de los sumandos serán negativos y la covarianza será negativa (Figura 4 a, Cov = -1098).
- Si los puntos se encuentran homogéneamente distribuidos por los cuatro cuadrantes, la covarianza será cercana a cero (Figura 4 b, Cov = -15).

Figura 4 a

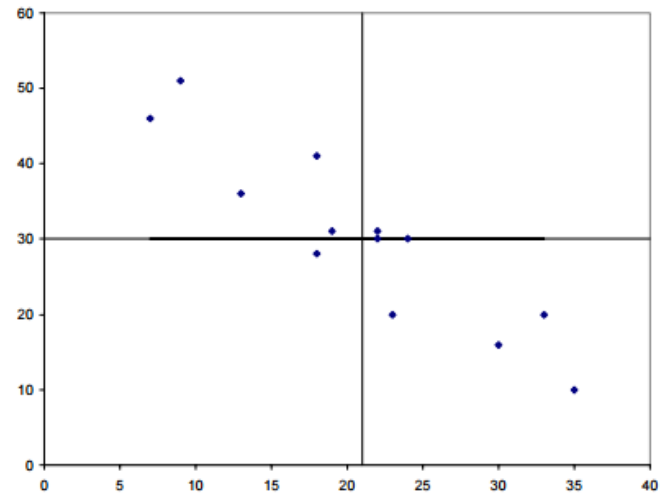
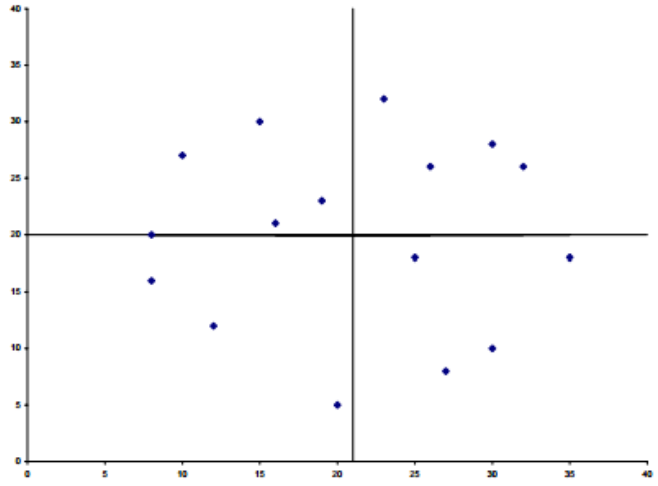



Figura 4 b




Podemos interpretar el signo de la covarianza, pero no su magnitud ya que ésta depende de las unidades de los datos. Si en el **Gráfico 4a**, cambiamos las unidades de la variable X dividiendo cada valor por 1000 (si X representara peso, sería equivalente a transformar el peso de gramos a Kg) la covarianza pasa de – 1098 a – 1.098. Por lo tanto, es importante estandarizar la covarianza de modo que no dependa de las unidades de las variables.

Definición Sean $(X_i, Y_i)$ las observaciones realizadas en cada uno de los n-sujetos de una muestra de tamaño n. Definimos el coeficiente de correlación muestral de Pearson entre X e Y como:

$$r = Corr(X,Y) = \frac{\text{cov}(X,Y)}{s_X\ s_Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)\ s_X\ s_Y}$$

donde $s_x$ y $s_y$ son los desvíos estándares muestrales de las variables X e Y respectivamente.

Ejemplo:

| | X | Y | $(X - \overline{X})$ | $(Y - \overline{Y})$ | $(X - \overline{X})(Y - \overline{Y})$ |
|---|---|---|---|---|---|
| | 3 | 10 | -3.86 | 3.14 | -12.12 |
| | 6 | 7 | -0.86 | 0.14 | -0.12 |
| | 5 | 9 | -1.86 | 2.14 | -3.98 |
| | 8 | 6 | 1.14 | -0.86 | -0.98 |
| | 9 | 8 | 2.14 | 1.14 | 2.45 |
| | 10 | 7 | 3.14 | 0.14 | 0.45 |
| | 7 | 8 | 0.14 | 1.14 | 0.16 |
| Media | 6.86 | 7.86 | | Suma = | -14.14 |
| DS | 2.41 | 1.35 | | | |

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)\ s_X\ s_Y} = \frac{-14.14}{(7-1)\ 2.41\ 1.35} = -0.73$$

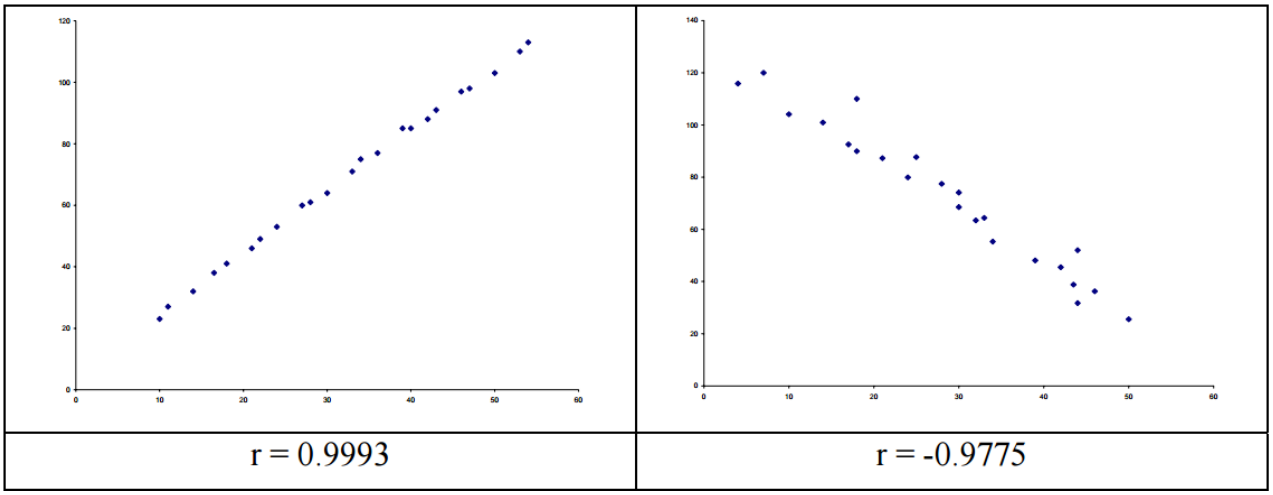
Propiedades del coeficiente de correlación de Pearson:

- $r$ toma valores entre –1 y 1 $(-1 \le r \le 1)$,
- $r$ mide la *fuerza* de la asociación LINEAL entre X e Y,
- $r = 0$ implica que no hay relación lineal entre las variables,
- r = + 1 implica que todos los puntos caen sobre una recta de pendiente positiva (asociación positiva),
- r = – 1 implica que todos los puntos caen sobre una recta de pendiente negativa (asociación negativa),
- mientras mayor el valor absoluto de $r$ mayor la fuerza de la asociación,
- el valor de $r$ no depende de las unidades de medición,
- el coeficiente de correlación trata a X e Y simétricamente, no identifica cual es la variable dependiente y cual la independiente.

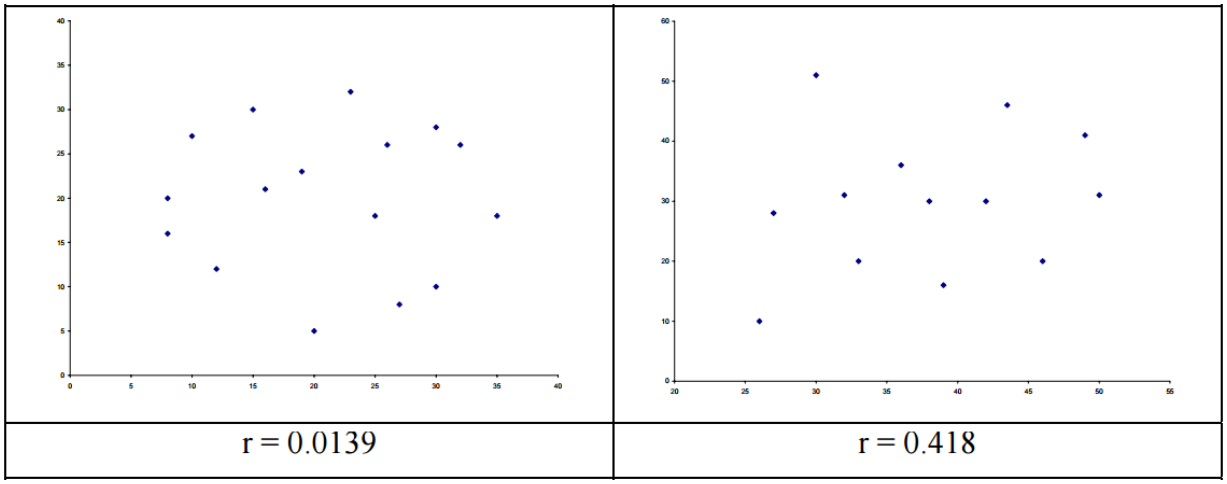
¿Qué mide exactamente el coeficiente de correlación de Pearson?

Cuán cercanos se encuentran los puntos alrededor de una LÍNEA RECTA que indique la tendencia general.

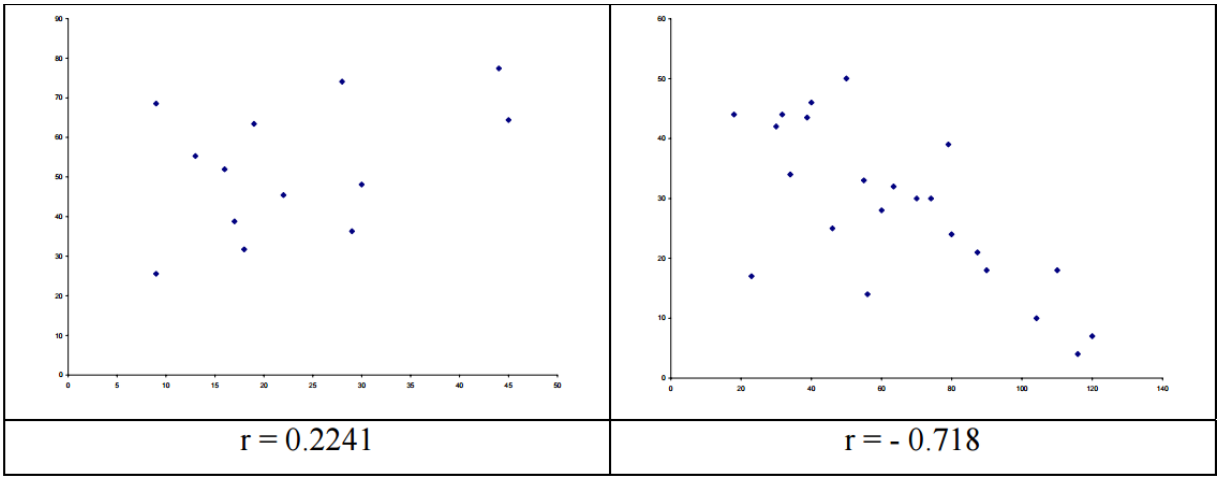
**Figura 5. Ejemplos de conjuntos de datos con diferente grado de correlación.**



| r = 0.9993 | r = -0.9775 |
|---|---|



| r = 0.0139 | r = 0.418 |
|---|---|

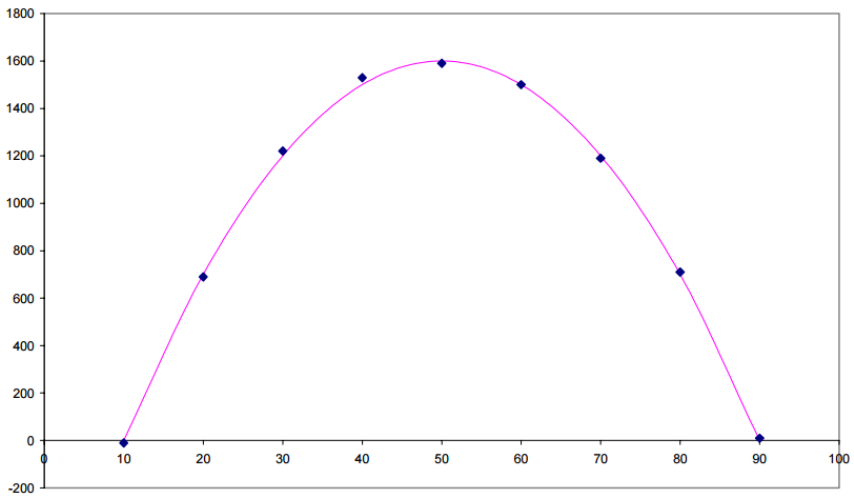


| r = 0.2241 | r = - 0.718 |
|---|---|

¿Qué ocurre si la relación entre las variables no es lineal?

Si el supuesto de linealidad no se cumple el valor del coeficiente de correlación puede ser engañoso. Consideremos el gráfico de la **Figura 6**, en que la relación entre las dos variables es en forma de U. En este caso el coeficiente r es cercano a cero, es decir, a partir de r concluiríamos que las variables NO están asociadas. Sin embargo, las variables están fuertemente asociadas ya que los valores de Y sigue una relación casi determinística con el valor de X, el problema es que esta relación no es lineal.

**Figura 6:**



¿Cómo afectan los datos outliers al coeficiente de correlación?

En principio el coeficiente de correlación de Pearson puede ser calculado para cualquier conjunto de datos en el que los pares ordenados sigan una relación aproximadamente lineal. Sin embargo, una observación outlier respecto de la relación, que no se encuentra en la tendencia general de los datos, puede influir notablemente en la magnitud del coeficiente. Una observación se denomina influyente cuando produce un cambio importante en el coeficiente de regresión lineal o en la recta que se propondría como modelo para la relación entre las variables.

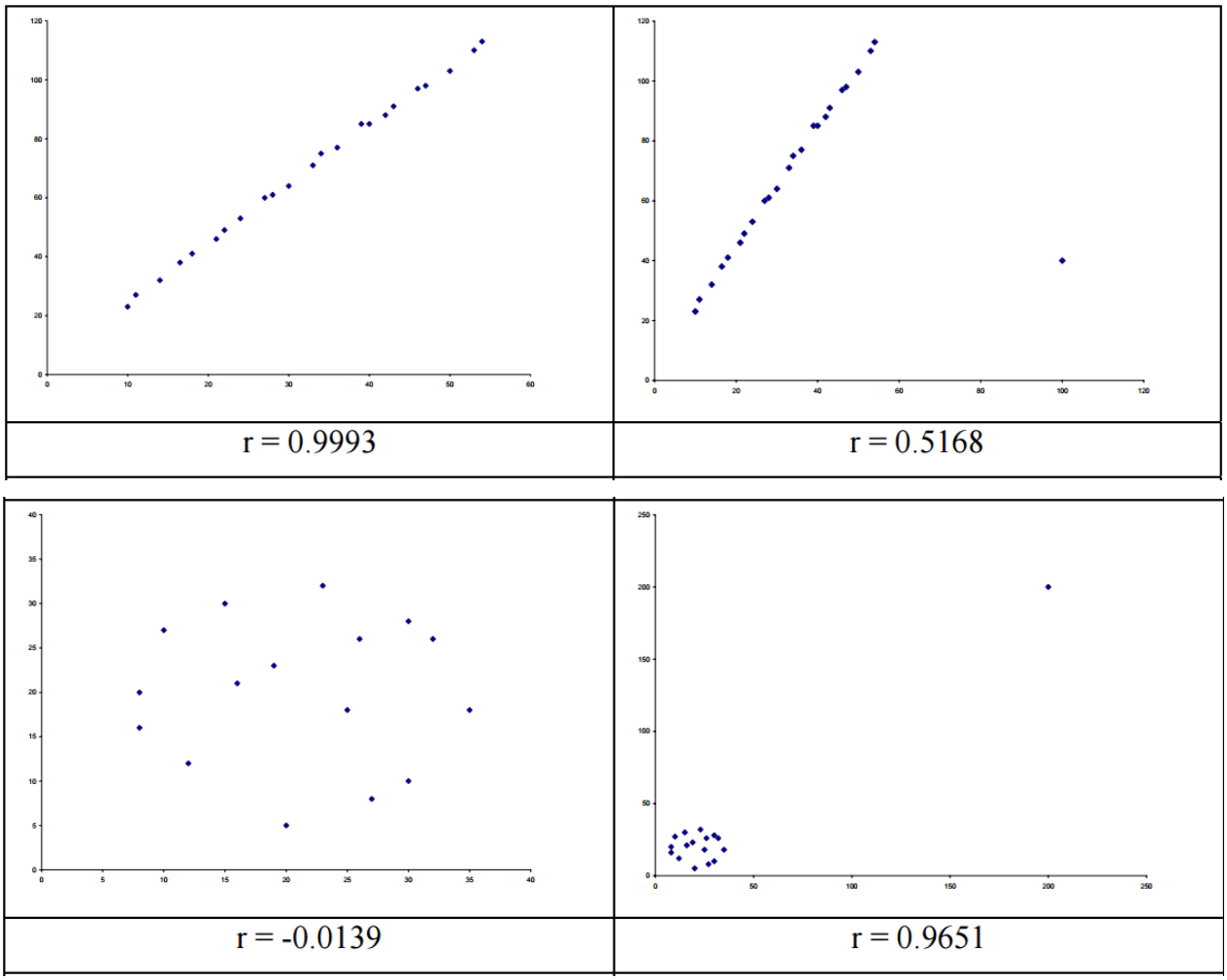
En la **Figura 7** se presentan tres de los gráficos de la **Figura 5** a los que se les agregó solamente 1 punto, que en cada caso logra modificar notablemente el valor del coeficiente de correlación de Pearson.

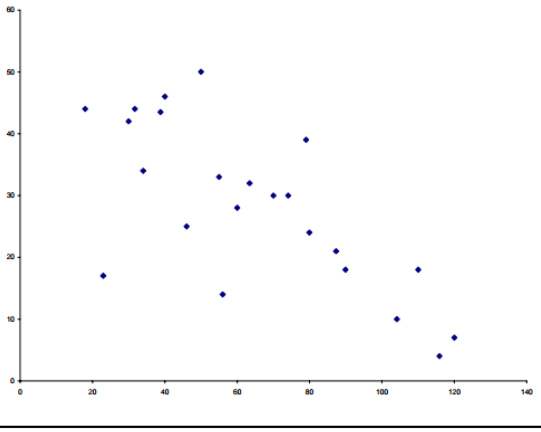
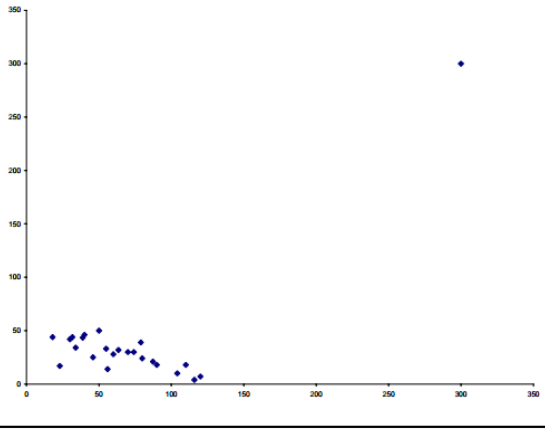
En conclusión:

- El coeficiente de correlación de Pearson es una medida muy sensible a la presencia de datos influyentes.
- El coeficiente de Pearson cuantifica la fuerza de la relación LINEAL entre las dos variables. Antes de calcularlo es necesario hacer un gráfico para decidir si la relación entre las variables es aproximadamente lineal y si no hay puntos influyentes. En general, el coeficiente de correlación de Pearson es una buena medida resumen del grado de asociación entre dos variables numéricas cuando el gráfico muestra una nube de puntos elíptica.

Por último, diremos que mostrar que dos variables están asociadas, no implica que exista relación de causalidad entre ellas.

**Figura 7. Efectos de datos influyentes sobre el coeficiente de correlación de Pearson.**



| | |
|---|---|
|  |  |
| r = -0.718 | r = 0.8971 |

Resumiendo, una medida de correlación entre dos variables X e Y debería satisfacer los siguientes requerimientos:

- Tomar valores entre −1 y 1.
- Si los valores más grandes de X tienden a aparecer con los valores más grandes de Y y los menores de X con los menores de Y, entonces la medida de correlación debería ser positiva y cercana a 1 cuando la tendencia sea muy fuerte. Decimos entonces que X e Y tienen correlación positiva.
- Si los mayores valores de X tienden a aparecer junto con los menores valores de Y y vice cersa, entonces la medida de correlación debería ser negativa, con −1 indicando que la tendencia es fuerte. Decimos entonces que X e Y están negativamente correlacionadas.
- Si los valores de X aparecen aleatoriamente apareados con los de Y, la medida de correlación debería ser próxima a cero. Decimos entonces, que X e Y no están correlacionados.

Existen otras medidas para resumir correlación que satisfacen los requerimientos anteriores pero que son robustas a la presencia de datos influyentes. Presentamos a continuación una propuesta alternativa para medir correlación que se construye ordenando los datos.

**COEFICIENTE DE CORRELACIÓN DE SPEARMAN**

Disponemos de n pares de observaciones $(X_1, Y_1)$, ..., $(X_n, Y_n)$. Las variables pueden ser numéricas o categóricas ordinales.

¿Cómo se calcula el coeficiente de Spearman?

1. Se ordenan los valores de cada variable por separado y se reemplaza cada observación por la posición (rango) que ésta ocupa en la muestra ordenada.
2. Se calcula el coeficiente de Pearson usando como datos los rangos.

Características

- Como el coeficiente de correlación de Spearman varía entre −1 y 1.
- Mide la fuerza de la correlación entre las dos variables. Valores positivos indican que la relación entre X e Y es creciente. Valores negativos indican que la relación es decreciente. Valores cercanos a cero indican que la relación no es creciente ni decreciente.
- No hace supuestos sobre la forma de la relación entre las dos variables.

Ejemplo:

Para los datos de la tabla siguiente calculamos el coeficiente de correlación de Spearman:

$$r_S = \frac{(1-4.5)(7-4.5)+(3-4.5)(3.5-4.5)+\ldots+(8-4.5)(8-4.5)}{2.45\cdot 2.45\cdot(8-1)} = 0.0000$$

Para estos datos el coeficiente de Pearson es r = 0.8355. ¿Por qué tanta diferencia entre ambos? La **Figura 8** muestra que la diferencia se debe a la presencia de un punto fuertemente influyente.

| | X | Y | Rango (X) | Rango(Y) |
|---|---|---|---|---|
| | 10 | 17 | 1 | 7 |
| | 13 | 14 | 3 | 3.5 |
| | 12 | 16 | 2 | 6 |
| | 15 | 13 | 5 | 2 |
| | 16 | 15 | 6 | 5 |
| | 17 | 14 | 7 | 3.5 |
| | 14 | 12 | 4 | 1 |
| | 30 | 30 | 8 | 8 |
| Media | 15.88 | 16.38 | 4.5 | 4.5 |
| DS | 6.13 | 5.73 | 2.45 | 2.43 |

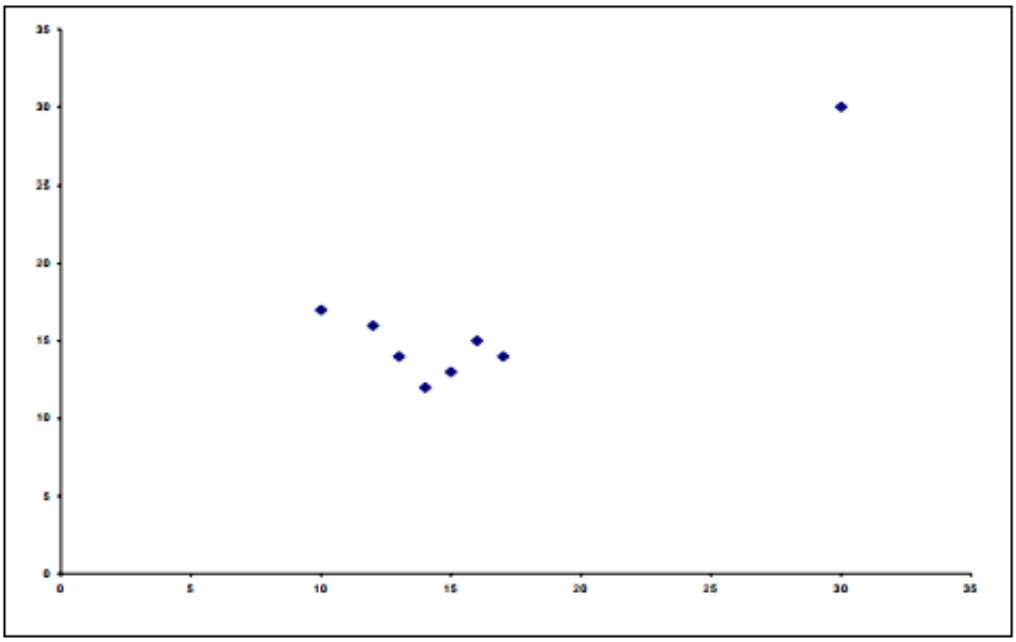



**Figura 8.** Efecto de un dato influyente sobre el coeficiente de correlación de Pearson.

¿Cuándo usar coeficiente de Spearman (u otro basado en rangos)?

- Cuando las variables tienen una relación creciente o decreciente pero no necesariamente lineal.
- Cuando hay datos influyentes.
- Cuando la forma de la nube de puntos no es elipsoidal.

**GRÁFICOS ENGAÑOSOS**

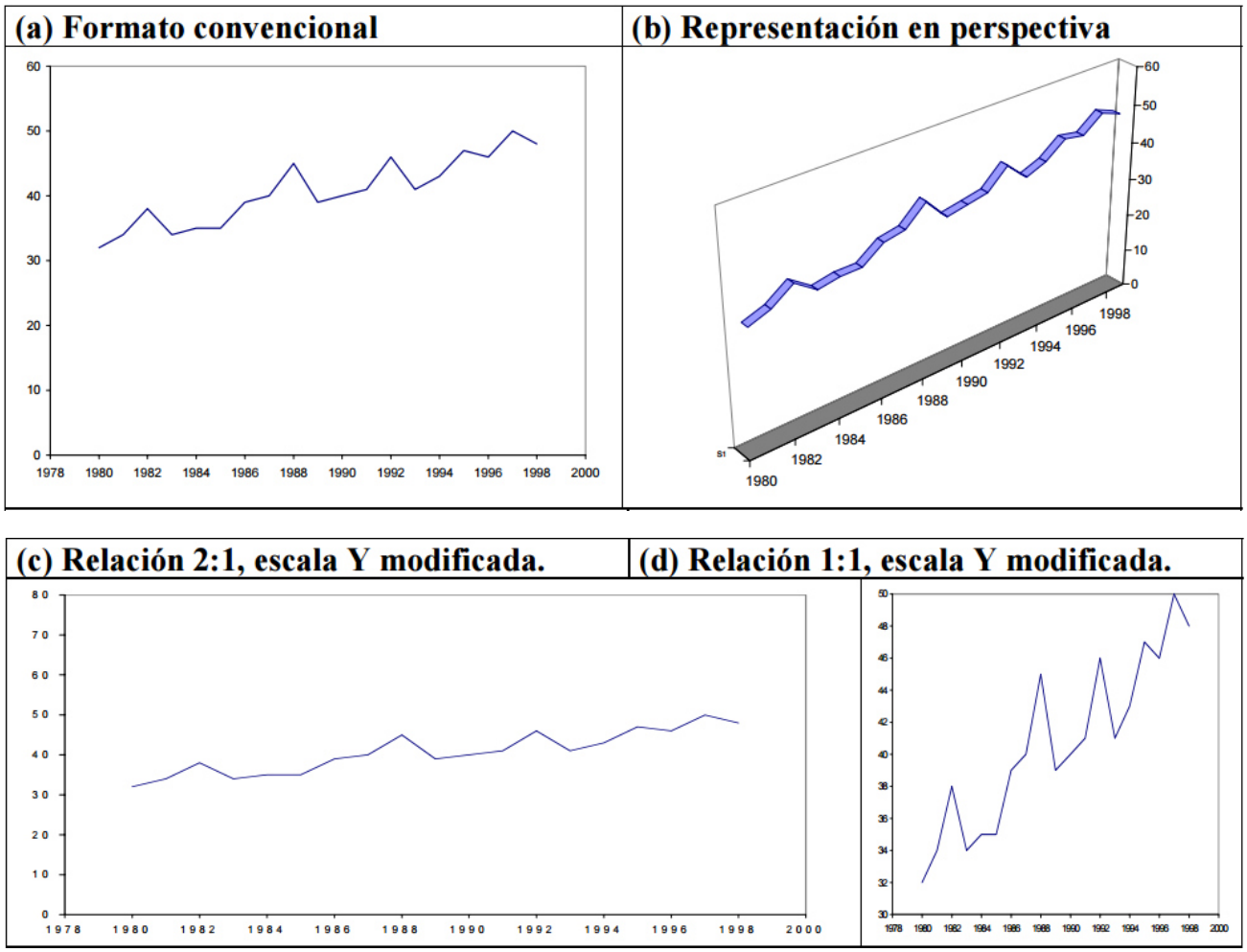
Cuando se trata de gráficos de dispersión o de series, la imagen visual puede modificarse notablemente usando uno o más de los siguientes recursos:

- Cambiando la escala de uno o ambos ejes,
- Eliminando el cero de la escala vertical en la representación,
- Cambiando la relación de longitud entre ambos ejes.

Los gráficos XY por convención se representan respetando una relación 4:3 entre el eje horizontal y el vertical, prácticamente todos los paquetes que construyen gráficos respetan esta convención.

La **Figura 14** muestra cuatro representaciones diferentes de los mismos datos de una serie anual donde se pretende mostrar cómo estos cambios pueden afectar la interpretación de la imagen.

**Figura 14.** Distintos formatos para la misma serie de tiempo:




La **Figura 14 (a)** muestra el gráfico obtenido respetando la relación 4:3 y usando la escala del eje vertical que comienza en cero. Se observa una tendencia moderadamente creciente y fluctuaciones moderadas.

En la **Figura 14 (b)** se realizó una "bonita" representación en perspectiva, respetando las escalas que se usaron en **(a).** Este gráfico puede producir una sensación de tendencia más marcada que el gráfico anterior o una impresión de que no hay tendencia, dependiendo del observador.

En **(c)** modificamos la relación horizontal:vertical, de 4:3 a 2:1, y aumentamos la escala del eje Y. Resultado: la tendencia y las fluctuaciones parecen poco importantes.

Finalmente en el gráfico **(d)** cambiamos la relación horizontal:vertical a 1:1 y modificamos la escala vertical logrando de este modo magnificar notablemente la tendencia y la importancia de las fluctuaciones.

Todos los gráficos de la **Figura 14** son correctos en el sentido que se construyeron usando la misma información (no hemos falseado o modificado los datos para construirlos). Sin embargo, algunos de ellos producen impresiones engañosas amplificando o disimulando diferencias que existen.

**INFORMACIÓN (INCLUÍDA EN ESTE DOCUMENTO EDUCATIVO) TOMADA DE:**

**Sitios web:**

**1.** http://www.dm.uba.ar/materias/estadistica_Q/2011/1/modulo%20descriptiva.pdf